CHAPTER

17

Correlation and Regression

CORRELATION

Correlation is a fundamental concept in statistics that allows us to measure the statistical relationship between two variables. It helps us understand the extent to which changes in one variable are associated with changes in another. Imagine it as a way to quantify how two things interact or influence each other.

BIVARIATE DATA

In real life, we generally deal with more than one variable.

Example:

	Math-Enthusiast	Math-Neutral	Math-Disliker	Marginal (English)
English-Lover	10	5	3	18
English-Neutral	3	8	2	13
English-Disliker	1	2	7	10
Marginal (Math)	14	15	12	41

- The above type of data is called a Bivariable data.
- When we involve more than two variables in our study, that type of data is called Multivariate data.

However, in real-life situations, we often deal with more than just two variables, leading to what we call multivariate data. Multivariate data involves the measurement or observation of multiple variables for each individual or object in a study.

Thus, in the given scenario, the data of the singing competition involves two variables which represent bivariate data. If we further consider additional variables such as the age, gender, or singing experience of the participants, it would be an example of multivariate data.

MARGINAL DISTRIBUTION

The marginal distribution of a variable in a bivariate frequency distribution is the distribution of that variable alone, without considering the other variable. It is obtained by summing (or adding) the frequencies along one of the margins (rows or columns) of the distribution table.

Example: Consider a bivariate frequency distribution table representing the preferences of students regarding two subjects, Math and English. The table might look like this:

	Math-Enthusiast	Math-Neutral	Math-Disliker	Marginal (English)
English-Lover	10	5	3	18
English-Neutral	3	8	2	13
English-Disliker	1	2	7	10
Marginal (Math)	14	15	12	41

1. The marginal distribution of the variable "English preference" is obtained by summing the frequencies along the "Marginal (English)" column.
2. The marginal distribution of the variable "Math preference" is obtained by summing the frequencies along the "Marginal (Math)" row.

CONDITIONAL DISTRIBUTION

The conditional distribution of a variable in a bivariate frequency distribution is the distribution of that variable under the condition that a specific value of the other variable is known. It is calculated by dividing the joint frequency by the corresponding marginal frequency.

To find the conditional distribution of "Math preference" given that a student is an "English Lover," divide each frequency in the "English Lover" row by the marginal frequency in the "Marginal (English)" column.

	Math-Enthusiast	Math-Neutral	Math-Disliker	Marginal (English)
English-Lover	$\frac{10}{18}$	$\frac{5}{18}$	$\frac{3}{18}$	18

CORRELATION

Correlation refers to the statistical relationship between two variables. It measures the extent to which the change in one variable is accompanied by a change in the other. Correlation helps us understand the association or connection between different variables and how they interact with each other.

For example, if we consider the height of students as variable X and their weight as variable Y, we would expect that large values of X (tall students) correspond to large values of Y (higher weight), while small values of X (short students) correspond to small values of Y (lower weight). This indicates a correlation between height and weight, as they tend to change together.

Let's see one more example to understand correlation. We might find a high degree of relationship between the price of a product and consumer demand for that product. As the price increases, the demand may decrease, and vice versa. This demonstrates the correlation between price and demand.

POSITIVE AND NEGATIVE CORRELATION

- **Positive Correlation:** It indicates that as one variable increases (or decreases), the other variable also increases (or decreases). For instance, in a business context, if the amount spent on digital advertising (X) by a firm increases, there is a corresponding increase in total annual sales (Y). Conversely, a reduction in advertising expenditure is associated with a decrease in total sales.

- **Negative Correlation:** It indicates that as one variable increases (or decreases), the other variable decreases (or increases). Consider the price of a product (X) and its demand (Y). If there is a negative correlation, an increase in the product's price leads to a decrease in demand. Conversely, a decrease in price tends to result in an increase in demand for the product.

METHODS OF STUDYING CORRELATION

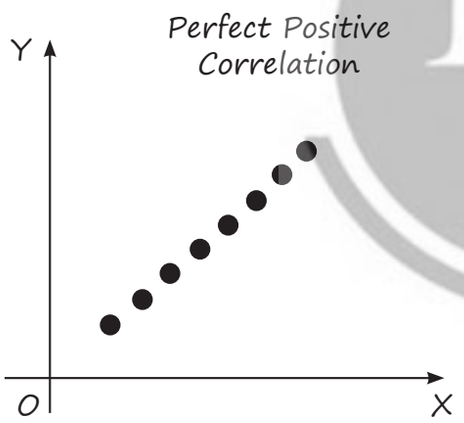
We shall discuss the following methods of measuring the linear relationship between two variables:

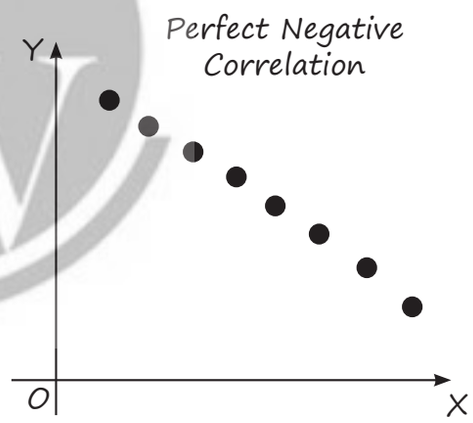
1. Scatter Diagram
2. Karl Pearson's Coefficient of Correlation
3. Spearman's Rank Correlation Coefficient
4. Coefficient of Concurrent deviations

SCATTER DIAGRAM

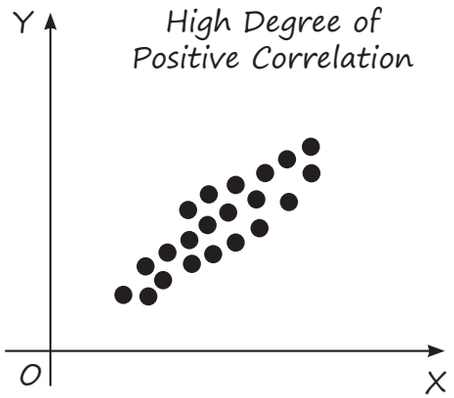
A scatter diagram is a graphical presentation of bivariate data $\{(X_i, Y_i): i = 1, 2, \ldots n\}$ on two quantitative variables X and Y that allows us to show two variables together, one on each axis, each pair being represented by a point on the graph as in coordinate geometry.

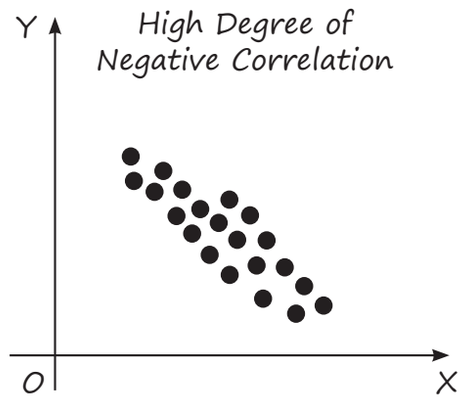
- **Perfect correlation**



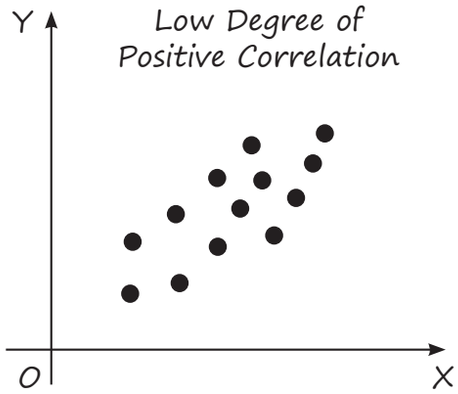


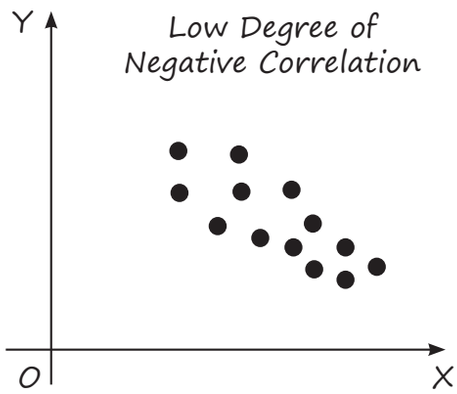
- **Very high degree of correlation**



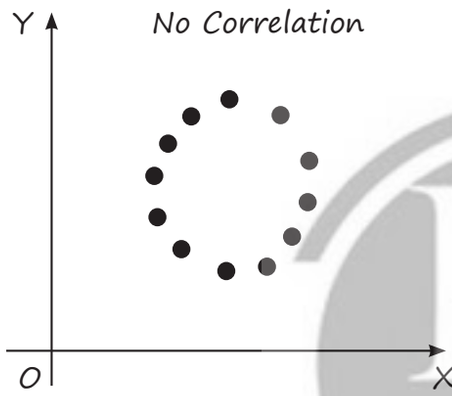


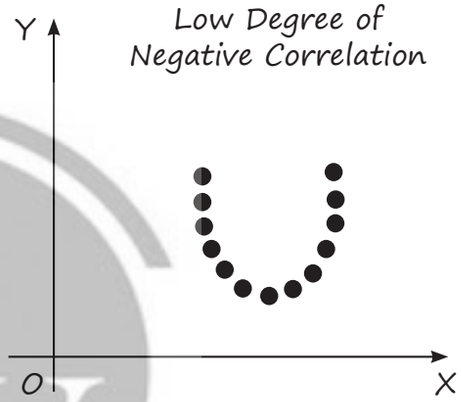
- **Low degree of correlation**





- **No correlation**





1. If r = 1, then all the points of the scatter diagram lie on a straight line having positive slope and we say that a perfect positive linear relationship exists between the two variables. Similarly, if r = −1, then all the points of the scatter diagram lie on a straight line having a negative slope and we say that a perfect negative linear relationship exists between the two variables.
2. If r is close to +1, then all the points of the scatter diagram closely follow a straight line having positive slope and we say that a high positive correlation exists between the two variables. Similarly, if r is close to −1, then all the points of the scatter diagram closely follow a straight line having a negative slope and we say that a high negative correlation exists between the two variables.
3. If r is close to 0, the linear relationship between the two variables is weak or perhaps non-existent.

THE CORRELATION COEFFICIENT

Correlation analysis attempts to measure the strength or closeness of linear relationships between two variables by means of a single number called a correlation coefficient.

Definition. The quantitative measure of strength in the linear relationship between two variables is called the correlation coefficient. It is denoted by r.

- Thus, the correlation coefficient r measures the extent to which the points cluster about a straight line.

