CHAPTER – 12

CORRELATION AND REGRESSION

LEARNING OBJECTIVES

After reading this chapter a student will be able to understand–

- The meaning of bivariate data and techniques of preparation of bivariate distribution;
- The concept of correlation between two variables and quantitative measurement of correlation including the interpretation of positive, negative and zero correlation;
- Concept of regression and its application in estimation of a variable from known set of data.

12.1 INTRODUCTION

In the previous chapter, we discussed many a statistical measure relating to Univariate distribution i.e. distribution of one variable like height, weight, mark, profit, wage and so on. However, there are situations that demand study of more than one variable simultaneously. A businessman may be keen to know what amount of investment would yield a desired level of profit or a student may want to know whether performing better in the selection test would enhance his or her chance of doing well in the final examination. With a view to answering this series of questions, we need to study more than one variable at the same time. Correlation Analysis and Regression Analysis are the two analyses that are made from a multivariate distribution i.e. a distribution of more than one variable. In particular when there are two variables, say x and y, we study bivariate distribution. We restrict our discussion to bivariate distribution only.

Correlation analysis, it may be noted, helps us to find an association or the lack of it between the two variables x and y. Thus if x and y stand for profit and investment of a firm or the marks in Statistics and Mathematics for a group of students, then we may be interested to know whether x and y are associated or independent of each other. The extent or amount of correlation between x and y is provided by different measures of Correlation namely Product Moment Correlation Coefficient or Rank Correlation Coefficient or Coefficient of Concurrent Deviations. In Correlation analysis, we must be careful about a cause and effect relation between the variables under consideration because there may be situations where x and y are related due to the influence of a third variable although no causal relationship exists between the two variables.

Regression analysis, on the other hand, is concerned with predicting the value of the dependent variable corresponding to a known value of the independent variable on the assumption of a mathematical relationship between the two variables and also an average relationship between them.

12.2 BIVARIATE DATA

When data are collected on two variables simultaneously, they are known as bivariate data and the corresponding frequency distribution, derived from it, is known as Bivariate Frequency Distribution. If x and y denote marks in Maths and Stats for a group of 30 students, then the corresponding bivariate data would be (x_i, y_i) for i = 1, 2, …. 30 where (x_1, y_1) denotes the marks in Mathematics and Statistics for the student with serial number or Roll Number 1, (x_2, y_2), that for the student with Roll Number 2 and so on and lastly (x_{30}, y_{30}) denotes the pair of marks for the student bearing Roll Number 30.

As in the case of a Univariate Distribution, we need to construct the frequency distribution for bivariate data. Such a distribution takes into account the classification in respect of both the variables simultaneously. Usually, we make horizontal classification in respect of x and vertical classification in respect of the other variable y. Such a distribution is known as Bivariate Frequency Distribution or Joint Frequency Distribution or Two way classification of the two variables x and y.

Illustration

Example 12.1 Prepare a Bivariate Frequency table for the following data relating to the marks in Statistics (x) and Mathematics (y):

(15, 13),	(1, 3),	(2, 6),	(8, 3),	(15, 10),	(3, 9),	(13, 19),
(10, 11),	(6, 4),	(18, 14),	(10, 19),	(12, 8),	(11, 14),	(13, 16),
(17, 15),	(18, 18),	(11, 7),	(10, 14),	(14, 16),	(16, 15),	(7, 11),
(5, 1),	(11, 15),	(9, 4),	(10, 15),	(13, 12)	(14, 17),	(10, 11),
(6, 9),	(13, 17),	(16, 15),	(6, 4),	(4, 8),	(8, 11),	(9, 12),
(14, 11),	(16, 15),	(9, 10),	(4, 6),	(5, 7),	(3, 11),	(4, 16),
(5, 8),	(6, 9),	(7, 12),	(15, 6),	(18, 11),	(18, 19),	(17, 16)
(10, 14),						

Take mutually exclusive classification for both the variables, the first class interval being 0-4 for both.

Solution

From the given data, we find that

Range for x = 19–1 = 18

Range for y = 19–1 = 18

We take the class intervals 0-4, 4-8, 8-12, 12-16, 16-20 for both the variables. Since the first pair of marks is (15, 13) and 15 belongs to the fourth class interval (12-16) for x and 13 belongs to the fourth class interval for y, we put a stroke in the (4, 4)-th cell. We carry on giving tally marks till the list is exhausted.

Table 12.1

Bivariate Frequency Distribution of Marks in Statistics and Mathematics.

		MARKS IN MATHS					
	X \ Y	0-4	4-8	8-12	12-16	16-20	Total
MARKS IN STATS	0–4	I (1)	I (1)	II (2)			4
	4–8	I (1)	IIII (4)	~~IIII~~ (5)	I (1)	I (1)	12
	8–12	I (1)	II (2)	IIII (4)	~~IIII~~ I (6)	I (1)	14
	12–16		I (1)	III (3)	II (2)	~~IIII~~ (5)	11
	16–20			I (1)	~~IIII~~ (5)	III (3)	9
	Total	3	8	15	14	10	50

We note, from the above table, that some of the cell frequencies (f_{ij}) are zero. Starting from the above Bivariate Frequency Distribution, we can obtain two types of univariate distributions which are known as:

(a) Marginal distribution.

(b) Conditional distribution.

If we consider the distribution of Statistics marks along with the marginal totals presented in the last column of Table 12-1, we get the marginal distribution of marks in Statistics. Similarly, we can obtain one more marginal distribution of Mathematics marks. The following table shows the marginal distribution of marks of Statistics.

Table 12.2

Marginal Distribution of Marks in Statistics

Marks	No. of Students
0-4	4
4-8	12
8-12	14
12-16	11
16-20	9
Total	50

We can find the mean and standard deviation of marks in Statistics from Table 12.2. They would be known as marginal mean and marginal SD of Statistics marks. Similarly, we can obtain the marginal mean and marginal SD of Mathematics marks. Any other statistical measure in respect of x or y can be computed in a similar manner.

If we want to study the distribution of Statistics Marks for a particular group of students, say for those students who got marks between 8 to 12 in Mathematics, we come across another univariate distribution known as conditional distribution.

Table 12.3

Conditional Distribution of Marks in Statistics for Students having Mathematics Marks between 8 to 12

Marks	No. of Students
0-4	2
4-8	5
8-12	4
12-16	3
16-20	1
Total	15

We may obtain the mean and SD from the above table. They would be known as conditional mean and conditional SD of marks of Statistics. The same result holds for marks in Mathematics. In particular, if there are m classifications for x and n classifications for y, then there would be altogether (m + n) conditional distribution.

12.3 CORRELATION ANALYSIS

While studying two variables at the same time, if it is found that the change in one variable is reciprocated by a corresponding change in the other variable either directly or inversely, then the two variables are known to be associated or correlated. Otherwise, the two variables are known to be dissociated or uncorrelated or independent. There are two types of correlation.

(i) Positive correlation

(ii) Negative correlation

If two variables move in the same direction i.e. an increase (or decrease) on the part of one variable introduces an increase (or decrease) on the part of the other variable, then the two variables are known to be positively correlated. As for example, height and weight yield and rainfall, profit and investment etc. are positively correlated.

On the other hand, if the two variables move in the opposite directions i.e. an increase (or a decrease) on the part of one variable results a decrease (or an increase) on the part of the other variable, then the two variables are known to have a negative correlation. The price and demand of an item, the profits of Insurance Company and the number of claims it has to meet etc. are examples of variables having a negative correlation.

The two variables are known to be uncorrelated if the movement on the part of one variable does not produce any movement of the other variable in a particular direction. As for example, Shoe-size and intelligence are uncorrelated.

12.4 MEASURES OF CORRELATION

We consider the following measures of correlation:

(a) Scatter diagram

(b) Karl Pearson's Product moment correlation coefficient

(c) Spearman's rank correlation co-efficient

(d) Co-efficient of concurrent deviations

(a) SCATTER DIAGRAM

This is a simple diagrammatic method to establish correlation between a pair of variables. Unlike product moment correlation co-efficient, which can measure correlation only when the variables are having a linear relationship, scatter diagram can be applied for any type of correlation – linear as well as non-linear i.e. curvilinear. Scatter diagram can distinguish between different types of correlation although it fails to measure the extent of relationship between the variables.

Each data point, which in this case a pair of values (x_i, y_i) is represented by a point in the rectangular axes of cordinates. The totality of all the plotted points forms the scatter diagram. The pattern of the plotted points reveals the nature of correlation. In case of a positive correlation, the plotted points lie from lower left corner to upper right corner, in case of a negative correlation the plotted points concentrate from upper left to lower right and in case of zero correlation, the plotted points would be equally distributed without depicting any particular pattern. The following figures show different types of correlation and the one to one correspondence between scatter diagram and product moment correlation coefficient.

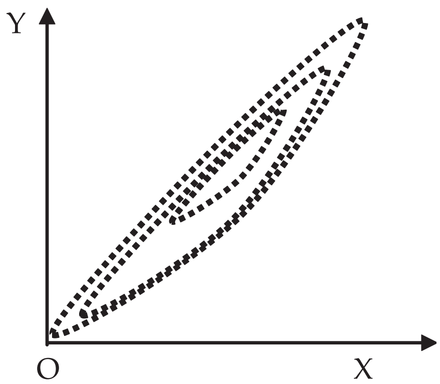

FIGURE 12.1
Showing Positive Correlation
(0 < r <1)

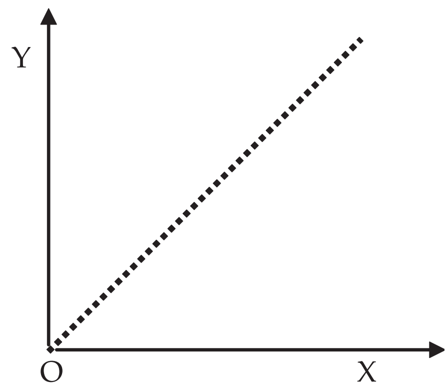

FIGURE 12.2
Showing perfect Correlation
(r = 1)

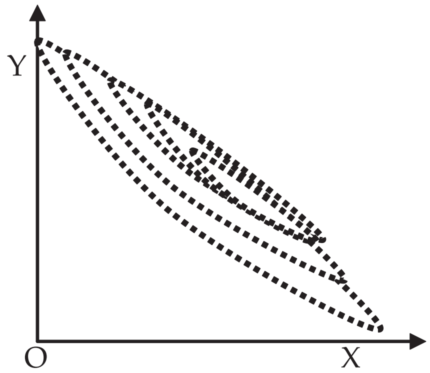

FIGURE 12.3
Showing Negative Correlation

(–1 < r <0)

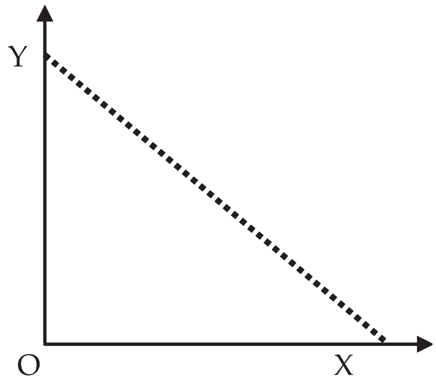

FIGURE 12.4
Showing perfect Negative Correlation
(r = –1)

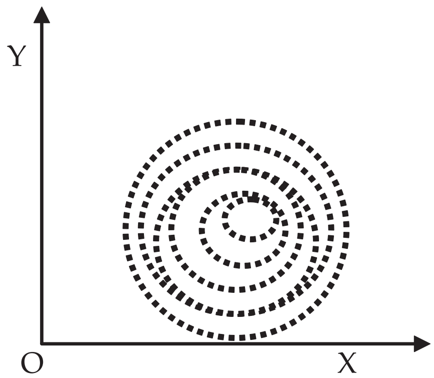

FIGURE 12.5
Showing No Correlation

(r = 0)