- The correlation coefficient ranges from + 1 to − 1 i.e. $-1 \le r \le 1$.
- If two variables have no linear relationship, the correlation between them is zero.
- Consequently, the more correlation differs from zero, the stronger the linear relationship between the two variables.

The following table shows degrees of correlation according to various values of r.

Degree of Correlation	Positive	Negative
Perfect correlation	+1	−1
Very high degree of correlation	+ 0.9 to + 1	− 0.9 to − 1
Fairly high degree of correlation	+ 0.75 to + 0.9	− 0.75 to − 0.9
Moderate degree of correlation	+ 0.50 to + 0.75	− 0.50 to − 0.75
Low degree of correlation	+ 0.25 to + 0.50	− 0.25 to − 0.5
Very low degree of correlation	0 to + 0.25	− 0.25 to 0
No correlation	0	0

Example 5. The covariance between two variables is **(ICAI)**

(a) Strictly positive (b) Strictly negative

(c) Always 0 (d) Either positive or negative or zero

Sol. (d) Correlation between two variables can be measured on a scale that varies from +1 through 0 to −1.

Covariance values are not standardized. Therefore, the covariance can range from negative infinity to positive infinity. Thus, the value for a perfect linear relationship depends on the data. Because the data are not standardized, it is difficult to determine the strength of the relationship between the variables.

When one variable increases as the other increases then the correlation is positive.and when one decreases as the other increases then the correlation is negative.

Also, complete absence of correlation is represented as 0 (zero) correlation.

So, the correlation between two variable is either positive or negative or zero.

Hence, the correct option is (d).

COVARIANCE

- **Definition:** Consider a set of n pairs of observations $(x_1, y_1), (x_2, y_2), \ldots\ldots (x_n, y_n)$ on two quantitative variables x and y, where $x_1, x_2, \ldots\ldots$ denote observed values of the variable x, and $y_1, y_2, \ldots\ldots$ those of y.

The covariance between x and y, denoted by Cov (x, y), is given by:

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

Example 1. Find Cov (x,y) between x and y if

x	3	4	5	6	7
y	8	7	6	5	4

(a) 2 (b) 3 (c) −2 (d) −1

Sol. (c) According to the question, we have

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
3	8	–2	2	–4
4	7	–1	1	–1
5	6	0	0	0
6	5	1	–1	–1
7	4	2	–2	–4
$\bar{x} = \frac{25}{5} = 5$	$\bar{y} = \frac{30}{5} = 6$			–10

Therefore, Cov $(x, y) = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n} = -10/5 = -2$

Hence, the correct option is (c).

Another Formula for Cov (x, y)

This formula is particularly useful when $\bar{x}$ or $\bar{y}$ is not an integer. The formula is:

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

Example 2. Find Cov (x,y) between x and y if

x	3	4	5	6	7
y	8	7	6	5	4

(a) 2 (b) 3 (c) –2 (d) –1

Sol. (c) $\text{COV}(x, y) = \dfrac{\sum xy}{n} - \left(\dfrac{\sum x}{n}\right)\left(\dfrac{\sum y}{n}\right)$

Therefore, the required table is,

x	y	xy
3	8	24
4	7	28
5	6	30
6	5	30
7	4	28
$\sum x = 25$	$\sum y = 30$	$\sum xy = 140$

Here,

$\sum x = 25$,

$\sum y = 30$

$\sum xy = 140$ and $n = 5$

Put these values in the above formula, then we get

$$Cov\,(x,y) = \frac{140}{5} - \left(\frac{25}{5}\right)\left(\frac{30}{5}\right)$$

$Cov\,(x, y) = 28 - (5)(6)$

$Cov\,(x, y) = -2$

Hence, the correct option is (c).

Example 3. Calculate the covariance between X and Y for the following data:

X	1	2	3	4	5	6	7	8	9	10
Y	6	9	6	7	8	5	12	3	17	1

(a) 2 (b) 0.2 (c) −0.4 (d) 0.4

Sol. (d) $COV(x,y) = \frac{\sum xy}{n} - \left(\frac{\sum x}{n}\right)\left(\frac{\sum y}{n}\right)$

Therefore, the required table is,

x	y	xy
1	6	6
2	9	18
3	6	18
4	7	28
5	8	40
6	5	30
7	12	84
8	3	24
9	17	153
10	1	10
$\sum x = 55$	$\sum y = 74$	$\sum xy = 411$

Here,

$\sum x = 55$,

$\sum y = 74$

$\sum xy = 411$ and $n = 10$

Put these values in the above formula,Then we get

$$Cov\,(x,y) = \frac{411}{10} - \left(\frac{55}{10}\right)\left(\frac{74}{10}\right) = \frac{411}{10} - \frac{407}{10} = \frac{4}{10} = 0.4$$

Hence, the correct option is (d).

Example 4. Find the covariance between X and Y, given that $\sum X = 60$, $\sum Y = 60$, $\sum XY = 574$, and n = 10.

(a) 2 (b) 3.4 (c) 3.2 (d) 3.6

Sol. (b) Given that, $\sum X = 60$, $\sum Y = 60$, $\sum XY = 574$, and n = 10

Put these values in $COV(x,y) = \frac{\sum xy}{n} - \left(\frac{\sum x}{n}\right)\left(\frac{\sum y}{n}\right)$, then we get

$$COV(x,y) = \frac{574}{10} - \left(\frac{60}{10}\right)\left(\frac{90}{10}\right)$$

$$COV(x,y) = \frac{574}{10} - \frac{540}{10} = \frac{34}{10} = 3.4$$

Hence, the correct option is (b).

KARL PEARSON'S COEFFICIENT OF CORRELATION

The Karl Pearson's coefficient of correlation, also called the Pearson's product - moment correlation coefficient, is the most widely used method of measuring the linear correlation between two variables.

Definition: The Karl Pearson's coefficient of correlation between two variables x and y, denoted by r, is defined by

$$r = \frac{Cov\ (xy)}{\sigma_x * \sigma_y}$$

where var. x and var. y are the variances of the values of x and y respectively, while σ_x and σ_y are their standard deviations.

Example 6. 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? (**ICAI**)

(a) 0.01 (b) 0.625 (c) 0.4 (d) 0.5

Sol. (b) We know that,

Correlation Coefficient, $r = \frac{cov(x,y)}{\sigma_x \times \sigma_y}$

where ,

r = Correlation Coefficient

cov(x, y) = Covariance of x and y

σ_x = Standard deviation of x

σ_y = Standard deviation of y

Also, Standard Deviation $= \sqrt{Variance}$

$\Rightarrow$ Standard Deviation of x, $\sigma_x = \sqrt{16} = 4$

$\Rightarrow$ Standard Deviation of y, $\sigma_y = \sqrt{256} = 16$

Calculating Correlation Coefficient, we get

$$(r) = \frac{40}{4 \times 16} = \frac{40}{64} = 0.625$$

Hence, the correct answer is option (b) i.e. 0.625.

Example 7. The covariance between the length and weight of five items is 6 and their standard deviations are 2.45 and 2.61 respectively. Find the coefficient of correlation between length and weight.

(a) 0.9383 (b) 1.9385 (c) 0.2583 (d) 3.6353

Sol. (a) Given: Cov (x, y) = 6, σ_x = 2.45 and σ_y = 2.61

Using, $r = \frac{Cov(xy)}{\sigma_x * \sigma_y}$

Substituting the values in the above formula, we get

$$r = \frac{6}{(2.45)(2.61)} = 0.9383$$

Hence, the correct option is (a).

Example 8. The coefficient of correlation between x and y is 0.5, the covariance is 16. If the standard deviation of x is 4 then the standard deviation of y is **(Jan 2021)**

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

Sol: (b) Given: r = 0.5, cov(x, y) = 16, σ_x= 4 and σ_y= ?

We know that,

Coefficient of correlation is given by:

$$r = \frac{cov\ (x, y)}{\sigma_x\ \sigma_y}$$

$$\Rightarrow 0.5 = \frac{16}{4 \times \sigma_y}$$

$$\Rightarrow \sigma_y = \frac{16}{4 \times 0.5} = 8$$

Hence, the correct option is (b).

Example 9. If covariance of 10 pairs of items is 7, variance of X is 36 and $\sum (Y - \bar{Y})^2 = 90$. Find the value of r.

(a) 1.389 (b) 0.389 (c) 0.258 (d) 1.635

Sol. (b) Given: $\sum (Y - \bar{Y})^2 = 90$, n = 10, Cov (x, y) = 7 and Var x = 36

Then, $\sigma_x = \sqrt{Var\ x} = \sqrt{36} = 6$ and $\sigma_y = \sqrt{\frac{\sum (Y - \bar{Y})^2}{n}} = \sqrt{\frac{90}{10}} = 3$

Using, $r = \frac{Cov\,(xy)}{\sigma_x * \sigma_y}$

Substituting the values in the above formula, we get

$r = \frac{7}{(6)(3)} = 0.389$

Hence, the correct option is (b).

Example 10. The coefficient of correlation between two variables x and y is 0.4 and their covariance is 10. If the variance of x series is 9, find the standard deviation of y series.

(a) 4.389 (b) 69.445 (c) 8.33 (d) 7.635

Sol. (c) Coefficient of correlation, r = 0.4, Cov (x, y) = 10, var (x) = 9

$\Rightarrow \sigma_x = 3$

As we know that, $\rho(x, y) = \frac{cov(x, y)}{\sigma_x \times \sigma_y}$

$\Rightarrow 0.4 = \frac{10}{\sqrt{9} \times \sigma_y}$

$\Rightarrow \sigma_y = \frac{100}{4 \times 3} = 8.33$

Hence, the correct option is (c).

CHARACTERISTICS OF THE CORRELATION COEFFICIENT

- Correlation coefficient is independent of the units of measurement.
- The correlation coefficient r is independent of change of origin and scale under consideration depending on the sign of scale factors.

That is, if the variables X and Y are replaced by the variables U and V such that U = aX + b and V = cY + d, where a and c are positive constants.
However, if a and c are arbitrary constants, then.

$$r(aX + b, cY + d) = \frac{a \times c}{|a \times c|} r_{XY}$$

'or'

$$r_{UV} = \frac{|a \times c|}{a \times c} r_{XY}$$

Thus, $r_{XY} = r_{UV}$ when a and c have the same signs and $r_{XY} = -r_{UV}$ when a and c have the opposite signs.

The correlation coefficient lies between –1 and 1, including both the limits.

That is, $-1 \le r \le 1$

Example 11. What are the limits of the correlation coefficient ?

(a) No limit (b) –1 and 1, excluding the limits

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

Sol. (d) We know that,

Correlation coefficient (r) lies between the range $-1 \le r \le 1$,

where –1 depicts the strong negative correlation and 1 depicts the strong positive correlation.