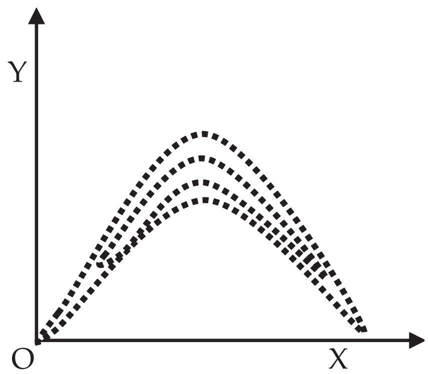

FIGURE 12.6
Showing Curvilinear Correlation
(r = 0)

(b) KARL PEARSON'S PRODUCT MOMENT CORRELATION COEFFICIENT

This is by for the best method for finding correlation between two variables provided the relationship between the two variables is linear. Pearson's correlation coefficient may be defined as the ratio of covariance between the two variables to the product of the standard deviations of the two variables. If the two variables are denoted by x and y and if the corresponding bivariate data are (x_i, y_i) for i = 1, 2, 3, ….., n, then the coefficient of correlation between x and y, due to Karl Pearson, in given by :

$$r = r_{xy} = \frac{\text{Cov}(x, y)}{S_x \times S_y} \qquad \text{.......(12.1)}$$

where

$$\text{cov}(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{\sum x_i y_i}{n} - \bar{x}\,\bar{y} \qquad \text{.......(12.2)}$$

$$S_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}} = \sqrt{\frac{\sum x_i^2}{n} - \bar{x}^2} \qquad \text{.......(12.3)}$$

and $$S_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n}} = \sqrt{\frac{\sum y_i^2}{n} - \bar{y}^2} \qquad \text{.......(12.4)}$$

A single formula for computing correlation coefficient is given by

$$r = \frac{n\sum x_i y_i - \sum x_i \times \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}} \qquad \text{.......(12.5)}$$

In case of a bivariate frequency distribution, we have

$$\text{Cov}(x,y) = \frac{\sum_{i,j} x_i y_i f_{ij}}{N} - \bar{x} \times \bar{y} \qquad \text{.......(12.6)}$$

$$S_x = \sqrt{\frac{\sum_i f_{io} x_i^2}{N} - \bar{x}^2} \qquad \text{.......(12.7)}$$

and $$S_y = \sqrt{\frac{\sum_j f_{oj} y_j^2}{N} - \bar{y}^2} \qquad \text{.......(12.8)}$$

where x_i = Mid-value of the i^{th} class interval of x.

y_j = Mid-value of the j^{th} class interval of y

f_{io} = Marginal frequency of x

f_{oj} = Marginal frequency of y

f_{ij} = frequency of the $(i, j)^{th}$ cell

$$N = \sum_{i,j} f_{ij} = \sum_i f_{io} = \sum_j f_{oj} = \text{Total frequency} \qquad \text{.............. (12.9)}$$

PROPERTIES OF CORRELATION COEFFICIENT

(i) **The Coefficient of Correlation is a unit-free measure.**

This means that if x denotes height of a group of students expressed in cm and y denotes their weight expressed in kg, then the correlation coefficient between height and weight would be free from any unit.

(ii) **The coefficient of correlation remains invariant under a change of origin and/or scale of the variables under consideration depending on the sign of scale factors.**

This property states that if the original pair of variables x and y is changed to a new pair of variables u and v by effecting a change of origin and scale for both x and y i.e.

$$u = \frac{x-a}{b} \text{ and } v = \frac{y-c}{d}$$

where a and c are the origins of x and y and b and d are the respective scales and then we have

$$r_{xy} = \frac{bd}{|b||d|} r_{uv} \qquad \text{...(12.10)}$$

r_{xy} and r_{uv} being the coefficient of correlation between x and y and u and v respectively, (12.10) established, numerically, the two correlation coefficients remain equal and they would have opposite signs only when b and d, the two scales, differ in sign.

(iii) **The coefficient of correlation always lies between –1 and 1, including both the limiting values i.e.**

$$-1 \le r \le 1 \qquad \text{………………… ....(12.11)}$$

Example 12.2 Compute the correlation coefficient between x and y from the following data n = 10, $\sum xy = 220$, $\sum x^2 = 200$, $\sum y^2 = 262$

$\sum x = 40$ and $\sum y = 50$

Solution

From the given data, we have by applying (12.5),

$$r = \frac{n\sum xy - \sum x \times \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} \times \sqrt{n\sum y^2 - (\sum y)^2}}$$

$$= \frac{10 \times 220 - 40 \times 50}{\sqrt{10 \times 200 - (40)^2 \times \sqrt{10 \times 262 - (50)^2}}}$$

$$= \frac{2200 - 2000}{\sqrt{2000 - 1600 \times \sqrt{2620 - 2500}}}$$

$$= \frac{200}{20 \times 10.9545}$$

$$= 0.91$$

Thus there is a good amount of positive correlation between the two variables x and y.

Alternately

As given, $\bar{x} = \frac{\sum x}{n} = \frac{40}{10} = 4$

$$\bar{y} = \frac{\sum y}{n} = \frac{50}{10} = 5$$

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

$$= \frac{220}{10} - 4.5 = 2$$

$$S_x = \sqrt{\frac{\sum x2}{n} - (\bar{x})^2}$$

$$= \sqrt{\frac{200}{10} - 4^2} = 2$$

$$S_y = \sqrt{\frac{\Sigma y_i^2}{n} - \bar{y}^2}$$

$$= \sqrt{\frac{262}{10} \quad 5^2}$$

$$= \sqrt{26.20 - 25} = 1.0954$$

Thus applying formula (12.1), we get

$$r = \frac{\text{cov}(x, y)}{S_x . S_y}$$

$$= \frac{2}{2 \times 1.0954} = 0.91$$

As before, we draw the same conclusion.

Example 12.3 Find product moment correlation coefficient from the following information:

X	:	2	3	5	5	6	8
Y	:	9	8	8	6	5	3

Solution

In order to find the covariance and the two standard deviation, we prepare the following table:

Table 12.3

Computation of Correlation Coefficient

x_i (1)	y_i (2)	$x_i y_i$ (3)= (1) x (2)	x_i^2 (4)= $(1)^2$	y_i^2 (5)= $(2)^2$
2	9	18	4	81
3	8	24	9	64
5	8	40	25	64
5	6	30	25	36
6	5	30	36	25
8	3	24	64	9
29	39	166	163	279

We have

$$\bar{x} = \frac{29}{6} = 4.8333 \quad \bar{y} = \frac{39}{6} = 6.50$$

$$\text{cov}(x, y) = \frac{\sum x_i y_i}{n} - \bar{x}\,\bar{y}$$

$$= 166/6 - 4.8333 \times 6.50 = -3.7498$$

$$= \sqrt{\frac{\Sigma x_i^2}{n} - (\bar{x})^2}$$

$$= \sqrt{\frac{163}{6} \quad (4.8333)^2}$$

$$= \sqrt{27.1667 - 23.3608} = 1.95$$

$$S_y = \sqrt{\frac{\sum y_i^2}{n} - (\bar{y})^2}$$

$$= \sqrt{\frac{279}{6} - (6.50)^2}$$

$$= \sqrt{46.50 - 42.25} = 2.0616$$

Thus the correlation coefficient between x and y in given by

$$r = \frac{\text{cov}(x, y)}{S_x \times s_y}$$

$$= \frac{-3.7498}{1.9509 \times 2.0616}$$

$$= -0.93$$

We find a high degree of negative correlation between x and y. Also, we could have applied formula (12.5) as we have done for the first problem of computing correlation coefficient.

Sometimes, a change of origin reduces the computational labor to a great extent. This we are going to do in the next problem.

Example 12.4 The following data relate to the test scores obtained by eight salesmen in an aptitude test and their daily sales in thousands of rupees:

Salesman :	1	2	3	4	5	6	7	8
scores :	60	55	62	56	62	64	70	54
Sales :	31	28	26	24	30	35	28	24

Solution

Let the scores and sales be denoted by x and y respectively. We take a, origin of x as the average of the two extreme values i.e. 54 and 70. Hence a = 62 similarly, the origin of y is taken

as $b = \frac{24+35}{2} \cong 30$

Table 12.4

Computation of Correlation Coefficient Between Test Scores and Sales.

Scores (x_i) (1)	Sales in Rs. 1000 (y_i) (2)	$u_i = x_i - 62$ (3)	$v_i = y_i - 30$ (4)	u_iv_i (5)=(3)x(4)	u_i^2 (6)=(3)^2	v_i^2 (7)=(4)^2
60	31	–2	1	–2	4	1
55	28	–7	–2	14	49	4
62	26	0	–4	0	0	16
56	24	–6	–6	36	36	36
62	30	0	0	0	0	0
64	35	2	5	10	4	25
70	28	8	–2	–16	64	4
54	24	–8	–6	48	64	36
Total	—	–13	–14	90	221	122

Since correlation coefficient remains unchanged due to change of origin, we have

$$r = r_{xy} = r_{uv} = \frac{n\sum u_iv_i - \sum u_i \times \sum v_i}{\sqrt{n\sum u_i^2 - \left(\sum u_i\right)^2} \times \sqrt{n\sum v_i^2 - \left(\sum v_i\right)^2}}$$

$$= \frac{8\times 90 - (-13)\times(-14)}{\sqrt{8\times 221 - (-13)^2} \times \sqrt{8\times 122 - (-14)^2}}$$

$$= \frac{538}{\sqrt{1768 - 169} \times \sqrt{976 - 196}}$$

$$= 0.48$$

In some cases, there may be some confusion about selecting the pair of variables for which correlation is wanted. This is explained in the following problem.

Example 12.5 Examine whether there is any correlation between age and blindness on the basis of the following data:

Age in years :	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of Persons (in thousands) :	90	120	140	100	80	60	40	20
No. of blind Persons :	10	15	18	20	15	12	10	06

Solution

Let us denote the mid-value of age in years as x and the number of blind persons per lakh as y. Then as before, we compute correlation coefficient between x and y.

Table 12.5

Computation of correlation between age and blindness

Age in years (1)	Mid-value x (2)	No. of Persons ('000) P (3)	No. of blind B (4)	No. of blind per lakh y=B/P × 1 lakh (5)	xy (2)×(5) (6)	x^2 $(2)^2$ (7)	y^2 $(5)^2$ (8)
0-10	5	90	10	11	55	25	121
10-20	15	120	15	12	180	225	144
20-30	25	140	18	13	325	625	169
30-40	35	100	20	20	700	1225	400
40-50	45	80	15	19	855	2025	361
50-60	55	60	12	20	1100	3025	400
60-70	65	40	10	25	1625	4225	625
70-80	75	20	6	30	2250	5625	900
Total	320	—	—	150	7090	17000	3120

The correlation coefficient between age and blindness is given by

$$r = \frac{n\sum xy - \sum x.\sum y}{\sqrt{n\sum x^2 - (\sum x)^2} \times \sqrt{n\sum y^2 - (\sum y)^2}}$$

$$= \frac{8.7090 - 320.150}{\sqrt{8.17000 - (320)^2} \times \sqrt{8.3120 - (150)^2}}$$

$$= \frac{8720}{183.3030.49.5984}$$

$$= 0.96$$

which exhibits a very high degree of positive correlation between age and blindness.

Example 12.6 Coefficient of correlation between x and y for 20 items is 0.4. The AM's and SD's of x and y are known to be 12 and 15 and 3 and 4 respectively. Later on, it was found that the pair (20, 15) was wrongly taken as (15, 20). Find the correct value of the correlation coefficient.

Solution

We are given that n = 20 and the original r = 0.4, $\bar{x}$ = 12, $\bar{y}$ = 15, S_x = 3 and S_y = 4

$$r = \frac{\text{cov}(x, y)}{S_x \times S_y} \qquad = 0.4 = \frac{\text{cov}(x, y)}{3 \times 4}$$

$$= \text{Cov}(x, y) = 4.8$$

$$= \frac{\sum xy}{n} - \bar{x}\,\bar{y} = 4.8$$

$$= \frac{\sum xy}{20} - 12 \times 15 = 4.8$$

$$= \sum xy = 3696$$

Hence, corrected $\sum xy$= 3696 – 20 × 15 + 15 × 20 = 3696

Also, $S_x^2 = 9$

$= (\sum x^2 / 20) - 12^2 = 9$

$\sum x^2 = 3060$

Similarly, $S_y^2 = 16$

$$S_y^2 = \frac{\sum y^2}{20} - 15^2 = 16$$

$\sum y^2 = 4820$

Thus corrected $\sum x = n\bar{x}$ – wrong x value + correct x value.

$$= 20 \times 12 - 15 + 20$$

$$= 245$$

Similarly corrected $\sum y = 20 \times 15 - 20 + 15 = 295$

Corrected $\sum x^2 = 3060 - 15^2 + 20^2 = 3235$

Corrected $\sum y^2 = 4820 - 20^2 + 15^2 = 4645$

Thus corrected value of the correlation coefficient by applying formula (12.5)

$$= \frac{20.3696 - 245.295}{\sqrt{20.3235 - (245)^2} \times \sqrt{20.4645 - (295)^2}}$$

$$= \frac{73920 - 72275}{68.3740 \times 76.6480}$$

$$= 0.31$$

Example 12.7 Compute the coefficient of correlation between marks in Statistics and Mathematics for the bivariate frequency distribution shown in table 12.1

Solution

For the sake of computational advantage, we effect a change of origin and scale for both the variable x and y.

Define $u_i = \frac{x_i - a}{b} = \frac{x_i - 10}{4}$

And $v_j = \frac{y_i - c}{d} = \frac{y_i - 10}{4}$

Where x_i and y_j denote respectively the mid-values of the x-class interval and y-class interval respectively. The following table shows the necessary calculation on the right top corner of each cell, the product of the cell frequency, corresponding u value and the respective v value has been shown. They add up in a particular row or column to provide the value of $f_{ij}u_iv_j$ for that particular row or column.

Table 12.6

Computation of Correlation Coefficient Between Marks of Mathematics and Statistics

Class Interval	Mid -value	Class Interval / Mid-value → v_j; u_i ↓	0-4	4-8	8-12	12-16	16-20	f_{io}	$f_{io}u_i$	$f_{io}u_i^2$	$f_{ij}u_iv_j$
		Mid-value	2	6	10	14	18				
		v_j / u_i	–2	–1	0	1	2				
0-4	2	–2	1 (4)	1 (2)	2 (0)			4	–8	16	6
4-8	6	–1	2 (4)	4 (4)	5 (0)	1 (–1)	1 (–2)	13	–13	13	5
8-12	10	0		2 (0)	4 (0)	6 (0)	1 (0)	13	0	0	0
12-16	14	1		1 (–1)	3 (0)	2 (2)	5 (10)	11	11	11	11
16-20	18	2			1 (0)	5 (10)	3 (12)	9	18	36	22
		f_{oj}	3	8	15	14	10	50	5	76	44
		$f_{oj}v_j$	–6	–8	0	14	20	20			
		$f_{oj}v_j^2$	12	8	0	14	40	74			
		$f_{ij}u_iv_j$	8	5	0	11	20	44		CHECK	

A single formula for computing correlation coefficient from bivariate frequency distribution is given by

$$r = \frac{N\sum_{i,j} f_{ij}u_iv_j - \sum f_{io}u_i \times \sum f_{oj}v_j}{\sqrt{N\sum f_{io}u_i^2 - \left(\sum f_{io}u_i\right)^2 \times \sum f_{oj}v_j^2 - \left(\sum f_{oj}v_j\right)^2}} \quad \text{...........................(12.10)}$$

$$= \frac{50 \times 44 - 8 \times 20}{\sqrt{50 \times 76 - 8^2}\sqrt{50 \times 74 - 20^2}}$$

$$= \frac{2040}{61.1228 \times 57.4456}$$

$$= 0.58$$

The value of r shown a good amount of positive correlation between the marks in Statistics and Mathematics on the basis of the given data.

Example 12.8 Given that the correlation coefficient between x and y is 0.8, write down the correlation coefficient between u and v where

(i) 2u + 3x + 4 = 0 and 4v + 16x + 11 = 0

(ii) 2u – 3x + 4 = 0 and 4v + 16x + 11 = 0

(iii) 2u – 3x + 4 = 0 and 4v – 16x + 11 = 0

(iv) 2u + 3x + 4 = 0 and 4v – 16x + 11 = 0

Solution

Using (12.10), we find that

$$r_{xy} = \frac{bd}{|b||d|} r_{uv}$$

i.e. $r_{xy} = r_{uv}$ if b and d are of same sign and $r_{uv} = -r_{xy}$ when b and d are of opposite signs, b and d being the scales of x and y respectively. In (i), u = (–2) + (-3/2) x and v = (–11/4) + (–4)y.

Since b = –3/2 and d = –4 are of same sign, the correlation coefficient between u and v would be the same as that between x and y i.e. $r_{xy} = 0.8 = r_{uv}$

In (ii), u = (–2) + (3/2)x and v = (–11/4) + (–4)y Hence b = 3/2 and d = –4 are of opposite signs and we have $r_{uv} = -r_{xy} = -0.8$

Proceeding in a similar manner, we have r_{uv} = 0.8 and – 0.8 in (iii) and (iv).

(c) SPEARMAN'S RANK CORRELATION COEFFICIENT

When we need finding correlation between two qualitative characteristics, say, beauty and intelligence, we take recourse to using rank correlation coefficient. Rank correlation can also be applied to find the level of agreement (or disagreement) between two judges so far as assessing a qualitative characteristic is concerned. As compared to product moment correlation coefficient, rank correlation coefficient is easier to compute, it can also be advocated to get a first hand impression about the correlation between a pair of variables.

Spearman's rank correlation coefficient is given by

$$r_R = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad \text{...} \quad (12.11)$$

where r_R denotes rank correlation coefficient and it lies between –1 and 1 inclusive of these two values.

$d_i = x_i - y_i$ represents the difference in ranks for the i-th individual and n denotes the number of individuals.

In case u individuals receive the same rank, we describe it as a tied rank of length u. In case of a tied rank, formula (12.11) is changed to

$$r_R = 1 - \frac{6\left[\sum_i d_i^2 + \sum_j \frac{(tj^3 - t_j)}{12}\right]}{n(n^2-1)} \quad \text{...} \quad (12.12)$$

In this formula, t_j represents the j^{th} tie length and the summation $\sum_j (t_j^3 - t_j)$ extends over the lengths of all the ties for both the series.

Example 12.9 compute the coefficient of rank correlation between sales and advertisement expressed in thousands of rupees from the following data:

Sales :	90	85	68	75	82	80	95	70
Advertisement :	7	6	2	3	4	5	8	1

Solution

Let the rank given to sales be denoted by x and rank of advertisement be denoted by y. We note that since the highest sales as given in the data, is 95, it is to be given rank 1, the second highest sales 90 is to be given rank 2 and finally rank 8 goes to the lowest sales, namely 68. We have given rank to the other variable advertisement in a similar manner. Since there are no ties, we apply formula (12.11).

Table 12.7

Computation of Rank correlation between Sales and Advertisement.

Sales (x_i)	Advertisement (y_i)	Rank for Sales (x_i)	Rank for Advertisement (y_i)	$d_i = x_i - y_i$	d_i^2
90	7	2	2	0	0
85	6	3	3	0	0
68	2	8	7	1	1
75	3	6	6	0	0
82	4	4	5	–1	1
80	5	5	4	1	1
95	8	1	1	0	0
70	1	7	8	–1	1
Total	—	—	—	0	4

Since n = 8 and $\sum d_i^2 = 4$, applying formula (12.11), we get.

$$r_R = 1 \quad \frac{6\sum d_i^2}{n(n^2 \quad 1)}$$

$$= 1 - \frac{6 \times 4}{8(8^2 - 1)}$$

$$= 1 – 0.0476$$

$$= 0.95$$

The high positive value of the rank correlation coefficient indicates that there is a very good amount of agreement between sales and advertisement.

Example 12.10 Compute rank correlation from the following data relating to ranks given by two judges in a contest:

Serial No. of Candidate :	1	2	3	4	5	6	7	8	9	10
Rank by Judge A :	10	5	6	1	2	3	4	7	9	8
Rank by Judge B :	5	6	9	2	8	7	3	4	10	1

Solution

We directly apply formula (12.11) as ranks are already given.

Table 12.8

Computation of Rank Correlation Coefficient between the ranks given by 2 Judges

Serial No.	Rank by A (x_i)	Rank by B (y_i)	$d_i = x_i - y_i$	d_i^2
1	10	5	5	25
2	5	6	–1	1
3	6	9	–3	9
4	1	2	–1	1
5	2	8	–6	36
6	3	7	–4	16
7	4	3	1	1
8	7	4	3	9
9	8	10	–2	4
10	9	1	8	64
Total	—	—	0	166

The rank correlation coefficient is given by

$$r_R = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 166}{10(10^2-1)}$$

$$= -0.006$$

The very low value (almost 0) indicates that there is hardly any agreement between the ranks given by the two Judges in the contest.

Example 12 .11 Compute the coefficient of rank correlation between Eco. marks and stats. Marks as given below:

Eco Marks :	80	56	50	48	50	62	60
Stats Marks :	90	75	75	65	65	50	65

Solution

This is a case of tied ranks as more than one student share the same mark both for Economics and Statistics. For Eco. the student receiving 80 marks gets rank 1 one getting 62 marks receives rank 2, the student with 60 receives rank 3, student with 56 marks gets rank 4 and since there are two students, each getting 50 marks, each would be receiving a common rank, the average of the next two ranks 5 and 6 i.e. $\frac{5+6}{2}$ i.e. 5.50 and lastly the last rank..

7 goes to the student getting the lowest Eco marks. In a similar manner, we award ranks to the students with stats marks.

Table 12.9

Computation of Rank Correlation Between Eco Marks and Stats Marks with Tied Marks

Eco Mark (x_i)	Stats Mark (y_i)	Rank for Eco (x_i)	Rank for Stats (y_i)	$d_i = x_i - y_i$	d_i^2
80	90	1	1	0	0
56	75	4	2.50	1.50	2.25
50	75	5.50	2.50	3	9
48	65	7	5	2	4
50	65	5.50	5	0.50	0.25
62	50	2	7	–5	25
60	65	3	5	–2	4
Total	—	—	—	0	44.50

For Economics mark there is one tie of length 2 and for stats mark, there are two ties of lengths 2 and 3 respectively.

$$\text{Thus } \frac{\Sigma\left(t_j^3 - t_j\right)}{12} = \frac{\left(2^3-2\right)+\left(2^3-2\right)+\left(3^3-3\right)}{12} = 3$$

$$\text{Thus } r_R = 1 - \frac{6\left[\sum_i d_i^2 + \sum_j \frac{\left(tj^3 - t_j\right)}{12}\right]}{n\left(n^2-1\right)}$$

$$= 1 - \frac{6 \times (44.50 + 3)}{7(7^2 - 1)}$$

$$= 0.15$$

Example 12.12 For a group of 8 students, the sum of squares of differences in ranks for Mathematics and Statistics marks was found to be 50 what is the value of rank correlation coefficient?

Solution

As given n = 8 and $\sum d_i^2$ = 50. Hence the rank correlation coefficient between marks in Mathematics and Statistics is given by

$$r_R = 1 - \frac{6\sum d_i^2}{n\left(n^2-1\right)}$$

$$= 1 - \frac{6 \times 50}{8(8^2-1)}$$

$$= 0.40$$

Example 12.13 For a number of towns, the coefficient of rank correlation between the people living below the poverty line and increase of population is 0.50. If the sum of squares of the differences in ranks awarded to these factors is 82.50, find the number of towns.