Thus, we can conclude that limits of the correlation coefficient are –1 and 1, including the limits.

Hence, the correct answer is option (d) i.e., –1 and 1, including the limits.

Example 12. Correlation coefficient is _____ of the units of measurement.

(a) Dependent (b) Independent (c) Both (d) None

Sol. (b) We know that,

The coefficient of Correlation is a unit-free measure.

Therefore, correlation coefficient is independent of the units of measurement.

Hence, the correct answer is option (b) i.e., Independent.

Example 913 If the correlation between X and Y is r, $U = \frac{X-5}{10}$ and V $\frac{Y\ 7}{2}$ then r_{uv} is

(a) r (b) –r (c) $\frac{r-5}{2}$ (d) $\frac{r-7}{10}$

Sol. (a) Given: $U = \frac{X-5}{10}$ and $V = \frac{Y-7}{2}$

$U = \frac{X}{10} - \frac{5}{10}$ and $V = \frac{Y}{2} - \frac{7}{2}$

Thus, $a = \frac{1}{10}$ and $c = \frac{1}{2}$

Therefore, $r_{uv} = \frac{a \times c}{|a \times c|} \times r_{xy}$

$\Rightarrow r_{uv} = \frac{\frac{1}{10} \times \frac{1}{2}}{\left|\frac{1}{10} \times \frac{1}{2}\right|} \times r_{xy}$

$\Rightarrow r_{uv} = r_{xy} = r$

Hence, the correct option is (a) i.e. r.

Example 14. If r = 0.58, correlation coefficient of u = –5x + 3 and v = y + 2 is _____

(a) 0.58 (b) –0.58 (c) 0.62 (d) None

Sol. (b) We know that the value of correlation coefficient does not change with change in origin and scale.

Also, if the sign of x & y in given equations are opposite, then sign of r also changes.

Here, sign of x,y are opposite so r_{uv} = –0.58

Hence, option (b) is correct.

Example 15. The coefficient of correlation between X and Y is 0.6. U and V are two variables defined as $U = \frac{X-3}{2}, V = \frac{Y-2}{3}$, then the coefficient of correlation between U and V is

(a) 0.6 (b) 0.4 (c) 0.8 (d) –0.6

Sol. (a) Given: Coefficient of correlation between X and Y = 0.6

Also, $U = \frac{X-3}{2}, V = \frac{Y-2}{3}$

We know that,

Correlation coefficient (Karl Pearson's) is independent of the change of scale and origin,

So, r(U, V) = r(X, Y) = 0.6

[Because X & Y have same sign]

Hence, the correct option is (a) i.e. 0.6.

PRACTICE QUESTIONS (PART A)

1. 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 shown

   (a) in kg (b) in cm (c) in kg and cm (d) free from any unit

2. The correlation coefficient between two variables X and Y is found to be 0.4. What is the correlation coefficient between 2X and (– Y)?

   (a) 0.4 (b) –0.8 (c) 0.8 (d) –0.4

3. If r = 0.28, cov (x, y) = 7.6, V(x) = 9 then σ_y =

   (a) 8.75 (b) 9.04 (c) 6.25 (d) None

4. If for two variables A and B, the covariance, variance of A, and variance of B are 60, 25, and 400 respectively, what is the value of the correlation coefficient?

   (a) 0.2 (b) 0.5 (c) 0.4 (d) 0.6

5. Coefficient of correlation between X and Y is 0.6. If both X and Y are multiplied by –1, then resultant correlation coefficient is

   (a) 0.6 (b) –0.6 (c) 1/0.6 (d) None

6. The covariance between two variable X and Y is 8.4 and their variances are 25 and 36 respectively. Calculate Karl Pearson's coefficient of correlation between them.

   (a) 0.82 (b) 0.28 (c) 0.01 (d) 0.09

7. If correlation coefficient between x and y is 0.5, then the correlation coefficient between 2x–3 and 3–5y is **(Dec 2019)**

   (a) 0.5 (b) –0.5 (c) 2.5 (d) –2.5

8. The coefficient of correlation between x & y when Cov(x, y) = –16.5, Var(x) = 2.89, Var(y) = 100 is

   (a) –0.97 (b) 0.97 (c) 0.89 (d) –0.89

9. The coefficient of correlation between X and Y series is –0.38. The linear relation between U and V are 3X + 5U = 3 and –8Y – 7V = 44, what is coefficient between U & V ?

(a) 0.38 (b) –0.38 (c) 0.40 (d) None of them

Answer Key

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

COMPUTING THE CORRELATION COEFFICIENT

Various formulas for computing the correlation coefficient between the two variables X and Y:

- The correlation coefficient, r, between two variables X and Y is given by

$$r = \frac{\text{cov}(X, Y)}{\sqrt{\text{Var } X}.\sqrt{\text{var } Y}}$$

If the variable x takes on the values $x_1, x_2, \ldots, x_n$ and the variable y takes on the values $y_1, y_2, \ldots, y_n$ then we have

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

$$\text{Var}(X) = \frac{\Sigma(x-\bar{x})^2}{n} \text{ Or } \frac{\Sigma x^2}{n} - (\bar{x})^2$$

$$\text{VarY} = \frac{\Sigma(y-\bar{y})^2}{n} \text{ Or } \frac{\Sigma y^2}{n} - (\bar{y})^2$$

We obtain another formula for r:

$$r = \frac{\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{\sqrt{\sum\left(x-\bar{x}\right)^2 \cdot \sum\left(y-\bar{y}\right)^2}}$$

Example 16. If the sum of the product of the deviation of X and Y from their means is zero, then the correlation coefficient between X and Y is **(July 2021)**

(a) Zero (b) Positive (c) Negative (d) 10

Sol. (a) According to the question, we have

$$\sum\left(X-\bar{X}\right)\left(Y-\bar{Y}\right) = 0$$

Thus, correlation coefficient, $r = \dfrac{\sum\left(X-\bar{X}\right)\left(y-\bar{Y}\right)}{\sqrt{\sum\left(X-\bar{X}\right)^2 \cdot \sum\left(Y-\bar{Y}\right)^2}}$

$$= \frac{0}{\sqrt{\sum\left(X-\bar{X}\right)^2 \cdot \sum\left(Y-\bar{Y}\right)^2}} = 0$$

Hence, the correct option is (a).

Example 17. Calculate coefficient of correlation from the following results: n = 10, $\sum X = 100, \sum Y = 150, \sum (X-10)^2 = 180 \sum (Y-15)^2 = 215$ and $\sum (X-10)(Y-15) = 60$

(a) 0.463 (b) 2.15 (c) 0.305 (d) –0.7618

Sol. (c) Given: $n = 10, \sum X = 100, \sum Y = 150, \sum (X-10)^2 = 180$

$\sum (Y-15)^2 = 215$ and $\sum (x-10)(y-15) = 60$

$$r = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum (X-\bar{X})^2 \sum (Y-\bar{Y})^2}} = \frac{60}{\sqrt{180 \times 215}} = \frac{60}{196.723} = 0.305$$

Therefore, the coefficient of correlation is 0.305.

Hence, the correct option is (c).

Example 18. Find the covariance between X and Y for the following data:

X:	66	67	68	69	70	71	72
Y:	68	67	70	70	69	70	69.

(a) 0.67 (b) 2.4 (c) 0.25 (d) 1.6

Sol. (a) $\bar{x} = \frac{\sum x}{n} = \frac{483}{7} = 69$

$\bar{y} = \frac{\sum y}{n} = 69$

x	66	67	68	69	70	71	72
y	68	67	70	70	69	70	69
$x - \bar{x}$	-3	-2	-1	0	1	2	3
$y - \bar{y}$	-1	-2	1	1	0	1	0
$(x-\bar{x})(y-\bar{y})$	3	4	1	0	0	2	0

$\sum (x-\bar{x})(y-\bar{y}) = 3+4+1+2 = 10$

$\sum (x-\bar{x})^2 = 9+4+1+0+1+4+9 = 28$

$\sum (y-\bar{y})^2 = 1+4+1+1+0+1+0 = 8$

Thus, $r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \cdot \sum (y-\bar{y})^2}} = \frac{10}{\sqrt{28.8}} = 0.67$

Hence, the correct option is (a).

Example 19. The coefficient of correlation between X and Y series from the following data is

	X series	Y series
Number of pairs of observations	15	15
Arithmetic mean	25	18
Standard Deviation	3.01	3.03
Sum of squares of dev. from mean	136	138

Sum of the product of the deviations of X and Y series from their respective means = 122

(a) 0.89 (b) 0.99 (c) 0.69 (d) 0.91

Sol. (a) Given that;

$\sigma_x = 3.01$

$\sigma_y = 3.03$

$\Sigma x^2 = 136$

$\Sigma y^2 = 138$

$\Sigma xy = 122$

where, $x = X - \bar{X}; y = Y - \bar{Y}$

Substituting the values in above formula, we get

$$r = \frac{\sum xy}{\sum x^2 \times \sum y^2} = \frac{122}{\sqrt{136 \times 138}} = 0.89$$

Hence, the correct option is (a).

Example 20. Given the following information: $r = 0.8$, $\Sigma xy = 60$, $\sigma_y = 2.5$ and $\Sigma x^2 = 90$. where x and y are the deviations from the respective means, find the number of items (n).

(a) 8 (b) 11 (c) 14 (d) 10

Sol. (d) From the question, we know that

$r = 0.8$, $\Sigma xy = 60$, $\sigma_y = 2.5$ and $\Sigma x^2 = 90$

The formula for the standard deviation is given by:

$$\sigma_y = \sqrt{\frac{\sum y^2}{n}}$$

From this, we have

$$2.5 = \sqrt{\frac{\sum y^2}{n}}$$

Squaring both sides,

$6.25n = \Sigma x^2$

As we know that the formula for the coefficient of correlation r is: $r = \frac{\sum xy}{\sqrt{\left|\sum x^2 \sum y^2\right|}}$

Substituting the given values, we get

$$0.8 = \frac{60}{\sqrt{(90)(6.25n)}}$$

$$\Rightarrow \sqrt{n} = \frac{60}{0.8\sqrt{|(90)(6.25)|}}$$

$$\Rightarrow \sqrt{n} = \frac{600\sqrt{10}}{8 \times 3 \times 25}$$

$$\Rightarrow \sqrt{n} = \sqrt{10}$$

$$\Rightarrow n = 10$$

Hence, the correct option is (d).

ANOTHER FORMULA

Also,

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

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

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

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

FORMULAE OF CORRELATION COEFFICIENT

- $r = \frac{Cov(X,Y)}{\sqrt{VarX} \cdot \sqrt{VarY}}$
- $r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$
- $r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$

Example 21. For the set of observations {(1, 2), (2, 5), (3, 7), (4, 8), (5, 10)}, the value of Karl-pearsons' coefficient of correlation is approximately given by **(Jan 2021)**

(a) 0.755 (b) 0.655 (c) 0.525 (d) 0.985

Sol. (d) According to the data of observations given, we have

X	Y	XY	X^2	Y^2
1	2	2	1	4
2	5	10	4	25
3	7	21	9	49
4	8	32	16	64
5	10	50	25	100
$\Sigma X = 15$	$\Sigma Y=32$	$\Sigma XY=115$	$\Sigma X^2=55$	$\Sigma Y^2=242$

Coefficient of correlation

$$= r = \frac{N\sum XY - \sum X \cdot \sum Y}{\sqrt{N\sum X^2 - \left(\sum X\right)} \cdot \sqrt{N\sum Y^2 - \left(\sum Y\right)^2}}$$

$$= \frac{5 \times 115 - 15 \times 32}{\sqrt{5 \times 55 - (15)^2} \cdot \sqrt{5 \times 242 - (32)^2}}$$

$$= \frac{575 - 480}{\sqrt{275 - 225} \cdot \sqrt{1210 - 1024}}$$

$$= \frac{+95}{\sqrt{50} \cdot \sqrt{186}} = +0.985$$

Therefore, the required value is 0.985.

Hence, the correct option is (d).

Example 22. From the following data, calculate Karl Pearson's coefficient of correlation:

Height of Fathers (in inches):	66	68	69	72	65	59	62	67	61	71
Height of Sons (in inches):	65	64	67	69	64	60	59	68	60	64

(a) 0.463 (b) 2.15 (c) 0.195 (d) 0.829

Sol. (d) Let Height of Fathers (in inches) be x

And Height of Sons (in inches) be y

Now, the required table is:

x	y	$u = x - 66$	$v = y - 64$	uv	u^2	v^2
66	65	0	1	0	0	1
68	64	2	0	0	4	0
69	67	3	3	9	9	9
72	69	6	5	30	36	25
65	64	-1	0	0	1	0
59	60	-7	-4	28	49	16
62	59	-4	-5	20	16	25
67	68	1	4	4	1	16
61	60	-5	-4	20	25	16
71	64	5	0	0	25	0
$\left(\sum_{i=1}^{N} x_i\right)$ = 660	$\left(\sum_{i=1}^{N} y_i\right)$ = 640	$\Sigma u = 0$	$\Sigma v = 0$	Σuv =111	(Σu^2) = 166	(Σv^2) = 108

Here, $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{n} = \frac{660}{10} = 66$

$\bar{y} = \frac{\sum_{i=1}^{N} y_i}{n} = \frac{640}{10} = 64$

Now, Correlation coefficient, $r = \frac{n \Sigma uv - \Sigma u \Sigma v}{\sqrt{n \Sigma u^2 - (\Sigma u)^2}\sqrt{n \Sigma v^2 - (\Sigma v)^2}}$

Substituting the values, we get

$r = \frac{111}{\sqrt{166 \times 108}} = 0.829$

Hence, the correct option is (d).

Example 23. Calculate Karl Pearson's coefficient of correlation between the variables X and Y using the following data:

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

(a) 2.463 (b) 2.15 (c) –0.915 (d) –0.229

Sol. (d) According to the data given in the question, we have

x	y	xy	x^2	y^2
5	1	5	25	1
4	2	8	16	4
3	4	12	9	16
2	5	10	4	25
1	2	2	1	4
5	4	20	25	16
1	8	8	1	64
7	2	14	49	4
9	6	54	81	36
8	2	16	64	4
$\Sigma x = 45$	$\Sigma y = 36$	$\Sigma xy = 149$	$\Sigma x^2 = 275$	$\Sigma y^2 = 174$

Then, we have to use, $r = \dfrac{\sum xy - \dfrac{(\sum x)(\sum y)}{N}}{\sqrt{\sum x^2 - \dfrac{(\sum x)^2}{N}}\sqrt{\sum y^2 - \dfrac{(\sum y)^2}{N}}}$

$$r = \frac{149 - \dfrac{(45)(36)}{10}}{\sqrt{275 - \dfrac{(45)^2}{10}}\sqrt{174 - \dfrac{(36)^2}{10}}}$$

$$r = \frac{-13}{\sqrt{3219}} = -0.229$$

Therefore, the Karl Pearson's coefficient of correlation between the variables X and Y is – 0.229.

Hence, the correct option is (d).

Example 24. The coefficient of correlation between x and y where

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

is **(ICAI)**

(a) 0.65 (b) 0.68 (c) 0.73 (d) 0.758

Sol. (a) Given, n = 5

To find correlation coefficient, we will prepare the table as follows:

x	y	$u = x - \bar{x}$ $= x - 64$	$v = y - \bar{y}$ $= y - 64$	uv	u^2	v^2
64	57	0	– 7	0	0	49
60	60	– 4	– 4	16	16	16
67	73	3	9	27	9	81
59	62	– 5	– 2	10	25	4
69	68	5	4	20	25	16
$\left(\sum_{i=1}^{N} x_i\right)$ $=319$	$\left(\sum_{i=1}^{N} y_i\right)$ $=320$	$\Sigma u = -1$	$(\Sigma v) = 0$	(Σuv) $= 73$	(Σu^2) $= 75$	(Σv^2) $=166$

Here,

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{n} = \frac{319}{5} = 63.8 \text{ (Let's assume mean as 64)}$$

$$\bar{y} = \frac{\sum_{i=1}^{N} y_i}{n} = \frac{320}{5} = 64$$

Correlation coefficient, $r = \dfrac{n\Sigma uv - \Sigma u \Sigma v}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\sqrt{n\Sigma v^2 - (\Sigma v)^2}} = 573$

$$= \frac{5(73) - (-1)(0)}{\sqrt{5(75) - (-1)^2}\sqrt{5(166) - (0)^2}}$$

$$= \frac{365 - 0}{\sqrt{375 - 1}\sqrt{830 - 0}}$$

$= 0.655$

Hence, the correct answer is option (a) i.e., 0.655 .

Example 25. What would be the correlation between u and v? **(ICAI)**

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

(a) – 0.6 (b) –0.3224 (c) –0.93 (d) 0.93

Sol. (c)

u	$x = u - \bar{u}$ $= u - 21$	v	$y = v - \bar{v}$ $= v + 42$	x^2	y^2	xy
10	-11	-24	18	121	324	-198
15	-6	-36	6	36	36	-36
25	4	-42	0	16	0	0
20	-1	-48	-6	1	36	-6
35	14	-60	-18	196	324	-252
$\Sigma u = 105$	$\Sigma x = 0$	$\Sigma v = -210$	$\Sigma y = 0$	$\Sigma x^2 = 370$	$\Sigma y^2 = 720$	$\Sigma xy = -492$

Here, $n = 5$

Then $\bar{u} = \dfrac{\sum u}{n} = \dfrac{105}{5} = 21$

And $\bar{v} = \dfrac{\sum v}{n} = \dfrac{-210}{5} = -42$

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

$$= \frac{5(-492) - (0)(0)}{\sqrt{\left(5(370) - (0)^2\right)\left(5(720) - (0)^2\right)}} - 0.93(\text{approx})$$

Hence the correct option is (c).

Example 26. Calculate coefficient of correlation from the following data:

X:	10,000	20,000	30,000	40,000	50,000	60,000	70,000
Y:	0.3	0.5	0.6	0.8	1.0	1.1	1.3

(a) 9.7289×10^{-1} (b) 12.7289

(c) -12.7289×10^{-1} (d) -97.289

Sol. (a) Assume mean for x series is, $\dfrac{270000}{6} = 45000.$

And assume mean of y series is, $\dfrac{5.3}{6} = 0.8833 \approx 0.9$

The required table is.

x	y	$u = x - 45000$	$v = y - 0.9$	uv	u^2	v^2
10,000	0.3	-35,000	-0.6	21000	1225000000	0.36
20,000	0.5	-25,000	-0.4	10000	625000000	0.16
30,000	0.6	-15,000	-0.3	4500	225000000	0.09
40,000	0.8	-5,000	-0.1	500	25000000	0.01
50,000	1.0	5,000	0.1	500	25000000	0.01
60,000	1.1	15,000	0.2	3000	225000000	0.04
70,000	1.3	25,000	0.4	10000	625000000	0.16
		$\Sigma u = -35,000$	$\Sigma v = -0.7$	$\Sigma uv = 49500$	$\Sigma u^2 = 2975000000$	$\Sigma v^2 = 0.83$

Now, Correlation coefficient, $r = \dfrac{n\Sigma uv - \Sigma u \Sigma v}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\sqrt{n\Sigma v^2 - (\Sigma v)^2}}$

Then, we get

$$r = \frac{7 \times 49500 - (-35000) \times (-0.7)}{\sqrt{7 \times 2975000000 - (-35000)^2}\sqrt{7 \times (0.83) - (-0.7)^2}} = 9.7289 \times 10^{-1}.$$

Hence, the correct option is (a).

Example 27. Find Karl Pearson's coefficient of correlation between the age and the playing habits of the people from the following information

Age groups (in years)	No. of people	No. of players
15 and less than 20	200	150
20 and less than 25	270	162
25 and less than 30	340	170
30 and less than 35	360	180
35 and less than 40	400	180
40 and less than 45	300	120

(a) 0.463 (b) −0.9395 (c) 0.15 (d) 0.7618

Sol. (b)

Age groups (in years)	Mid Value (x)	No. of people	No. of players	Percentage of Players (%)	(y)	$dx = x - 30$	$dy = x - 54$	dxdy	dx^2	dy^2
15 - 19	17	200	150	$\frac{150}{200} \times 100 = 75\%$	75	-13	21	-273	169	441
20 - 24	22	270	162	$\frac{162}{270} \times 100 = 60\%$	60	-8	6	-48	64	36
25 - 29	27	340	170	$\frac{170}{340} \times 100 = 50\%$	50	-3	-4	12	9	16
30 - 34	32	360	180	$\frac{180}{360} \times 100 = 50\%$	50	2	-4	-8	4	16
35 - 39	37	400	180	$\frac{180}{400} \times 100 = 45\%$	45	7	-9	-63	49	81
40 - 44	42	300	120	$\frac{120}{300} \times 100 = 40\%$	40	12	-14	-168	144	196
	$\Sigma x = 177$				$\Sigma y = 320$	$\Sigma dx = -3$	$\Sigma dy = -4$	$\Sigma dxdy = -548$	$\Sigma dx^2 = 439$	$\Sigma dy^2 = 786$

Now we have to find the Mean values of X and Y,

So, mean value of $\bar{x} = \frac{\sum x}{N} = \frac{177}{6} = 29.5 \approx 30$, then assumed mean is 30.