Solution

As given $r_R = 0.50$, $\sum d_i^2$ = 82.50.

$$\text{Thus } r_R = 1 - \frac{6\sum d_i^2}{n\left(n^2-1\right)}$$

$$0.50 = 1 - \frac{6 \times 82.50}{n(n^2 - 1)}$$

$$= n(n^2 - 1) = 990$$

$$= n(n^2 - 1) = 10(10^2 - 1)$$

$\therefore$ n = 10 as n must be a positive integer.

Example 12.14 While computing rank correlation coefficient between profits and investment for 10 years of a firm, the difference in rank for a year was taken as 7 instead of 5 by mistake and the value of rank correlation coefficient was computed as 0.80. What would be the correct value of rank correlation coefficient after rectifying the mistake?

Solution:

We are given that n = 10,

r_R = 0.80 and the wrong d_i = 7 should be replaced by 5.

$$r_R = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

$$0.80 = 1 - \frac{6 \sum d_i^2}{10(10^2 - 1)}$$

$$\sum d_i^2 = 33$$

Corrected $\sum d_i^2 = 33 - 7^2 + 5^2 = 9$

Hence rectified value of rank correlation coefficient

$$= 1 - \frac{6 \times 9}{10 \times (10^2 - 1)}$$

$$= 0.95$$

(d) COEFFICIENT OF CONCURRENT DEVIATIONS

A very simple and casual method of finding correlation when we are not serious about the magnitude of the two variables is the application of concurrent deviations. This method involves in attaching a positive sign for a x-value (except the first) if this value is more than the previous value and assigning a negative value if this value is less than the previous value. This is done for the y-series as well. The deviation in the x-value and the corresponding y-value is known to be concurrent if both the deviations have the same sign.

Denoting the number of concurrent deviation by c and total number of deviations as m (which must be one less than the number of pairs of x and y values), the coefficient of concurrent deviation is given by

$$r_C = \pm\sqrt{\pm\frac{(2c-m)}{m}} \quad \text{...(12.13)}$$

If (2c–m) >0, then we take the positive sign both inside and outside the radical sign and if (2c–m) <0, we are to consider the negative sign both inside and outside the radical sign.

Like Pearson's correlation coefficient and Spearman's rank correlation coefficient, the coefficient of concurrent deviations also lies between –1 and 1, both inclusive.

Example 12.15 Find the coefficient of concurrent deviations from the following data.

Year :	1990	1991	1992	1993	1994	1995	1996	1997
Price :	25	28	30	23	35	38	39	42
Demand :	35	34	35	30	29	28	26	23

Table 12.10

Solution:

Computation of Coefficient of Concurrent Deviations.

Year	Price	Sign of deviation from the previous figure (a)	Demand	Sign of deviation from the previous figure (b)	Product of deviation (ab)
1990	25		35		
1991	28	+	34	–	–
1992	30	+	35	+	+
1993	23	–	30	–	+
1994	35	+	29	–	–
1995	38	+	28	–	–
1996	39	+	26	–	–
1997	42	+	23	–	–

In this case, m = number of pairs of deviations = 7

c = No. of positive signs in the product of deviation column = Number of concurrent deviations = 2

Thus r_C $= \pm\sqrt{\pm\frac{(2c-m)}{m}}$

$= \pm\sqrt{\pm\frac{(4-7)}{m}}$

$= \pm\sqrt{\pm\frac{(-3)}{7}}$

$= -\sqrt{\frac{3}{7}} = -0.65$

(Since $\frac{2c \quad m}{m} = \frac{3}{7}$ we take negative sign both inside and outside of the radical sign)

Thus there is a negative correlation between price and demand.

12.5 REGRESSION ANALYSIS

In regression analysis, we are concerned with the estimation of one variable for a given value of another variable (or for a given set of values of a number of variables) on the basis of an average mathematical relationship between the two variables (or a number of variables). Regression analysis plays a very important role in the field of every human activity. A businessman may be keen to know what would be his estimated profit for a given level of investment on the basis of the past records. Similarly, an outgoing student may like to know her chance of getting a first class in the final University Examination on the basis of her performance in the college selection test.

When there are two variables x and y and if y is influenced by x i.e. if y depends on x, then we get a simple linear regression or simple regression. y is known as dependent variable or regression or explained variable and x is known as independent variable or predictor or explanator. In the previous examples since profit depends on investment or performance in the University Examination is dependent on the performance in the college selection test, profit or performance in the University Examination is the dependent variable and investment or performance in the selection test is the In-dependent variable.

In case of a simple regression model if y depends on x, then the regression line of y on x in given by

$y = a + bx$ …………………… (12.14)

Here a and b are two constants and they are also known as regression parameters. Furthermore, b is also known as the regression coefficient of y on x and is also denoted by b_{yx}. We may define the regression line of y on x as the line of best fit obtained by the method of least squares and used for estimating the value of the dependent variable y for a known value of the independent variable x.

The method of least squares involves in minimizing

$$\Sigma e_i^2 = \Sigma (y_i^2 - y_i)^2 = \Sigma (y_i - a - bx_i)^2 \;\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \;(12.15)$$

where y_i demotes the actual or observed value and $y_i = a + b_{xi}$, the estimated value of y_i for a given value of x_i, e_i is the difference between the observed value and the estimated value and e_i is technically known as error or residue. This summation intends over n pairs of observations of $(x_{i,} y_i)$. The line of regression of y or x and the errors of estimation are shown in the following figure.

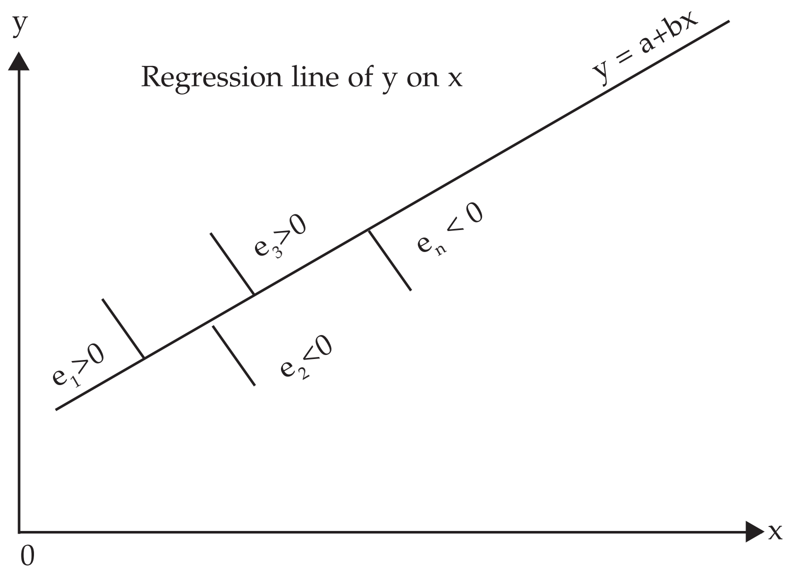

FIGURE 12.7

SHOWING REGRESSION LINE OF y ON x

AND ERRORS OF ESTIMATION

Minimisation of (12.15) yields the following equations known as 'Normal Equations'

. $\Sigma y_i = na + b\Sigma x_i$ ……………….. (12.16)

$\Sigma x_i y_i = a\Sigma x_i + b\Sigma x_i^2$ ……………….. (12.17)

Solving there two equations for b and a, we have the "least squares" estimates of b and a as

$$b = \frac{Cov(x, y)}{S_x^{\,2}}$$

$$= \frac{r.S_x.S_y}{S_x^2}$$

$$= \frac{r.S_y}{S_x} \quad \text{..(12.18)}$$

After estimating b, estimate of a is given by

$$a = \bar{y} - b\bar{x} \quad \text{.......................................(12.19)}$$

Substituting the estimates of b and a in (12.14), we get

$$\frac{(y - \bar{y})}{S_y} = \frac{r(x - \bar{x})}{S_x} \quad \text{..(12.20)}$$

There may be cases when the variable x depends on y and we may take the regression line of x on y as

$x = a\hat{} + b\hat{}y$

Unlike the minimization of vertical distances in the scatter diagram as shown in figure (12.7) for obtaining the estimates of a and b, in this case we minimize the horizontal distances and get the following normal equation in a^ and b^, the two regression parameters :

$$\Sigma x_i = na\hat{} + b\hat{}\Sigma y_i \quad \text{.......................................(12.21)}$$

$$\Sigma x_i y_i = a\hat{}\Sigma y_i + b\hat{}\Sigma y_i^2 \quad \text{.......................................(12.22)}$$

or solving these equations, we get

$$b\hat{} = b_{xy} = \frac{\text{cov}(x, y)}{S_y^2} = \frac{r.S_x}{S_y} \quad \text{..........................(12.23)}$$

$$\text{and } a\hat{} = \bar{x} - b\hat{}\,\bar{y} \quad \text{.......................................(12.24)}$$

A single formula for estimating b is given by

$$b\hat{} = b_{yx} = \frac{n\Sigma x_i y_i - \Sigma x_i.\Sigma y_i}{n\Sigma y_i^2 - (\Sigma y_i)^2} \quad \text{....................(12.25)}$$

$$\text{Similarly, } b\hat{} = b_{yx} = \frac{n\Sigma x_i y_i - \Sigma x_i.\Sigma y_i}{n\Sigma y_i^2 - (\Sigma y_i)^2} \quad \text{...........(12.26)}$$

The standardized form of the regression equation of x on y, as in (12.20), is given by

$$\frac{x-\bar{x}}{S_x} = r\frac{(y-\bar{y})}{S_y} \quad \text{……………………................} \quad (12.27)$$

Example 12.15 Find the two regression equations from the following data:

x:	2	4	5	5	8	10
y:	6	7	9	10	12	12

Hence estimate y when x is 13 and estimate also x when y is 15.

Solution

Table 12.11

Computation of Regression Equations

x_i	y_i	$x_i\, y_i$	x_i^2	y_i^2
2	6	12	4	36
4	7	28	16	49
5	9	45	25	81
5	10	50	25	100
8	12	96	64	144
10	12	120	100	144
34	56	351	234	554

On the basis of the above table, we have

$$\bar{x} = \frac{\sum x_i}{n} = \frac{34}{6} = 5.6667$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{56}{6} = 9.3333$$

$$\text{cov}\,(x, y) = \frac{x_i y_i}{n} \quad \bar{x}\,\bar{y}$$

$$= \frac{351}{6} - 5.6667 \times 9.3333$$

$$= 58.50 – 52.8890$$

$$= 5.6110$$

$$S_x^{\,2} = \frac{\sum x_i^2}{n} - (\bar{x})^2$$

$$= \frac{234}{6} - (5.6667)^2$$

$= 39 - 32.1115$

$= 6.8885$

$$S_y^{\,2} = \frac{\sum y_i^2}{n} - \left(\bar{y}\right)^2$$

$$= \frac{554}{6} - (9.3333)^2$$

$= 92.3333 - 87.1105$

$= 5.2228$

The regression line of y on x is given by

$y = a + bx$

$$\text{Where } \hat{b} = \frac{\text{cov}(x, y)}{S_x^2}$$

$$= \frac{5.6110}{6.8885}$$

$= 0.8145$

and $\hat{a} = \bar{y} - b\bar{x}$

$= 9.3333 - 0.8145 \times 5.6667$

$= 4.7178$

Thus the estimated regression equation of y on x is

$y = 4.7178 + 0.8145x$

When x = 13, the estimated value of y is given by $\hat{y} = 4.7178 + 0.8145 \times 13 = 15.3063$

The regression line of x on y is given by

$x = \hat{a} + \hat{b}\, y$

$$\text{Where } \hat{b} = \frac{\text{cov}\ x, y}{S_y^{\,2}}$$

$$= \frac{5.6110}{5.2228}$$

$= 1.0743$

and $\hat{a} = \bar{x} - b\,\bar{y}$

$= 5.6667 - 1.0743 \times 9.3333$

$= -4.3601$

Thus the estimated regression line of x on y is

$x = -4.3601 + 1.0743y$

When y = 15, the estimate value of x is given by

$\hat{x} = -4.3601 + 1.0743 \times 15$

$= 11.75$

Example 12.16 Marks of 8 students in Mathematics and statistics are given as:

Mathematics:	80	75	76	69	70	85	72	68
Statistics:	85	65	72	68	67	88	80	70

Find the regression lines. When marks of a student in Mathematics are 90, what are his most likely marks in statistics?