Now the Mean value of $\bar{y} = \frac{\Sigma y}{N} = \frac{320}{6} = 53.33 \approx 54$, then assumed mean is 54.

As we know that,

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

$$r = \frac{-548 - \frac{(-3) \times (-4)}{6}}{\sqrt{439 - \frac{(-3)^2}{6}} \sqrt{786 - \frac{(-4)^2}{6}}} = \frac{550}{\sqrt{439 - \frac{3}{2}} \sqrt{786 - \frac{8}{3}}} = -\frac{22\sqrt{4935}}{1645} = -0.9395$$

Hence, the correct option is (b).

PRACTICE QUESTIONS (PART B)

1. The coefficient of correlation between X and Y if cov (x, y) = 16.5, Var (X) = 8.25 and Var(Y) = 33.

   (a) 0 (b) 1 (c) 2.5 (d) 4.3

2. The coefficient of correlation between X and Y for the following data:

   $n = 25, \sum X = 55, \sum Y = 40, \sum X^2 = 385, \sum Y^2 = 192, \sum XY = 185$

   (a) 0.068 (b) –0.068 (c) 0.186 (d) None of these

3. Find the covariance between X and Y, given that $\sum X = 60, \sum Y = 90, \sum XY = 574$ and n = 10.

   (a) 1.5 (b) 2.7 (c) 3.4 (d) None of these

4. Take 200 and 150 respectively as the assumed mean for X and Y series of 11 values, then dx = X – 200, dy = Y – 150, $\Sigma dx = 13$, $\Sigma dx^2 = 2667$, $\Sigma dy = 42$, $\Sigma dy^2 = 6964$; $\Sigma dxdy = 3943$

   The value of r is

   (a) 0.77 (b) 0.98 (c) 0.92 (d) 0.82

5. What is the coefficient of correlation between the ages of husbands and wives from the following data? **(ICAI)**

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

6. Calculate the covariance between X and Y for the following data:

X:	1	2	3	4	5	6	7	8	9	10
Y:	6	9	6	7	8	5	12	3	17	1

   (a) 0.5 (b) 0.75 (c) 0.4 (d) None of these

7. Find correlation between age and blindness using following data:

Age (years)	50 - 55	55-60	60-65	65-70	70-75	75-80
Number of Persons	25000	20000	15000	12000	10000	8000
Number of Blind	200	150	90	48	30	12

   (a) 0.463 (b) 0.959 (c) –0.9894 (d) 0.7618

Answer Key

1. (b) 2. (b) 3. (c) 4. (c) 5. (b) 6. (c) 7. (b)

SPEARMAN'S COEFFICIENT OF RANK CORRELATION

- In situations where we have a series of items that cannot be measured numerically, such as qualities like beauty, intelligence, leadership ability, honesty, etc., we can still compare the rankings of individuals within a group.

- But the individuals in the group can be arranged in order thereby obtaining for each individual a number indicating its rank in the group.
- If we have a group of individuals ranked according to two different qualities, it is natural to ask the following question: "Is there an association between the rankings?"
- To answer this question, we need to use a formula known as Spearman's coefficient of rank correlation.
- The Spearman's correlation coefficient is nothing but Karl Pearson's correlation coefficient between the ranks and is interpreted in much the same way.

The Spearman's correlation coefficient, denoted by "ρ" will range from – 1 to + 1. A value of + 1 indicates perfect association for identical rankings and a value of – 1 indicates perfect association for reverse rankings.

Rank R_1	Rank R_2
1	7
2	8
3	6
4	1
6	2
8	4
7	3
5	5

In this example, we have two sets of rankings, R_1 and R_2. The values in the table represent the ranks assigned to the individuals for each quality. By applying Spearman's coefficient of rank correlation formula, we can determine the extent of association between these rankings.

COMPUTING THE RANK CORRELATION COEFFICIENT:

We shall consider the following three cases to compute the rank correlation coefficient.

Case I: When actual ranks are given

In this case the following steps are involved:

1. Compute D, the difference between the two ranks given to each individual.
2. Compute D^2 and obtain the ΣD^2.

Apply the formula: $\rho = 1 - \frac{6\sum D^2}{n(n^2 - 1)}$ where, n is the number of observations.

Example 28. Determine Spearman's rank correlation coefficient from the given data $\Sigma D^2 = 30$, $N = 10$. **(June 2019)**

(a) R = 0.82 (b) R = 0.32 (c) R = 0.40 (d) None of these

Sol. (a) Here, $\Sigma D^2 = 30$, $N=10$

Spearman's rank correlation is;

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)} = 1 - \frac{6 \times 30}{10(10^2 - 1)} = 1 - \frac{180}{990} = 1 - \frac{2}{11} = \frac{9}{11} = 0.82$$

Therefore, R = 0.82

Hence, the correct answer is option (a), i.e, R = 0.82

Example 29. If the sum of squares of the rank differences of 9 pairs of values is 80, find the correlation coefficient between them.

(a) 0.33 (b) 0.39 (c) –0.33 (d) –0.039

Sol. (a) Given,

Sum of squares of the rank differences = $\Sigma D^2 = 80$

Number of pairs n = 9

We know,

$$\rho = 1 - \frac{6\sum D^2}{n(n^2 - 1)}$$

$$\rho = 1 - \frac{6(80)}{9(9^2 - 1)} = 1 - \frac{480}{720} = 0.33 \text{ (approax)}$$

The correlation coefficient between them is 0.33.

Hence, the correct option is (a).

Example 30. Ten competitors in following in a beauty contest are ranked by two judges in the following order:

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

Calculate the Spearman's rank correlation coefficient.

(a) 0.6364 (b) 1.395 (c) –0.9894 (d) 0.7618

Sol. (a)

I Judge R_1	II Judge R_2	$D = R_1 - R_2$	D^2
1	6	-5	25
6	4	2	4
5	9	-4	16
10	8	2	4
3	1	2	4
2	2	0	0
4	3	1	1
9	10	-1	1
7	5	2	4
8	7	1	1
			$\sum D^2 = 60$

So, $\Sigma D^2 = 60$ and n = 10

Now we have to use, $\rho = 1 - \frac{6\sum D^2}{n(n^2-1)}$

$\rho = 1 - \frac{6(60)}{10(10^2-1)} = \frac{7}{11} = 0.6364$

The Spearman's rank correlation coefficient is 0.6364.
Hence, the correct option is (a).

Example 31. Rankings of 10 trainees at the beginning and at the end of a certain course are given as follows:

Trainees:	A	B	C	D	E	F	G	H	I	J
Rank at the beginning:	1	6	3	9	5	2	7	10	8	4
Rank at the end:	6	8	3	7	2	1	5	9	4	10

Calculate spearman's rank correlation coefficient

(a) 0.636 (b) 0.394 (c) –0.394 (d) 0.762

Sol. (b)

Trainees	Rank at the beginning R_1	Rank at the end R_2	$D = R_1 - R_2$	D^2
A	1	6	–5	25
B	6	8	–2	4
C	3	3	0	0
D	9	7	2	4
E	5	2	3	9
F	2	1	1	1
G	7	5	2	4
H	10	9	1	1
I	8	4	4	16
J	4	10	–6	36
				$\Sigma D^2 = 100$

So, $\Sigma D^2 = 100$ and $n = 10$

Now we have to use, $\rho = 1 - \frac{6\sum D^2}{n(n^2-1)}$

$\rho = 1 - \frac{6(100)}{10(10^2-1)} = \frac{13}{33} = 0.3939$

The Spearman's rank correlation coefficient is 0.394.
Hence, the correct option is (b).

Example 32. Ten competitors in a beauty contest are ranked by three judges in the following order:

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

Use the rank correlation method to determine which pair of judges has the nearest approach to common taste in beauty

(a) Judge I and Judge II (b) Judge I and Judge III

(c) Judge II and Judge III (d) None of these

Sol. (c)

I Judge R_1	II Judge R_1	III Judge R_1	$D_1 = R_1 - R_2$	$D_2 = R_1 - R_3$	$D_3 = R_2 - R_3$	D_1^2	D_2^2	D_3^2
1	4	6	−3	−5	−2	9	25	4
4	8	7	−4	−3	1	16	9	1
8	7	1	1	−7	6	1	49	36
9	5	8	4	1	−3	16	1	9
6	9	10	−3	−4	−1	9	16	1
10	6	5	4	5	1	16	25	1
7	10	9	−3	−2	1	9	4	1
3	2	2	1	1	0	1	1	0
2	3	3	−1	−1	0	1	1	0
5	1	4	4	1	−3	16	1	9
						$\sum D_1^2 = 94$	$\sum D_2^2 = 132$	$\sum D_3^2 = 62$

Now we have, $\sum D_1^2 = 94$, $\sum D_2^2 = 132$, $\sum D_3^2 = 62$ and $n = 10$

Then correlation coefficient between Judge I and Judge II is $p_1 = 1 - \frac{6\sum D_1^2}{n(n^2-1)}$

$$p_1 = 1 - \frac{6(94)}{10(10^2-1)} = \frac{71}{165} = 0.4303$$

Then correlation coefficient between Judge I and Judge III is $p_2 = 1 - \frac{6\sum D_2^2}{n(n^2-1)}$

$$p_2 = 1 - \frac{6(132)}{10(10^2-1)} = \frac{1}{5} = 0.2$$

Then correlation coefficient between Judge II and Judge III is $p_3 = 1 - \frac{6\sum D_3^2}{n(n^2-1)}$

$$p_3 = 1 - \frac{6(62)}{10(10^2-1)} = \frac{103}{165} = 0.6242$$

The Judge II and the Judge III have the nearest approach to common taste in beauty since the correlation coefficient is maximum in that case.

Hence, the correct option is (c).

Example 33. In a bivariate data of n pairs of observations, the sum of squares of differences between the ranks of observed values of two variables is 231 and the rank correlation coefficient is – 0.4. Find the value of n.

(a) 10 (b) 13 (c) $-5 \pm \sqrt{74}i$ (d) 5

Sol. (a) Given: $\Sigma D^2 = 231$ and $p = -0.4$

Using formula, $\rho = 1 - \frac{6\sum D^2}{n\left(n^2-1\right)}$

$$\Rightarrow -0.4 = 1 - \frac{6(231)}{n(n^2-1)}$$

$$\Rightarrow -1.4 = -\frac{6(231)}{n(n^2-1)}$$

$$\Rightarrow -1.4n\,(n^2-1) = -1386$$

$$\Rightarrow n\,(n^2-1) = 990$$

$$\Rightarrow n\,(n^2-1) = 10 \times 900 = 10\,(10^2-1)$$

$$\Rightarrow n = 10$$

Therefore, the value of n is 10.

Hence, the correct option is (a).

Example 34. The coefficient of rank correlation of the marks obtained by 10 students in Statistics and Accountancy was found to be 0.2. It was later discovered that the difference in ranks in the two subjects obtained by one of the students was wrongly taken as 9 instead of 7. Find the correct value of coefficient of rank correlation.

(a) 0.3334 (b) 0.3939 (c) –0.3334 (d) –0.3939

Sol. (b) Given: $\rho = 0.2$, $n=10$

As we know, $\rho = 1 - \frac{6\sum D^2}{n\left(n^2-1\right)}$

$$\Rightarrow 0.2 = 1 - \frac{6\sum D^2}{10(10^2-1)}$$

$$\Rightarrow 0.8 = \frac{6\sum D^2}{990}$$

$$\Rightarrow \sum D^2 = \frac{0.8 \times 990}{6} = 132$$

Now Correct $\Sigma D^2 = 132 - 9^2 + 7^2 = 100$

Again using, $\rho = 1 - \frac{6\sum D^2}{n\left(n^2-1\right)}$

$$p = 1 - \frac{6(100)}{10(10^2-1)} = 1 - \frac{60}{99} = \frac{13}{33} = 0.3939$$

Therefore, the correct value of coefficient of rank correlation is 0.3939.

Hence, the correct option is (b).

CASE 2:

When actual ranks are not given:

- Sometimes we are given the actual bivariate data on two variables and not the ranks. In such situations, it is necessary to assign the ranks.
- Ranks can be assigned by taking either the highest value as 1 or the lowest value as 1. The next highest or the next lowest value is given rank 2 and so on.
- But whether we start with the lowest value or the highest value, we must follow the same method in case of both the variables.

Example 35. The marks obtained by 9 students in Mathematics and Accountancy are as follow:

Marks in Mathematics (X):	30	33	45	23	8	49	12	4	31
Marks in Accountancy (Y):	35	23	47	17	10	43	9	6	28

Calculate Spearman's rank correlation coefficient.

(a) 0.4 (b) 0.39 (c) –0.34 (d) 0.9

Sol. (d) Here we does not have ranks so we have to assign ranks first,

Marks in Mathematics (X)	Marks in Accountancy (Y)	Rank of (X) R_1	Rank of (Y) R_2	$D = R_1 - R_2$	D^2
30	35	5	3	2	4
33	23	3	5	–2	4
45	47	2	1	1	1
23	17	6	6	0	0
8	10	8	7	1	1
49	43	1	2	–1	1
12	9	7	8	–1	1

Marks in Mathematics (X)	Marks in Accountancy (Y)	Rank of (X) R_1	Rank of (Y) R_2	$D = R_1 - R_2$	D^2
4	6	9	9	0	0
31	28	4	4	0	0
					$\sum D^2 = 12$

Now we have $\Sigma D^2 = 12$, and $n = 9$

Using $\rho = 1 - \frac{6\sum D^2}{n\left(n^2-1\right)}$

$$p = 1 - \frac{6(12)}{9(9^2 - 1)} = \frac{9}{10} = 0.9$$

Therefore, the spearman's rank correlation coefficient is 0.9.
Hence, the correct option is (d).

WHEN RANKS ARE EQUAL:

- If there are two or more items with the same rank in either series, then it is customary to assign common rank to each repeated item.
- The common rank is the average of the ranks which these items would have got if they were different from each other and the next item will get the rank next to the rank used in computing the common rank.
- For example, suppose there are two items at rank 4. In this case, the common rank assigned to each item would be 4.5.
- The next item in line would then be assigned rank 5.
- Similarly, if there are three items at rank 7, the common rank assigned to each item will be 7.
- The next rank to be assigned will be 10.
- When equal ranks are assigned to certain items, an adjustment is made in the formula for calculating the Spearman's rank correlation coefficient. This adjustment involves adding a correction factor for each repeated item in both series.
- This correction factor is to be added for each repeated item in both the series. The formula can thus be written as follows:

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

Example 36. Calculate Spearman's coefficient of rank correlation from the following data:

X:	57	16	24	65	16	16	9	40	33	48
Y:	19	6	9	20	4	15	6	24	13	13

(a) 0.7515 (b) 0.3009 (c) -0.3674 (d) 0.7899

Sol. (a)

(X)	(Y)	Rank of (X) R_1	Rank of (Y) R_2	$D = R_1 - R_2$	D^2
57	19	2	3	–1	1
16	6	8	8.5	–0.5	0.25
24	9	6	7	–1	1
65	20	1	2	–1	1
16	4	8	10	–2	4
16	15	8	4	4	16
9	6	10	8.5	1.5	2.25
40	24	4	1	3	9
33	13	5	5.5	–0.5	0.25
48	13	3	5.5	–2.5	6.25
					$\sum D^2 = 41$

Now we have, $\sum D^2 = 41$ and $n = 10$

Using, $\rho = 1 - \frac{6\sum D^2}{n\left(n^2-1\right)}$ we get

$$p = 1 - \frac{6(41)}{10(10^2 - 1)} = \frac{124}{165} = 0.7515$$

Therefore, the spearman's rank correlation coefficient is 0.7515.

Hence, the correct option is (a).

MERITS AND DEMERITS OF SPEARMAN'S RANK CORRELATION METHOD

Merits

The rank correlation method has the following merits

1. It is easy to understand and simple to apply.
2. The Spearman's rank correlation method is the only method that can be used to find correlation coefficients if we are dealing with data of qualitative characteristics like beauty, intelligence, honesty, etc.
3. This is the only method that can be used where we are given the ranks and not the actual bivariate data on two variables.

DEMERITS:

The rank correlation method has the following limitations:

1. This method cannot be used for finding correlation in the case of bivariate frequency distribution.
2. This method is very difficult to apply when the number of items is more than 30.

PRACTICE QUESTIONS (PART C)

1. If the sum of square of differences of rank is 50 and the number of items is 8, then what is the value of the rank correlation coefficient?

   (a) 0.59 (b) 0.40 (c) 0.36 (d) 0.63

2. If the rank correlation coefficient between marks in Management and Mathematics for a group of students is 0.6 and the sum of the squares of the difference in rank is 66 . Then what is the number of students in the group?

   (a) 9 (b) 10 (c) 11 (d) 12

3. What is the value of the 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

4. The ranks of five participants given by two judges are:

   Participants

	A	B	C	D	E
Judge I	1	2	3	4	5
Judge II	5	4	3	2	1

   (a) 1 (b) 0 (c) -1 (d) 12

5. The coefficient of rank correlation between debenture prices and share prices of a company is found to be 0.143. If the sum of the squares of the differences in ranks is 48, find the value of n.

   (a) 5 (b) 6 (c) 7 (d) None of these

6. Calculate Spearman's rank correlation coefficient between advertisement cost and sales from the following data:

Advertisement cost (000 ₹)	39	65	62	90	82	75	25	98	35	78
Sales (lakhs ₹)	47	53	58	86	62	68	60	91	51	84

   (a) 0.5 (b) 0.734 (c) 0.65 (d) None of these

Answer Key

1. (b) **2.** (b) **3.** (d) **4.** (c) **5.** (c) **6.** (b)

COEFFICIENT OF CONCURRENT DEVIATION

The coefficient of concurrent deviation is a statistical measure used to assess the degree of simultaneous variation or deviation in two variables. It helps in understanding how two sets of data move together or diverge from each other.

The formula for the coefficient of concurrent deviation is as follows

$$r_c = \pm\sqrt{\pm\frac{2c-m}{m}}$$

where, 'n' represents the number of concurrent deviation, 'm' represents the total number of deviations (which must be one less than the number of pairs of x and y values)

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

Example 37. What is the coefficient of concurrent deviations for the following data: (**ICAI**)

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

Sol. (c) Now prepare a table to find the coefficient of concurrence:

If the value increases then we will write +ve and –ve when decreases for both the variables.

Supply	Demand	Deviation sign for X	Deviation sign for Y	Deviation sign for XY
68	65			
43	60	–ve	–ve	+ve
38	55	–ve	–ve	+ve
78	61	+ve	+ve	+ve
66	35	–ve	–ve	+ve
83	75	+ve	+ve	+ve
38	45	–ve	–ve	+ve
23	40	–ve	–ve	+ve
83	85	+ve	+ve	+ve
63	80	–ve	–ve	+ve
53	85	–ve	+ve	–ve

The number of positive signs is the value of c.

Here n = 11, m = n – 1 = 11 – 1 = 10, c = 9

We know that $r_c = \pm\sqrt{\pm\frac{(2c-m)}{m}}$

$$\Rightarrow r_c = \pm\sqrt{\pm\frac{(2(9)-10)}{10}}$$

$$\Rightarrow r_c = \pm\sqrt{\pm\frac{(18-10)}{10}}$$

$$\Rightarrow r_c = \sqrt{\frac{8}{10}}$$

$$\Rightarrow r_c = 0.8944$$

Hence, the correct answer is option (c) i.e., 0.89.

Example 38. For 10 pairs of observations, number of concurrent deviations was found to be 4. What is the value of the coefficient of concurrent deviation? **(ICAI)**

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

Sol. (d) Given that,

No. of Concurrent deviations (c) = 4

Total number of observations (n) = 10

We know that the coefficient of concurrent deviation $\Rightarrow r_c = \pm\sqrt{\pm\frac{2c-m}{m}}$

Here, m = n – 1 = 10 – 1 = 9

Then,

$$\Rightarrow r_c = \pm\sqrt{\pm\frac{2c-m}{m}}$$

$$\Rightarrow r_c = \pm\sqrt{\pm\frac{2\times 4-9}{9}}$$

$$\Rightarrow r_c = \pm\sqrt{\pm\frac{8-9}{9}}$$

$$\Rightarrow r_c = -\sqrt{\frac{1}{9}}$$

$$\Rightarrow r_c = \frac{-1}{3}$$

If (2c – m) < 0, then we take the negative sign both inside and outside the radical sign

Hence, the correct answer is option (d) i.e., –13.