Solution

We denote the marks in Mathematics and Statistics by x and y respectively. We are to find the regression equation of y on x and also of x or y. Lastly, we are to estimate y when x = 90. For computation advantage, we shift origins of both x and y.

Table 12.12

Computation of regression lines

Maths mark (x_i)	Stats mark (y_i)	$u_i = x_i - 74$	$v_i = y_i - 76$	$u_i v_i$	u_i^2	v_i^2
80	85	6	9	54	36	81
75	65	1	–11	–11	1	121
76	72	2	–4	–8	4	16
69	68	–5	–8	40	25	64
70	67	–4	–9	36	16	81
85	88	11	12	132	121	144
72	80	–2	4	–8	4	16
68	70	–6	–6	36	36	36
595	595	3	–13	271	243	559

The regression coefficients b (or b_{yx}) and b′ (or b_{xy}) remain unchanged due to a shift of origin.

Applying (12.25) and (12.26), we get

$$b = b_{yx} = b_{vu} = \frac{n\sum u_i v_i - \sum u_i . \sum v_i}{n\sum u_i^2 - (\sum u_i)^2}$$

$$= \frac{8.(271) - (3).(-13)}{8.(243) - (3)^2}$$

$$= \frac{2168 + 39}{1944 - 9}$$

$$= 1.1406$$

$$\text{and } b^{\wedge} = b_{xy} = b_{uv} = \frac{n\sum u_i v_i - \sum u_i . \sum v_i}{n\sum v_i^2 - (\sum v_i)^2}$$

$$= \frac{8.(271) - (3).(-13)}{8.(559) - (-13)^2}$$

$$= \frac{2168 + 39}{4472 - 169}$$

$$= 0.5129$$

Also $a^{\wedge} = \bar{y} - b^{\wedge}\bar{x}$

$$= \frac{(595)}{8} - 1.1406 \frac{(595)}{8}$$

$= 74.375 - 1.1406 \times 74.375$

$= -10.4571$

and $a^{\wedge} = \bar{x} - b^{\wedge}\bar{y}$

$= 74.375 - 0.5129 \times 74.375$

$= 36.2280$

The regression line of y on x is

$y = -10.4571 + 1.1406x$

and the regression line of x on y is

$x = 36.2281 + 0.5129y$

For x = 90, the most likely value of y is

$$\hat{y} = -10.4571 + 1.1406 \times 90$$
$$= 92.1969$$
$$\cong 92$$

Example 12.17 The following data relate to the mean and SD of the prices of two shares in a stock Exchange:

Share	Mean (in Rs.)	SD (in Rs.)
Company A	44	5.60
Company B	58	6.30

Coefficient of correlation between the share prices = 0.48

Find the most likely price of share A corresponding to a price of Rs. 60 of share B and also the most likely price of share B for a price of Rs. 50 of share A.

Solution

Denoting the share prices of Company A and B respectively by x and y, we are given that

$\bar{x}$ = Rs. 44, $\bar{y}$ = Rs. 58

S_x = Rs. 5.60, S_y = Rs. 6.30

and r = 0.48

The regression line of y on x is given by

$$y = a + bx$$

Where
$$b = r \times \frac{S_y}{S_x}$$
$$= 0.48 \times \frac{6.30}{5.60}$$
$$= 0.54$$
$$a = \bar{y} - b\bar{x}$$
$$= \text{Rs. } (58 - 0.54 \times 44)$$
$$= \text{Rs. } 34.24$$

Thus the regression line of y on x i.e. the regression line of price of share B on that of share A is given by

$$y = \text{Rs. } (34.24 + 0.54x)$$

When x = Rs. 50, = Rs. $(34.24 + 0.54 \times 50)$

$= \text{Rs. } 61.24$

= The estimated price of share B for a price of Rs. 50 of share A is Rs. 61.24

Again the regression line of x on y is given by

$x = a\hat{} + b\hat{}y$

Where $b\hat{} = r \times \frac{S_x}{S_y}$

$= 0.48 \times \frac{5.60}{6.30}$

$= 0.4267$

$a\hat{} = \bar{x} - b\hat{}\,\bar{y}$

$= \text{Rs. } (44 - 0.4267 \times 58)$

$= \text{Rs. } 19.25$

Hence the regression line of x on y i.e. the regression line of price of share A on that of share B in given by

$x = \text{Rs. } (19.25 + 0.4267y)$

When y = Rs. 60, $\hat{x} = \text{Rs. } (19.25 + 0.4267 \times 60)$

$= \text{Rs. } 44.85$

Example 12.18 The following data relate the expenditure or advertisement in thousands of rupees and the corresponding sales in lakhs of rupees.

Expenditure on Ad :	8	10	10	12	15
Sales :	18	20	22	25	28

Find an appropriate regression equation.

Solution

Since sales (y) depend on advertisement (x), the appropriate regression equation is of y on x i.e. of sales on advertisement. We have, on the basis of the given data,

$n = 5,\ \sum x = 8+10+10+12+15 = 55$

$\sum y = 18+20+22+25+28 = 113$

$\sum xy = 8\times18+10\times20+10\times22+12\times25+15\times28 = 1284$

$\sum x^2 = 8^2+10^2+10^2+12^2+15^2 = 633$

$$\therefore b = \frac{n\sum xy - \sum x \times \sum y}{n\sum x^2 - (\sum x)^2}$$

$$= \frac{5\times1284-55\times113}{5\times633-(55)^2}$$

$$= \frac{205}{140}$$

$$= 1.4643$$

$$a = \bar{y}-b\bar{x}$$

$$= \frac{113}{5}-1.4643\times\frac{55}{5}$$

$$= 22.60 - 16.1073$$

$$= 6.4927$$

Thus, the regression line of y or x i.e. the regression line of sales on advertisement is given by

$$y = 6.4927 + 1.4643x$$

12.6 PROPERTIES OF REGRESSION LINES

We consider the following important properties of regression lines:

(i) **The regression coefficients remain unchanged due to a shift of origin but change due to a shift of scale.**

This property states that if the original pair of variables is (x, y) and if they are changed to the pair (u, v) where

$$u=\frac{x\quad a}{p} \text{ and } v=\frac{y\quad c}{q}$$

$$b_{yx} = \frac{q}{p}\times b_{vu} \quad\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(12.28)}$$

$$\text{and } bxy = \frac{p}{q}\times b_{uv} \quad\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(12.29)}$$

(ii) **The two lines of regression intersect at the point $\bar{x},\bar{y}$, where x and y are the variables under consideration.**

According to this property, the point of intersection of the regression line of y on x and the regression line of x on y is $\bar{x},\bar{y}$ i.e. the solution of the simultaneous equations in x and y.

(iii) **The coefficient of correlation between two variables x and y in the simple geometric**

mean of the two regression coefficients. The sign of the correlation coefficient would be the common sign of the two regression coefficients.

This property says that if the two regression coefficients are denoted by b_{yx} (=b) and b_{xy} (=b') then the coefficient of correlation is given by

$$r = \pm \sqrt{b_{yx} \times b_{xy}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(12.30)}$$

If both the regression coefficients are negative, r would be negative and if both are positive, r would assume a positive value.

Example 12.19 If the relationship between two variables x and u is u + 3x = 10 and between two other variables y and v is 2y + 5v = 25, and the regression coefficient of y on x is known as 0.80, what would be the regression coefficient of v on u?

Solution

u + 3x = 10

$$u = \frac{(x - 10/3)}{-1/3}$$

and 2y + 5v = 25

$$\Rightarrow \quad v = \frac{(y - 25/2)}{-5/2}$$

From (12.28), we have

$$b_{yx} = \frac{q}{p} \times b_{vu}$$

or,

$$0.80 = \frac{-5/2}{-1/3} \times b_{vu}$$

$$\Rightarrow \quad 0.80 = \frac{15}{2} \times b_{vu}$$

$$\Rightarrow \quad b_{vu} = \frac{2}{15} \times 0.80 = \frac{8}{75}$$

Example 12.20 For the variables x and y, the regression equations are given as 7x – 3y – 18 = 0 and 4x – y – 11 = 0

(i) Find the arithmetic means of x and y.

(ii) Identify the regression equation of y on x.

(iii) Compute the correlation coefficient between x and y.

(iv) Given the variance of x is 9, find the SD of y.

Solution

(i) Since the two lines of regression intersect at the point $(\bar{x}, \bar{y})$, replacing x and y by $\bar{x}$ and $\bar{y}$ respectively in the given regression equations, we get

$$7\bar{x}-3\bar{y}-18=0$$

and $$4\bar{x}-\bar{y}\quad 11=0$$

Solving these two equations, we get $\bar{x}$ = 3 and $\bar{y}$ = 1

Thus the arithmetic means of x and y are given by 3 and 1 respectively.

(ii) Let us assume that 7x – 3y – 18 = 0 represents the regression line of y on x and 4x – y – 11 = 0 represents the regression line of x on y.

Now 7x – 3y – 18 = 0

$$\Rightarrow \qquad y=(-6)+\frac{(7)}{3}x$$

$$\therefore \qquad b_{yx}=\frac{7}{3}$$

Again 4x – y – 11 = 0

$$\Rightarrow \quad x=\frac{(11)}{4}+\frac{(1)}{4}y \qquad \therefore b_{xy}=\frac{1}{4}$$

Thus $r^2 = b_{yx} \times b_{xy}$

$$=\frac{7}{3}\times\frac{1}{4}$$

$$=\frac{7}{12}<1$$

Since $|r| \le 1 \Rightarrow r^2 \le 1$, our assumptions are correct. Thus, 7x – 3y – 18 = 0 truly represents the regression line of y on x.

(iii) Since $r^2 = \frac{7}{12}$

$\therefore \quad r = \sqrt{\frac{7}{12}}$ (We take the sign of r as positive since both the regression coefficients are positive)

$= 0.7638$

(iv) $b_{yx} = r \times \frac{S_y}{S_x}$

$\Rightarrow \frac{7}{3} = 0.7638 \times \frac{S_y}{3} \quad (\therefore S_x^2 = 9 \text{ as given})$

$\Rightarrow S_y = \frac{7}{0.7638}$

$= 9.1647$

12.7 REVIEW OF CORRELATION AND REGRESSION ANALYSIS

So far we have discussed the different measures of correlation and also how to fit regression lines applying the method of 'Least Squares'. It is obvious that we take recourse to correlation analysis when we are keen to know whether two variables under study are associated or correlated and if correlated, what is the strength of correlation. The best measure of correlation is provided by Pearson's correlation coefficient. However, one severe limitation of this correlation coefficient, as we have already discussed, is that it is applicable only in case of a linear relationship between the two variables.

If two variables x and y are independent or uncorrelated then obviously the correlation coefficient between x and y is zero. However, the converse of this statement is not necessarily true i.e. if the correlation coefficient, due to Pearson, between two variables comes out to be zero, then we cannot conclude that the two variables are independent. All that we can conclude is that no linear relationship exists between the two variables. This, however, does not rule out the existence of some non linear relationship between the two variables. For example, if we consider the following pairs of values on two variables x and y.

(–2, 4), (–1, 1), (0, 0), (1, 1) and (2, 4), then cov (x, y) = (–2+ 4) + (–1+1) + (0×0) + (1×1) + (2×4) = 0

as $\bar{x} = 0$

Thus $r_{xy} = 0$

This does not mean that x and y are independent. In fact the relationship between x and y is $y = x^2$. Thus it is always wiser to draw a scatter diagram before reaching conclusion about the existence of correlation between a pair of variables.