Example 39. The coefficient of concurrent deviation for P pairs of observations was found to be 13. If the number of concurrent deviations was found to be 6, then the value of P is **(ICAI)**

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

Sol. (a) Given:

Coefficient of concurrent deviation $(r_c) = \frac{1}{\sqrt{3}}$

No. of Concurrent deviations (c) = 6

Total number of observations (n) = ρ

We know that, coefficient of concurrent deviation

$$r_c = \pm\sqrt{\pm\frac{2c-m}{m}}$$

Here, m = P – 1

Then,

$$\Rightarrow r_c = \pm\sqrt{\pm\frac{2c - m}{m}}$$

$$\Rightarrow \frac{1}{\sqrt{3}} = \pm\sqrt{\pm\frac{2 \times 6 - m}{m}}$$

$$\Rightarrow \frac{1}{\sqrt{3}} = \pm\sqrt{\pm\frac{12 - m}{m}}$$

Squaring on both sides, we get

$$\frac{1}{3} = \frac{12 - m}{m}$$

$$\Rightarrow m = 3\,(12 - m)$$

$$\Rightarrow m = 36 - 3m$$

$$\Rightarrow 4m = 36$$

$$\Rightarrow m = \frac{36}{4} = 9$$

Here, $m = P - 1$

$\Rightarrow P = m + 1 = 9 + 1 = 10$

Hence, the correct answer is option (a) i.e., 10.

REGRESSION

Regression, in a literal sense, refers to a process of moving backward or returning to an average value. In the context of statistics and mathematics, it involves creating a mathematical equation to predict the value of one variable based on known values of one or more other variables. It is a method for understanding relationships and making forecasts.

For example, consider predicting a student's final exam score based on the number of hours spent studying. Regression analysis could help develop an equation that models this relationship, allowing educators to make predictions about a student's potential performance based on their study hours. The regression equation would provide a tool for forecasting final exam scores using the known variable of study hours.

- **Dependent Variable:** The variable whose value is to be predicted is called the dependent vairbale or explained variable.
- **Independent Variable:** The vairbales which are used to preduct the values of a dependent variable are called independent vairbales or explanatory variables
- **Simple Regression Analysis and Simple Linear Analysis:** The regression analysis confined to the study of only two variables, a dependent varibale and an independent variable, is called simple regression analysis.

When the relationship between the dependent varibale and the independent variable is linear, the technique for prediction is called simple linear regression.

If let say y depends on x, then equation will be:-

$\Rightarrow y = a + bx$

- **METHOD OF LEAST SQUARE: REGRESSION**: Thus, if a line of best fit approximating the given data has the equation Y = a + bX then the method of least squares requires that we must determine constants a and b the so as to minimize

$e_i = \sum (y_i - \widehat{y_1})^2 = \sum (y_i - (a + bx_i)^2$ where $\widehat{y_i} = a + bx_i$

Where, e_i is the error of estimation.

- The equations known as the normal equation for estimating a and b, are given by

$$\sum y = an + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

Multiply first eq. by Σx and second eq. by n, we get

$$\sum x \sum y = an\sum x + b\left(\sum x\right)^2$$

and $n\sum xy = an\sum x + bn\sum x^2$

On subtracting the above equations, we get

$$\sum x\sum y = an\sum x + b\left(\sum x\right)^2$$

$$n\sum xy = an\sum x + nb\sum x^2$$

$$\sum x\sum y - n\sum xy = b\left(\sum x\right)^2 - nb\sum x^2$$

$$\left(n\sum xy - \sum x\sum y\right) = b\left(n\sum x^2 - \left(\sum x\right)^2\right)$$

$$\Rightarrow b = \frac{n\sum xy - \sum x\sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

$$\Rightarrow b = \frac{n\sum xy - \sum x\sum y / n^2}{n\sum x^2 - \left(\sum x\right)^2 / n^2}$$

$$\Rightarrow b = \frac{\frac{\sum xy}{n^x} - \frac{\sum x\sum y}{n^2}}{\frac{\sum x^2}{n^2} - \frac{\left(\sum x\right)^2}{x^2}}$$

$$\Rightarrow b = \frac{\frac{\sum xy}{n} - \frac{\sum x}{n}\frac{\sum y}{n}}{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} = \frac{cov(x, y)}{\sqrt{x}^2}$$

Solving equation simultaneously for a and b, we obtain

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

Hence, the line of best fit approximating n pair of observations (x_1, y_1), (x_2, y_2) ……,

(x_n, y_n) is y on x

$y = a + bx$, where a, b are the constants.

The line of best fit given is called the least squares line of regression of y on x. The constant b is called the regression coefficient of y on x is denoted by b_{yx}.

It measures the change in y corresponding to a unit change in x.

- Thus, b_{yx} represent the slope of the line of regression of Y on X and is given by

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

On the other hand, if we wish to estimate a value of X for a given value of Y, we have to obtain a regression line of X on Y:

$X = bY + a$.

The line of best fit given is called the least squares line of regression of X on Y. The constant b is called the regression coefficient of X on Y is denoted by b_{XY}.

It measures the change in X corresponding to a unit change in Y.

Thus, b_{xy} represent the slope of the line of regression of Y on X and is given by

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{N\sum y^2 - \left(\sum y\right)^2}$$

The equation of the line of regression of Y on X can also be written as $(y - \bar{y}) = b_{yx}(x - \bar{x})$

REGRESSION LINES:

Since, $\sum y = an + b\sum x$

$$\left(\frac{\sum y}{n}\right) = \frac{na + b\sum x}{n}$$

$$\left(\frac{\sum y}{x}\right) = \frac{na}{n} + b\frac{\sum x}{n}$$

$\Rightarrow \bar{y} = a + b\bar{x}$

$\Rightarrow a = \bar{y} - b\bar{x}$ ...(i)

$$\Rightarrow \frac{b_{yx} = b}{y \text{ on } x} = \frac{x\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

$$b_{yx} = \frac{\cos(x, y)}{(\sigma_x)^2}$$

$$b \text{ or } b_{yx} = r\left(\frac{\sigma_y}{\sigma_x}\right) \quad ...(ii)$$

Regression line in

y on x

$y = a + bx$

From (i)

$\Rightarrow y = (\bar{y} - b\bar{x}) + bx$

$\Rightarrow y - \bar{y} - b(x - \bar{x})$

From (ii)

$$\Rightarrow y - \bar{y} = r\left(\frac{\sigma_y}{\sigma_x}\right)(x - \bar{x})$$

Regression line in correlation coefficient

$$\frac{r}{\sigma_x} = \frac{cov(x, y)}{\sigma_x \sigma_y (\sigma_x)}$$

$$\frac{r}{\sigma_x} = \frac{cov(x, y)}{(\sigma_x)^2}\sigma_y$$

$$\Rightarrow \frac{r}{\sigma_x} = \frac{b_{yx}}{\sigma_y}$$

$$\Rightarrow by = \frac{\sigma_y r}{\sigma_x}$$

Example 40. Calculate the regression coefficients from the following information:

$\Sigma X = 50, \Sigma Y = 30, \Sigma XY = 1000, \Sigma X^2 = 3000, \Sigma Y^2 = 1800$ and $n = 10$

(a) 0.497 and 0.309 (b) 0.307 and 0.3009

(c) 0.586 and –0.367 (d) None of these

Sol. (a) Given: $\Sigma X = 50, \Sigma Y = 30, \Sigma XY = 1000, \Sigma X^2 = 3000, \Sigma Y^2 = 1800$, and $n = 10$

Thus,

$$b_{xy} = \frac{n\sum XY - \sum X \sum Y}{n\sum Y^2 - \left(\sum Y\right)^2} = \frac{10(1000) - (50)(30)}{10(1800) - (30)^2} = \frac{10000 - 1500}{18000 - 900} = 0.497$$

Also, $b_{yx} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2} = \frac{10\left(1000\right) - \left(50\right)\left(30\right)}{10\left(3000\right) - \left(50\right)^2} \frac{8500}{30000 - 2500} = \frac{17}{55} = 0.309$

Therefore, the regression coefficients are 0.497 and 0.309.

Hence, the correct option is (a).

Example 31. In the regression equation X on Y, $X = \frac{35}{8} - \frac{2Y}{5}$ b_{xy} is equal to

(a) $-\frac{2}{5}$ (b) $\frac{35}{8}$ (c) $\frac{2}{5}$ (d) $\frac{5}{2}$

Sol. (a) We know that, equation of regression line X on Y is given by $X = a + bY$

where, $b = b_{xy}$ = slope of line and a= intercept of line

Comparing the given equation $X = \frac{35}{8} - \frac{2Y}{5}$ to the standard equation, we get

$$b_{xy} = -\frac{2}{5}$$

Hence, the correct answer is option (a) i.e., -2/5.

Example 41. In the estimation of egression equation of two variables X and Y, the following results were obtained:

$\Sigma X = 900, \Sigma Y = 700, \Sigma X^2 = 6360, \Sigma Y^2 = 2860, \Sigma XY = 3900, n = 10$

Obtain two regression equations.

(a) $x - 1.81y + 6.67 = 0$ and $3.792x - y - 1.28 = 0$

(b) $4x - 1.31y + 6.67 = 0$ and $0.72x - 1.30y - 1.238 = 0$

(c) $1.3230x - 1.32y + 6.67 = 0$ and $0.792x - 1.033y - 1.228 = 0$

(d) $x - 1.28y + 0.4 = 0$ and $0.792x - y - 1.28 = 0$

Sol. (d) Given: $\Sigma X = 900, \Sigma Y = 700, \Sigma X^2 = 6360, \Sigma Y^2 = 2860, \Sigma XY = 390$

As we know,

$$\bar{X} = \frac{\sum X}{n} = \frac{900}{10} = 90 \text{ and } \bar{Y} = \frac{\sum Y}{n} = \frac{700}{10} = 70$$

y on x is $(y - \bar{y}) = b_{yx}(x - \bar{x})$ .....(i)

x on y is $(x - \bar{x}) = b_{xy}(y - \bar{y})$ .....(ii)

Since,

$$b_{yx} = \frac{n\sum XY - \sum X\sum Y}{n\sum X^2 - \left(\sum X\right)^2} = \frac{10(3900) - (900)(700)}{10(6360) - (900)^2} \quad \frac{591000}{63600 - 810000}$$

$$= \frac{985}{1244} = 0.792$$

Therefore the equation of regression y on x is:

$(y - 70) = 0.792(x - 90)$

$y = 0.792x - 1.28$

$0.792x - y - 1.28 = 0$ .....(iii)

Also,

$$b_{xy} = \frac{n\sum XY - \sum X\sum Y}{n\sum Y^2 - \left(\sum Y\right)^2} = \frac{10(3900) - (900)(700)}{10(2860) - (700)^2} = \frac{3900 - 630000}{28600 - 490000} = 1.28$$

Therefore the equation of regression x on y is:

$(x - 90) = 1.381(y - 70)$

$x = 1.28y + 0.4$ .....(iv)

Therefore, the two equations are:

$x = 1.28y + 0.4$ and $0.792x - y - 1.28 = 0$

Hence, the correct option is (d).