There are some cases when we may find a correlation between two variables although the two variables are not causally related. This is due to the existence of a third variable which is related to both the variables under consideration. Such a correlation is known as spurious correlation or non-sense correlation. As an example, there could be a positive correlation between production of rice and that of iron in India for the last twenty years due to the effect of a third variable time on both these variables. It is necessary to eliminate the influence of the third variable before computing correlation between the two original variables.

Correlation coefficient measuring a linear relationship between the two variables indicates the amount of variation of one variable accounted for by the other variable. A better measure for this purpose is provided by the square of the correlation coefficient, Known as 'coefficient of determination'. This can be interpreted as the ratio between the explained variance to total variance i.e.

$$r^2 = \frac{\text{Explained variance}}{\text{Total variance}}$$

Thus a value of 0.6 for r indicates that $(0.6)^2 \times 100\%$ or 36 per cent of the variation has been accounted for by the factor under consideration and the remaining 64 per cent variation is due to other factors. The 'coefficient of non-determination' is given by $(1-r^2)$ and can be interpreted as the ratio of unexplained variance to the total variance.

Coefficient of non-determination = $(1-r^2)$

Regression analysis, as we have already seen, is concerned with establishing a functional relationship between two variables and using this relationship for making future projection. This can be applied, unlike correlation for any type of relationship linear as well as curvilinear. **The two lines of regression coincide i.e. become identical when r = –1 or 1 or in other words, there is a perfect negative or positive correlation between the two variables under discussion. If r = 0 Regression lines are perpendicular to each other.**

EXERCISE

Set A

Write the correct answers. Each question carries 1 mark.

1. Bivariate Data are the data collected for
   (a) Two variables
   (b) More than two variables
   (c) Two variables at the same point of time
   (d) Two variables at different points of time.

2. For a bivariate frequency table having (p + q) classification the total number of cells is
   (a) p (b) p + q
   (c) q (d) pq

3. Some of the cell frequencies in a bivariate frequency table may be
   (a) Negative (b) Zero
   (c) a or b (d) Non of these

4. For a p x q bivariate frequency table, the maximum number of marginal distributions is
   (a) p (b) p + q
   (c) 1 (d) 2

5. For a p x q classification of bivariate data, the maximum number of conditional distributions is
   (a) p (b) p + q
   (c) pq (d) p or q

6. Correlation analysis aims at
   (a) Predicting one variable for a given value of the other variable
   (b) Establishing relation between two variables
   (c) Measuring the extent of relation between two variables
   (d) Both (b) and (c).

7. Regression analysis is concerned with
   (a) Establishing a mathematical relationship between two variables
   (b) Measuring the extent of association between two variables
   (c) Predicting the value of the dependent variable for a given value of the independent variable
   (d) Both (a) and (c).

8. What is spurious correlation?
   (a) It is a bad relation between two variables.
   (b) It is very low correlation between two variables.
   (c) It is the correlation between two variables having no causal relation.
   (d) It is a negative correlation.
9. Scatter diagram is considered for measuring
   (a) Linear relationship between two variables
   (b) Curvilinear relationship between two variables
   (c) Neither (a) nor (b)
   (d) Both (a) and (b).
10. If the plotted points in a scatter diagram lie from upper left to lower right, then the correlation is
   (a) Positive (b) Zero
   (c) Negative (d) None of these.
11. If the plotted points in a scatter diagram are evenly distributed, then the correlation is
   (a) Zero (b) Negative
   (c) Positive (d) (a) or (b).
12. If all the plotted points in a scatter diagram lie on a single line, then the correlation is
   (a) Perfect positive (b) Perfect negative
   (c) Both (a) and (b) (d) Either (a) or (b).
13. The correlation between shoe-size and intelligence is
   (a) Zero (b) Positive
   (c) Negative (d) None of these.
14. The correlation between the speed of an automobile and the distance travelled by it after applying the brakes is
   (a) Negative (b) Zero
   (c) Positive (d) None of these.
15. Scatter diagram helps us to
   (a) Find the nature correlation between two variables
   (b) Compute the extent of correlation between two variables
   (c) Obtain the mathematical relationship between two variables
   (d) Both (a) and (c).

16. Pearson's correlation coefficient is used for finding
    (a) Correlation for any type of relation
    (b) Correlation for linear relation only
    (c) Correlation for curvilinear relation only
    (d) Both (b) and (c).
17. Product moment correlation coefficient is considered for
    (a) Finding the nature of correlation
    (b) Finding the amount of correlation
    (c) Both (a) and (b)
    (d) Either (a) and (b).
18. If the value of correlation coefficient is positive, then the points in a scatter diagram tend to cluster
    (a) From lower left corner to upper right corner
    (b) From lower left corner to lower right corner
    (c) From lower right corner to upper left corner
    (d) From lower right corner to upper right corner.
19. When r = 1, all the points in a scatter diagram would lie
    (a) On a straight line directed from lower left to upper right
    (b) On a straight line directed from upper left to lower right
    (c) On a straight line
    (d) Both (a) and (b).
20. Product moment correlation coefficient may be defined as the ratio of
    (a) The product of standard deviations of the two variables to the covariance between them
    (b) The covariance between the variables to the product of the variances of them
    (c) The covariance between the variables to the product of their standard deviations
    (d) Either (b) or (c).
21. The covariance between two variables is
    (a) Strictly positive (b) Strictly negative
    (c) Always 0 (d) Either positive or negative or zero.
22. The coefficient of correlation between two variables
    (a) Can have any unit.
    (b) Is expressed as the product of units of the two variables

(c) Is a unit free measure

(d) None of these.

23. What are the limits of the correlation coefficient?

(a) No limit (b) –1 and 1

(c) 0 and 1, including the limits (d) –1 and 1, including the limits

24. In case the correlation coefficient between two variables is 1, the relationship between the two variables would be

(a) $y = a + bx$ (b) $y = a + bx, b > 0$

(c) $y = a + bx, b < 0$ (d) $y = a + bx$, both a and b being positive.

25. If the relationship between two variables x and y in given by $2x + 3y + 4 = 0$, then the value of the correlation coefficient between x and y is

(a) 0 (b) 1

(c) –1 (d) negative.

26. For finding correlation between two attributes, we consider

(a) Pearson's correlation coefficient

(b) Scatter diagram

(c) Spearman's rank correlation coefficient

(d) Coefficient of concurrent deviations.

27. For finding the degree of agreement about beauty between two Judges in a Beauty Contest, we use

(a) Scatter diagram (b) Coefficient of rank correlation

(c) Coefficient of correlation (d) Coefficient of concurrent deviation.

28. If there is a perfect disagreement between the marks in Geography and Statistics, then what would be the value of rank correlation coefficient?

(a) Any value (b) Only 1

(c) Only –1 (d) (b) or (c)

29. When we are not concerned with the magnitude of the two variables under discussion, we consider

(a) Rank correlation coefficient (b) Product moment correlation coefficient

(c) Coefficient of concurrent deviation (d) (a) or (b) but not (c).

30. What is the quickest method to find correlation between two variables?

(a) Scatter diagram (b) Method of concurrent deviation

(c) Method of rank correlation (d) Method of product moment correlation

31. What are the limits of the coefficient of concurrent deviations?

(a) No limit

(b) Between –1 and 0, including the limiting values

(c) Between 0 and 1, including the limiting values

(d) Between –1 and 1, the limiting values inclusive

32. If there are two variables x and y, then the number of regression equations could be

(a) 1 (b) 2

(c) Any number (d) 3.

33. Since Blood Pressure of a person depends on age, we need consider

(a) The regression equation of Blood Pressure on age

(b) The regression equation of age on Blood Pressure

(c) Both (a) and (b)

(d) Either (a) or (b).

34. The method applied for deriving the regression equations is known as

(a) Least squares (b) Concurrent deviation

(c) Product moment (d) Normal equation.

35. The difference between the observed value and the estimated value in regression analysis is known as

(a) Error (b) Residue

(c) Deviation (d) (a) or (b).

36. The errors in case of regression equations are

(a) Positive (b) Negative

(c) Zero (d) All these.

37. The regression line of y on x is derived by

(a) The minimisation of vertical distances in the scatter diagram

(b) The minimisation of horizontal distances in the scatter diagram

(c) Both (a) and (b)

(d) (a) or (b).

38. The two lines of regression become identical when

(a) $r = 1$ (b) $r = -1$

(c) $r = 0$ (d) (a) or (b).

39. What are the limits of the two regression coefficients?

(a) No limit (b) Must be positive

(c) One positive and the other negative

(d) Product of the regression coefficient must be numerically less than unity.

40. The regression coefficients remain unchanged due to a

(a) Shift of origin (b) Shift of scale

(c) Both (a) and (b) (d) (a) or (b).

41. If the coefficient of correlation between two variables is –0 9, then the coefficient of determination is

(a) 0.9 (b) 0.81

(c) 0.1 (d) 0.19.

42. If the coefficient of correlation between two variables is 0.7 then the percentage of variation unaccounted for is

(a) 70% (b) 30%

(c) 51% (d) 49%

Set B

Answer the following questions by writing the correct answers. Each question carries 2 marks.

1. If for two variable x and y, the covariance, variance of x and variance of y are 40, 16 and 256 respectively, what is the value of the correlation coefficient?

(a) 0.01 (b) 0.625

(c) 0.4 (d) 0.5

2. If cov(x, y) = 15, what restrictions should be put for the standard deviations of x and y?

(a) No restriction.

(b) The product of the standard deviations should be more than 15.

(c) The product of the standard deviations should be less than 15.

(d) The sum of the standard deviations should be less than 15.

3. If the covariance between two variables is 20 and the variance of one of the variables is 16, what would be the variance of the other variable?

(a) More than 100 (b) More than 10

(c) Less than 10 (d) More than 1.25

4. If $y = a + bx$, then what is the coefficient of correlation between x and y?

(a) 1 (b) –1

(c) 1 or –1 according as $b > 0$ or $b < 0$ (d) none of these.

5. If $r = 0.6$ then the coefficient of non-determination is

(a) 0.4 (b) –0.6

(c) 0.36 (d) 0.64

6. If u + 5x = 6 and 3y – 7v = 20 and the correlation coefficient between x and y is 0.58 then what would be the correlation coefficient between u and v?

(a) 0.58 (b) –0.58

(c) –0.84 (d) 0.84

7. If the relation between x and u is 3x + 4u + 7 = 0 and the correlation coefficient between x and y is –0.6, then what is the correlation coefficient between u and y?

(a) –0.6 (b) 0.8

(c) 0.6 (d) –0.8

8. From the following data

x:	2	3	5	4	7
y:	4	6	7	8	10

Two coefficient of correlation was found to be 0.93. What is the correlation between u and v as given below?

u:	–3	–2	0	–1	2
v:	–4	–2	–1	0	2

(a) –0.93 (b) 0.93 (c) 0.57 (d) –0.57

9. Referring to the data presented in Q. No. 8, what would be the correlation between u and v?

u:	10	15	25	20	35
v:	–24	–36	–42	–48	–60

(a) –0.6 (b) 0.6 (c) –0.93 (d) 0.93

10. If the sum of squares of difference of ranks, given by two judges A and B, of 8 students in 21, what is the value of rank correlation coefficient?

(a) 0.7 (b) 0.65 (c) 0.75 (d) 0.8

11. If the rank correlation coefficient between marks in management and mathematics for a group of student in 0.6 and the sum of squares of the differences in ranks in 66, what is the number of students in the group?

(a) 10 (b) 9 (c) 8 (d) 11

12. While computing rank correlation coefficient between profit and investment for the last 6 years of a company the difference in rank for a year was taken 3 instead of 4. What is the rectified rank correlation coefficient if it is known that the original value of rank correlation coefficient was 0.4?

(a) 0.3 (b) 0.2 (c) 0.25 (d) 0.28

13. For 10 pairs of observations, No. of concurrent deviations was found to be 4. What is the value of the coefficient of concurrent deviation?

(a) $\sqrt{0.2}$ (b) $-\sqrt{0.2}$ (c) 1/3 (d) –1/3

14. The coefficient of concurrent deviation for p pairs of observations was found to be $1/\sqrt{3}$. If the number of concurrent deviations was found to be 6, then the value of p is.

    (a) 10 (b) 9 (c) 8 (d) none of these

15. What is the value of correlation coefficient due to Pearson on the basis of the following data:

x:	–5	–4	–3	–2	–1	0	1	2	3	4	5
y:	27	18	11	6	3	2	3	6	11	18	27

    (a) 1 (b) –1 (c) 0 (d) –0.5

16. Following are the two normal equations obtained for deriving the regression line of y and x:

    $5a + 10b = 40$

    $10a + 25b = 95$

    The regression line of y on x is given by

    (a) $2x + 3y = 5$ (b) $2y + 3x = 5$ (c) $y = 2 + 3x$ (d) $y = 3 + 5x$

17. If the regression line of y on x and of x on y are given by $2x + 3y = -1$ and $5x + 6y = -1$ then the arithmetic means of x and y are given by

    (a) (1, –1) (b) (–1, 1) (c) (–1, –1) (d) (2, 3)

18. Given the regression equations as $3x + y = 13$ and $2x + 5y = 20$, which one is the regression equation of y on x?

    (a) 1st equation (b) 2nd equation (c) both (a) and (b) (d) none of these.

19. Given the following equations: $2x - 3y = 10$ and $3x + 4y = 15$, which one is the regression equation of x on y ?

    (a) 1st equation (b) 2nd equation (c) both the equations (d) none of these

20. If $u = 2x + 5$ and $v = -3y - 6$ and regression coefficient of y on x is 2.4, what is the regression coefficient of v on u?

    (a) 3.6 (b) –3.6 (c) 2.4 (d) –2.4

21. If $4y - 5x = 15$ is the regression line of y on x and the coefficient of correlation between x and y is 0.75, what is the value of the regression coefficient of x on y?

    (a) 0.45 (b) 0.9375 (c) 0.6 (d) none of these

22. If the regression line of y on x and that of x on y are given by $y = -2x + 3$ and $8x = -y + 3$ respectively, what is the coefficient of correlation between x and y?

    (a) 0.5 (b) $-1/\sqrt{2}$ (c) –0.5 (d) none of these

23. If the regression coefficient of y on x, the coefficient of correlation between x and y and variance of y are –3/4, $\frac{\sqrt{3}}{2}$ and 4 respectively, what is the variance of x?

    (a) $2/\sqrt{3/2}$ (b) 16/3 (c) 4/3 (d) 4

24. If $y = 3x + 4$ is the regression line of y on x and the arithmetic mean of x is –1, what is the arithmetic mean of y?

(a) 1 (b) –1 (c) 7 (d) none of these

SET C

Write down the correct answers. Each question carries 5 marks.

1. What is the coefficient of correlation from the following data?

x:	1	2	3	4	5
y:	8	6	7	5	5

(a) 0.75 (b) –0.75 (c) –0.85 (d) 0.82

2. The coefficient of correlation between x and y where

x:	64	60	67	59	69
y:	57	60	73	62	68

is

(a) 0.655 (b) 0.68 (c) 0.73 (d) 0.758

3. What is the coefficient of correlation between the ages of husbands and wives from the following data?

Age of husband (year):	46	45	42	40	38	35	32	30	27	25
Age of wife (year):	37	35	31	28	30	25	23	19	19	18

(a) 0.58 (b) 0.98 (c) 0.89 (d) 0.92

4. Given that for twenty pairs of observations, $\sum xu = 525$, $\sum x = 129$, $\sum u = 97$, $\sum x^2 = 687$, $\sum u^2 = 427$ and $y = 10 - 3u$, the coefficient of correlation between x and y is

(a) –0.7 (b) 0.74 (c) –0.74 (d) 0.75

5. The following results relate to bivariate data on (x, y):

$\sum xy = 414$, $\sum x = 120$, $\sum y = 90$, $\sum x^2 = 600$, $\sum y^2 = 300$, n = 30. Later or, it was known that two pairs of observations (12, 11) and (6, 8) were wrongly taken, the correct pairs of observations being (10, 9) and (8, 10). The corrected value of the correlation coefficient is

(a) 0.752 (b) 0.768 (c) 0.846 (d) 0.953

6. The following table provides the distribution of items according to size groups and also the number of defectives:

Size group:	9-11	11-13	13-15	15-17	17-19
No. of items:	250	350	400	300	150
No. of defective items:	25	70	60	45	20

The correlation coefficient between size and defectives is

(a) 0.25 (b) 0.12 (c) 0.14 (d) 0.07

7. For two variables x and y, it is known that cov (x, y) = 80, variance of x is 16 and sum of squares of deviation of y from its mean is 250. The number of observations for this bivariate data is

(a) 7 (b) 8 (c) 9 (d) 10

8. Eight contestants in a musical contest were ranked by two judges A and B in the following manner:

Serial Number of the contestants:	1	2	3	4	5	6	7	8
Rank by Judge A:	7	6	2	4	5	3	1	8
Rank by Judge B:	5	4	6	3	8	2	1	7

The rank correlation coefficient is

(a) 0.65 (b) 0.63 (c) 0.60 (d) 0.57

9. Following are the marks of 10 students in Botany and Zoology:

Serial No.:	1	2	3	4	5	6	7	8	9	10
Marks in Botany:	58	43	50	19	28	24	77	34	29	75
Marks in Zoology:	62	63	79	56	65	54	70	59	55	69

The coefficient of rank correlation between marks in Botany and Zoology is

(a) 0.65 (b) 0.70 (c) 0.72 (d) 0.75

10. What is the value of Rank correlation coefficient between the following marks in Physics and Chemistry:

Roll No.:	1	2	3	4	5	6
Marks in Physics:	25	30	46	30	55	80
Marks in Chemistry:	30	25	50	40	50	78

(a) 0.782 (b) 0.696 (c) 0.932 (d) 0.857

11. What is the coefficient of concurrent deviations for the following data:

Supply:	68	43	38	78	66	83	38	23	83	63	53
Demand:	65	60	55	61	35	75	45	40	85	80	85

(a) 0.82 (b) 0.85 (c) 0.89 (d) –0.81

12. What is the coefficient of concurrent deviations for the following data:

Year:	1996	1997	1998	1999	2000	2001	2002	2003
Price:	35	38	40	33	45	48	49	52
Demand:	36	35	31	36	30	29	27	24

(a) –0.43 (b) 0.43 (c) 0.5 (d) $\sqrt{2}$

13. The regression equation of y on x for the following data:

x	41	82	62	37	58	96	127	74	123	100
y	28	56	35	17	42	85	105	61	98	73

Is given by

(a) y = 1.2x – 15 (b) y = 1.2x + 15 (c) y = 0.93x – 14.64 (d) y = 1.5x – 10.89

14. The following data relate to the heights of 10 pairs of fathers and sons:

(175, 173), (172, 172), (167, 171), (168, 171), (172, 173), (171, 170), (174, 173), (176, 175) (169, 170), (170, 173)

The regression equation of height of son on that of father is given by

(a) y = 100 + 5x (b) y = 99.708 + 0.405x (c) y = 89.653 + 0.582x (d) y = 88.758 + 0.562x

15. The two regression coefficients for the following data:

x:	38	23	43	33	28
y:	28	23	43	38	8

are

(a) 1.2 and 0.4 (b) 1.6 and 0.8 (c) 1.7 and 0.8 (d) 1.8 and 0.3

16. For y = 25, what is the estimated value of x, from the following data:

X:	11	12	15	16	18	19	21
Y:	21	15	13	12	11	10	9

(a) 15 (b) 13.926 (c) 13.588 (d) 14.986

17. Given the following data:

Variable:	x	y
Mean:	80	98
Variance:	4	9

Coefficient of correlation = 0.6

What is the most likely value of y when x = 90 ?

(a) 90 (b) 103 (c) 104 (d) 107

18. The two lines of regression are given by

8x + 10y = 25 and 16x + 5y = 12 respectively.

If the variance of x is 25, what is the standard deviation of y?

(a) 16 (b) 8 (c) 64 (d) 4

19. Given below the information about the capital employed and profit earned by a company over the last twenty five years:

	Mean	SD
Capital employed (0000 Rs)	62	5
Profit earned (000 Rs)	25	6

Correlation coefficient between capital employed and profit = 0.92. The sum of the Regression coefficients for the above data would be:

(a) 1.871 (b) 2.358 (c) 1.968 (d) 2.346

20. The coefficient of correlation between cost of advertisement and sales of a product on the basis of the following data:

Ad cost (000 Rs):	75	81	85	105	93	113	121	125
Sales (000 000 Rs):	35	45	59	75	43	79	87	95

is

(a) 0.85 (b) 0.89 (c) 0.95 (d) 0.98

ANSWERS

Set A											
1.	(c)	2.	(d)	3.	(b)	4.	(d)	5.	(b)	6.	(d)
7.	(d)	8.	(c)	9.	(d)	10.	(c)	11.	(a)	12.	(d)
13.	(a)	14.	(a)	15.	(a)	16.	(b)	17.	(c)	18.	(a)
19.	(a)	20.	(c)	21.	(d)	22.	(c)	23.	(d)	24.	(b)
25.	(c)	26.	(c)	27.	(b)	28.	(c)	29.	(c)	30.	(b)
31.	(d)	32.	(b)	33.	(a)	34.	(a)	35.	(d)	36.	(d)
37.	(a)	38.	(d)	39.	(d)	40.	(a)	41.	(b)	42.	(c)
Set B											
1.	(b)	2.	(b)	3.	(a)	4.	(c)	5.	(d)	6.	(b)
7.	(c)	8.	(b)	9.	(c)	10.	(c)	11.	(a)	12.	(b)
13.	(d)	14.	(a)	15.	(c)	16.	(c)	17.	(a)	18.	(b)
19.	(d)	20.	(b)	21.	(a)	22.	(c)	23.	(b)	24.	(a)
Set C											
1.	(c)	2.	(a)	3.	(b)	4.	(c)	5.	(c)	6.	(d)
7.	(d)	8.	(d)	9.	(d)	10.	(d)	11.	(c)	12.	(a)
13.	(c)	14.	(b)	15.	(a)	16.	(c)	17.	(d)	18.	(b)
19.	(a)	20.	(c)								

ADDITIONAL QUESTION BANK

1. ————— is concerned with the measurement of the "strength of association" between variables.

   (a) correlation (b) regression (c) both (d) none

2. ————— gives the mathematical relationship of the variables.

   (a) correlation (b) regression (c) both (d) none

3. When high values of one variable are associated with high values of the other & low values of one variable are associated with low values of another, then they are said to be

   (a) positively correlated (b) directly correlated
   (c) both (d) none

4. If high values of one tend to low values of the other, they are said to be

   (a) negatively correlated (b) inversely correlated
   (c) both (d) none

5. Correlation coefficient between two variables is a measure of their linear relationship .

   (a) true (b) false (c) both (d) none

6. Correlation coefficient is dependent of the choice of both origin & the scale of observations.

   (a) True (b) false (c) both (d) none

7. Correlation coefficient is a pure number.

   (a) true (b) false (c) both (d) none

8. Correlation coefficient is ————— of the units of measurement.

   (a) dependent (b) independent (c) both (d) none

9. The value of correlation coefficient lies between

   (a) –1 and +1 (b) –1 and 0
   (c) 0 and 1 Inclusive of these two values (d) none.

10. Correlation coefficient can be found out by

    (a) Scatter Diagram (b) Rank Method (c) both (d) none.

11. Covariance measures _________ variations of two variables.

    (a) joint (b) single (c) both (d) none

12. In calculating the Karl Pearson's coefficient of correlation it is necessary that the data should be of numerical measurements. The statement is

    (a) valid (b) not valid (c) both (d) none

13. Rank correlation coefficient lies between

    (a) 0 to 1 (b) –1 to +1 inclusive of these value
    (c) –1 to 0 (d) both

14. A coefficient near +1 indicates tendency for the larger values of one variable to be associated with the larger values of the other.

(a) true (b) false (c) both (d) none

15. In rank correlation coefficient the association need not be linear.

(a) true (b) false (c) both (d) none

16. In rank correlation coefficient only an increasing/decreasing relationship is required.

(a) false (b) true (c) both (d) none

17. Great advantage of ____________ is that it can be used to rank attributes which can not be expressed by way of numerical value .

(a) concurrent correlation (b) regression
(c) rank correlation (d) none

18. The sum of the difference of rank is

(a) 1 (b) –1 (c) 0 (d) none.

19. Karl Pearson's coefficient is defined from

(a) ungrouped data (b) grouped data (c) both (d) none.

20. Correlation methods are used to study the relationship between two time series of data which are recorded annually, monthly, weekly, daily and so on.

(a) True (b) false (c) both (d) none

21. Age of Applicants for life insurance and the premium of insurance – correlation is

(a) positive (b) negative (c) zero (d) none

22. "Unemployment index and the purchasing power of the common man" ——Correlation is

(a) positive (b) negative (c) zero (d) none

23. Production of pig iron and soot content in Durgapur – Correlations are

(a) positive (b) negative (c) zero (d) none

24. "Demand for goods and their prices under normal times" —— Correlation is

(a) positive (b) negative (c) zero (d) none

25. ___________ is a relative measure of association between two or more variables.

(a) Coefficient of correlation (b) Coefficient of regression
(c) both (d) none

26. The lines of regression passes through the points, bearing _________ no. of points on both sides

(a) equal (b) unequal (c) zero (d) none

27. Under Algebraic Method we get ———— linear equations .

(a) one (b) two (c) three (d) none

28. In linear equations Y = a + bX and X= a + bY 'a' is the

(a) intercept of the line (b) slope
(c) both (d) none

29. In linear equations Y = a + bX and X = a + bY ' b ' is the

(a) intercept of the line (b) slope of the line
(c) both (d) none

30. The regression equations Y = a + bX and X = a + bY are based on the method of

(a) greatest squares (b) least squares (c) both (d) none

31. The line Y = a + bX represents the regression equation of

(a) Y on X (b) X on Y (c) both (d) none

32. The line X = a + bY represents the regression equation of

(a) Y on X (b) X onY (c) both (d) none

33. Two regression lines always intersect at the means.

(a) true (b) false (c) both (d) none

34. r, b_{xy}, b_{yx} all have ______ sign.

(a) different (b) same (c) both (d) none

35. The regression coefficients are zero if r is equal to

(a) 2 (b) –1 (c) 1 (d) 0

36. The regression lines are identical if r is equal to

(a) +1 (b) –1 (c) ± 1 (d) 0

37. The regression lines are perpendicular to each other if r is equal to
(a) 0 (b) +1 (c) –1 (d) ± 1

38. Feature of Least Square regression lines are——— The sum of the deviations at the Y's or the X's from their regression lines are zero.

(a) true (b) false (c) both (d) none

39. The coefficient of determination is defined by the formula

(a) $r^2 = 1 - \dfrac{\text{unexplained variance}}{\text{total variance}}$ (b) $r^2 = \dfrac{\text{explained variance}}{\text{total variance}}$

(c) both (d) none

40. If the line Y = 13 –3X /2 is the regression equation of y on x then byx is

(a) $\dfrac{2}{3}$ (b) $\dfrac{-2}{3}$ (c) $\dfrac{3}{2}$ (d) $\dfrac{-3}{2}$

41. In the line Y = 19 – 5X/2 is the regresson equation x on y then bxy is,
(a) 19/2 (b) 5/2 (c) –5/2 (d) –2/5

42. The line X = 31/6 — Y/6 is the regression equation of
(a) Y on X (b) X on Y (c) both (d) we can not say

43. In the regression equation x on y, X = 35/8 – 2Y /5, b_{xy} is equal to
(a) –2/5 (b) 35/8 (c) 2/5 (d) 5/2

44. The square of coefficient of correlation 'r' is called the coefficient of
(a) determination (b) regression (c) both (d) none

45. A relationship $r^2 = 1 - \frac{500}{300}$ is not possible
(a) true (b) false (c) both (d) none

46. Whatever may be the value of r, positive or negative, its square will be
(a) negative only (b) positive only (c) zero only (d) none only

47. Simple correlation is called
(a) linear correlation (b) nonlinear correlation
(c) both (d) none

48. A scatter diagram indicates the type of correlation between two variables.
(a) true (b) false (c) both (d) none

49. If the pattern of points (or dots) on the scatter diagram shows a linear path diagonally across the graph paper from the bottom left- hand corner to the top right, correlation will be
(a) negative (b) zero (c) positive (d) none

50. The correlation coefficient being +1 if the slope of the straight line in a scatter diagram is
(a) positive (b) negative (c) zero (d) none

51. The correlation coefficient being –1 if the slope of the straight line in a scatter diagram is
(a) positive (b) negative (c) zero (d) none

52. The more scattered the points are around a straight line in a scattered diagram the _______ is the correlation coefficient.
(a) zero (b) more (c) less (d) none

53. If the values of y are not affected by changes in the values of x, the variables are said to be
(a) correlated (b) uncorrelated (c) both (d) zero

54. If the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable, then correlation is said to be
(a) non linear (b) linear (c) both (d) none

55. Variance may be positive, negative or zero.

(a) true (b) false (c) both (d) none

56. Covariance may be positive, negative or zero.

(a) true (b) false (c) both (d) none

57. Correlation coefficient between x and y = correlation coefficient between u and v

(a) true (b) false (c) both (d) none

58. In case ' The ages of husbands and wives' ———— correlation is

(a) positive (b) negative (c) zero (d) none

59. In case 'Shoe size and intelligence'

(a) positive correlation (b) negative correlation
(c) no correlation (d) none

60. In case 'Insurance companies' profits and the no of claims they have to pay "——

(a) positive correlation (b) negative correlation
(c) no correlation (d) none

61. In case 'Years of education and income'———

(a) positive correlation (b) negative correlation
c) no correlation (d) none

62. In case 'Amount of rainfall and yield of crop'——

(a) positive correlation (b) negative correlation
(c) no correlation (d) none

63. For calculation of correlation coefficient, a change of origin is

(a) not possible (b) possible (c) both (d) none

64. The relation $r_{xy} = \text{cov}(x,y)/\sigma_x.\sigma_y$ is

(a) true (b) false (c) both (d) none

65. A small value of r indicates only a _________ linear type of relationship between the variables.

(a) good (b) poor (c) maximum (d) highest

66. Two regression lines coincide when

(a) r = 0 (b) r = 2 (c) r = $\pm$ 1 (d) none

67. Neither y nor x can be estimated by a linear function of the other variable when r is equal to

(a) + 1 (b) – 1 (c) 0 (d) none

68. When r = 0 then cov (x,y) is equal to

(a) + 1 (b) – 1 (c) 0 (d) none

69. When the variables are not independent, the correlation coefficient may be zero

(a) true (b) false (c) both (d) none

70. b_{xy} is called regression coefficient of

(a) x on y (b) y on x (c) both (d) none

71. b_{yx} is called regression coefficient of

(a) x on y (b) y on x (c) both (d) none

72. The slopes of the regression line of y on x is

(a) b_{yx} (b) b_{xy} (c) b_{xx} (d) b_{yy}

73. The slopes of the regression line of x on y is

(a) b_{yx} (b) b_{xy} (c) $1/b_{xy}$ (d) $1/b_{yx}$

74. The angle between the regression lines depends on

(a) correlation coefficient (b) regression coefficient
(c) both (d) none

75. If x and y satisfy the relationship y = –5 + 7x, the value of r is

(a) 0 (b) – 1 (c) + 1 (d) none

76. If b_{yx} and b_{xy} are negative, r is

(a) positive (b) negative (c) zero (d) none

77. Correlation coefficient r lie between the regression coefficients b_{yx} and b_{xy}

(a) true (b) false (c) both (d) none

78. Since the correlation coefficient r cannot be greater than 1 numerically, the product of the regression must

(a) not exceed 1 (b) exceed 1 (c) be zero (d) none

79. The correlation coefficient r is the __________ of the two regression coefficients b_{yx} and b_{xy}

(a) A.M (b) G.M (c) H.M (d) none

80. Which is true?

(a) $b_{yx} = r\frac{\sigma_x}{\sigma_y}$ (b) $b_{yx} = r\frac{\sigma_y}{\sigma_x}$

(c) $b_{yx} = r\frac{\sigma_{xy}}{\sigma_x}$ (d) $b_{yx} = r\frac{\sigma_{yy}}{\sigma_x}$

81. Maximum value of Rank Correlation coefficient is

(a) –1 (b) + 1 (c) 0 (d) none

82. The partial correlation coefficient lies between

(a) –1 and +1 inclusive of these two value (b) 0 and + 1
(c) –1 and (d) none

83. r_{12} is the correlation coefficient between

(a) x_1 and x_2 (b) x_2 and x_1 (c) x_1 and x_3 (d) x_2 and x_3

84. r_{12} is the same as r_{21}

(a) true (b) false (c) both (d) none

85. In case of employed persons 'Age and income' correlation is

(a) positive (b) negative (c) zero (d) none

86. In case 'Speed of an automobile and the distance required to stop the car often applying brakes' – correlation is

(a) positive (b) negative (c) zero (d) none

87. In case 'Sale of woolen garments and day temperature'—— correlation is

(a) positive (b) negative (c) zero (d) none

88. In case 'Sale of cold drinks and day temperature' ——— correlation is

(a) positive (b) negative (c) zero (d) none

89. In case of 'Production and price per unit' – correlation is

(a) positive (b) negative (c) zero (d) none

90. If slopes at two regression lines are equal them r is equal to

(a) 1 (b) ± 1 (c) 0 (d) none

91. Co–variance measures the joint variations of two variables.

(a) true (b) false (c) both (d) none

92. The minimum value of correlation coefficient is

(a) 0 (b) –2 (c) 1 (d) –1

93. The maximum value of correlation coefficient is

(a) 0 (b) 2 (c) 1 (d) –1

94. When r = 0 , the regression coefficients are

(a) 0 (b) 1 (c) –1 (d) none

95. The regression equation of Y on X is, 2x + 3Y + 50 = 0. The value of b_{YX} is

(a) 2/3 (b) – 2/3 (c) –3/2 (d) none

96. In Method of Concurrent Deviations, only the directions of change (Positive direction / Negative direction) in the variables are taken into account for calculation of

(a) coefficient of S.D (b) coefficient of regression.
(c) coefficient of correlation (d) none

ANSWERS

1	(a)	2	(b)	3	(c)	4	(c)	5	(a)
6	(b)	7	(a)	8	(b)	9	(a)	10	(b)
11	(a)	12	(a)	13	(b)	14	(a)	15	(a)
16	(b)	17	(c)	18	(c)	19	(b)	20	(a)
21	(a)	22	(b)	23	(a)	24	(b)	25	(a)
26	(d)	27	(b)	28	(a)	29	(b)	30	(b)
31	(a)	32	(b)	33	(a)	34	(b)	35	(d)
36	(c)	37	(a)	38	(a)	39	(c)	40	(d)
41	(d)	42	(b)	43	(a)	44	(a)	45	(a)
46	(b)	47	(a)	48	(a)	49	(c)	50	(a)
51	(b)	52	(c)	53	(b)	54	(b)	55	(b)
56	(a)	57	(a)	58	(a)	59	(c)	60	(b)
61	(a)	62	(a)	63	(b)	64	(a)	65	(b)
66	(c)	67	(c)	68	(c)	69	(a)	70	(a)
71	(b)	72	(a)	73	(b)	74	(a)	75	(c)
76	(b)	77	(a)	78	(a)	79	(b)	80	(b)
81	(b)	82	(a)	83	(a)	84	(a)	85	(a)
86	(a)	87	(b)	88	(a)	89	(b)	90	(b)
91	(a)	92	(d)	93	(c)	94	(a)	95	(b)
96	(c)								