Example 42. 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 **(ICAI)**

(a) $y = 1.2x - 15$ (b) $y = 1.2x + 15$

(c) $y = 0.93x - 14.68$ (d) $y = 1.5x - 10.89$

Sol. (c) To find: regression equation y on x i.e. b_{yx}

The data according to the information is as follows :

x	y	$dx = x - \bar{X}$	$dy = y - \bar{Y}$	$dxdy = (x - \bar{X})(y - \bar{Y})$	$dx^2 = (x - \bar{X})^2$
41	28	- 39	- 32	1248	1521
82	56	2	- 4	- 8	4
62	35	- 18	- 25	450	324
37	17	- 43	- 43	1849	1849
58	42	- 22	- 18	396	484
96	85	16	25	400	256

x	y	$dx = x - \bar{X}$	$dy = y - \bar{Y}$	$dxdy = (x - \bar{X})(y - \bar{Y})$	$dx^2 = (x - \bar{X})^2$
127	105	47	45	2115	2209
74	61	– 6	1	– 6	36
123	98	43	38	1634	1849
100	73	20	13	260	400
$\sum x_i = 800$	$\sum y_i = 600$	$\left(\sum dx\right) = 0$	$\left(\sum dy\right) = 0$	$\sum dx\, dy = 8338$	$\sum dx^2 = 8932$

Here,

$$\bar{X} = \frac{\sum x_i}{N} = \frac{800}{10} = 80$$

$$\bar{Y} = \frac{\sum y_i}{N} = \frac{600}{10} = 60$$

We know that,

$$b_{yx} = \frac{N\sum dxdy - \sum dx \sum dy}{N\sum dx^2 - (\sum dx)^2} = \frac{10(8338) - 0}{10(8932) - (0)^2}$$

$= 10\ (8338) - 010\ (8932) - (0)^2$

$= 83388932$

$= 0.9333$

$\bar{Y} = a + b_{yx}\ \bar{X}$

$60 = a + (0.933380)$

$a = 60 - 74.64$

$a = - 14.24$

So, regression equation y on x is given by $y = a + bx$

$y = 0.933x + 14.24$

Hence, the correct answer is option (c) i.e., $y = 0.93x - 14.68$.

Example 43. Following table gives the age of cars of a certain make and annual maintenance costs. Obtain the regression equation for costs related to age:

Age of cars in years:	2	4	6	8
Maintenance cost (in hundred)	10	20	25	30

Also estimate the annual maintenance cost for the ten-year-old car.

(a) $3.792x - y - 1.28 = 0$ and 30

(b) $0.72x - 1.30y - 1.238 = 0$ and 40

(c) $3.25x - y + 5.0 = 0$ and 37.5

(d) $0.792x - y - 1.28 = 0$ and $- 37.5$

Sol. (c) We have to obtain the regression equation for costs related to age. That means cost is dependent on age.

Let age of cares in years be x and Maintenance cost (in hundred) is y

Therefore, the regression equation for costs related to age will be given by $(y - \bar{y}) = b_{yx}(x - \bar{x})$

Age of cars in years (x)	Maintenance cost (in hundred) (y)	xy	x^2	y^2
2	10	20	4	100
4	20	80	16	400
6	25	150	36	625
8	30	240	64	900
$\Sigma x = 20$	$\Sigma y = 85$	$\Sigma xy = 490$	$\Sigma x^2 = 120$	$\Sigma y^2 = 2025$

Here, n = 4

Now, Now, $\bar{X} = \frac{\sum X}{n} = \frac{20}{4} = 5$ and $\bar{Y} = \frac{\sum Y}{n} = \frac{85}{4} = 21.25$

Thus,

$$b_{yx} = \frac{n\sum XY - \sum X \sum Y}{N\sum X^2 - \left(\sum X\right)^2} = \frac{4(490) - (20)(85)}{4(120) - (20)^2} \frac{260}{360 - 400} = \frac{13}{4} = 3.25$$

Therefore, the equation of regression y on x is $(y - 21.25) = 3.25(x - 5)$.

$3.25x - y + 5.0 = 0$

Now, the annual maintenance cost for the ten-year-old car is $3.25(10) - y + 5.0 = 0$

$\Rightarrow y = 37.5$

Hence, the correct option is (c).

Example 44. 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 (ICAI)

(a) $y = 100 + 5x$ (b) $y = 102.60 + 0.4055x$

(c) $y = 89.653 + 0.582x$ (d) $y = 88.758 + 0.562x$

Sol. (b) Let height of father be x and that of son be y.

To find the regression equation y on x, we will find b_{yx}. For that we prepare the table as follows:

Height of father (x)	Height of son (y)	u = x – 171	v = y – 172	uv	u^2
175	173	4	1	4	16
172	172	1	0	0	1
167	171	– 4	– 1	4	16

Height of father (x)	Height of son (y)	$u = x - 171$	$v = y - 172$	uv	u^2
168	171	– 3	– 1	3	9
172	173	1	1	1	1
171	170	0	– 2	0	0
174	173	3	1	3	9
176	175	5	3	15	25
169	170	– 2	– 2	4	4
170	173	– 1	1	– 1	1
(Σx_i) =1714	(Σy_i) =1721	$(\Sigma u) = 4$	$(\Sigma v) = 1$	(Σuv) = 33	(Σu^2) = 82

Here, $\bar{x} = \frac{\Sigma x_i}{N} = \frac{1714}{10} = 171.4$

$\bar{y} = \frac{\Sigma y_i}{N} = \frac{1721}{10} = 172.1$

$b_{yx} = \frac{N\Sigma uv - \Sigma u \Sigma v}{N\Sigma u_i^2 - (\Sigma u)^2} = \frac{10(33) - 4(1)}{10(82) - (4)^2} = \frac{326}{804} = 0.4054$

$\bar{y} = a + b_{yx}\bar{x}$

$172.1 = a + 0.4054(171.4)$

$a = 172.1 - 69.486$

$a = 102.61$

Since, the regression equation is given by $y = a + bx$

Thus, the regression equation y on x is $y = 102.61 + 0.4054x$

Hence, the correct option is (b) i.e. $y = 102.60 + 0.405x$.

Here,

$\bar{x} = \frac{\Sigma x_i}{N} = \frac{1714}{10} = 171.4$

$\bar{y} = \frac{\Sigma y_i}{N} = \frac{1721}{10} = 172.1$

$b_{yx} = \frac{N\Sigma uv - \Sigma u \Sigma v}{N\Sigma u_i^2 - (\Sigma u)^2}$

$= \frac{10(33) - 4(1)}{10(82) - (4)^2} = \frac{326}{804}$

$= 0.4054$

$\bar{y} = a + b_{yx}\bar{x}$

$172.1 = a + 0.4054(171.4)$

$a = 172.1 - 69.486$

$a = 102.61$

PRACTICE QUESTIONS (PART D)

1. If mean of X and Y variables is 20 and 40 respectively and the regression coefficient Y on X is 1.608, then the regression line of Y on X is:

   (a) Y = 1.608 X + 7.84 (b) Y = 1.56 X + 4.84

   (c) Y =1.608 X + 4.84 (d) Y =1.56 X + 7.84

2. If Y = 18X + 5 is the regression line of X on Y, the value of b_{xy} is

   (a) $\frac{5}{18}$ (b) 18 (c) 5 (d) $\frac{1}{18}$

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

   (a) $\frac{2}{3}$ (b) $\frac{-2}{3}$ (c) $\frac{-3}{2}$ (d) None

4. The two regression coefficients for the following data are :

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

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

5. From the following data, obtain the two regression equations and estimate the value of X when Y is 130 and estimate the value of Y when X is 30.

X	39	33	30	31	32	36	41	49	46	43
Y	132	134	138	129	136	131	132	135	128	125

(a) X = 29.38 and Y = 143.66 (b) X = 35.15 and Y = 153.80
(c) X = 39.32 and Y = 133.84 (d) None of these

Answer Key

1. (a) **2.** (d) **3.** (b) **4.** (a) **5.** (c)

SOME MORE FORMULAS – REGRESSION:

I. Formulas for Regression Coefficients in terms of Covariance and Variances:

By definition, the regression coefficient of Y on X is given by

$$b_{yx} = \frac{Cov(x, y)}{\sigma_x^2} \text{ or } b_{yx} = r.\frac{\sigma_y}{\sigma_x}$$

Similarly, the regression coefficient of X on Y is given by

$$b_{yx} = \frac{Cov(x, y)}{\sigma_y^2} \text{ or } b_{yx} = r.\frac{\sigma_x}{\sigma_Y}$$

The reader may also recall that the covariance between X and Y & the variances of X and Y values are respectively given by

$$Cov(X, Y) = \frac{\sum (x - \bar{x})(y - \bar{y})}{n} \text{ or } \frac{\sum xy}{n} - \bar{x} \cdot \bar{y}$$

$$\sigma_x^2 = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} \text{ or } \sqrt{\frac{\sum x^2}{n} - (\bar{x})^2}$$

$$\sigma_y^2 = \sqrt{\frac{\sum (y - \bar{y})^2}{n}} \text{ or } \sqrt{\frac{\Sigma y^2}{n} - (\bar{y})^2}$$

Formulas for regression coefficients in terms of deviation of X– and Y– values from their respective means:

By definition, the covariance between x and y is given by

$$Cov(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N} = \frac{\sum x_i y_i}{N} - \bar{x}\,\bar{y}$$

Further, the variances of x and y values are respectively given by

$$\sigma_X^2 = \frac{Cov\,(X, Y)}{b_{YX}} \text{ and } \sigma_y^2 = \frac{Cov\,(X, Y)}{b_{XY}}$$

Thus, using formulas, we obtain

$$\sigma_y^2 = \frac{Cov\,(X, Y)}{b_{XY}} \text{ and } b_{XY} = \frac{Cov\,(X, Y)}{\sigma_y^2}$$

RELATION OF REGRESSION COEFFICIENTS AND CORRELATION COEFFICIENT:

$$r = \pm\sqrt{\pm b_{XY} \times b_{YX}}. \text{ and } r = \frac{Cov\,(x.y)}{\sigma_x \sigma_y}$$

Example 45. Given $\sigma_x = 20$, $\sigma_y = 20$ and cov (X, Y) = –100 find:

1. Correlation coefficient 2. Both the regression Coefficients.

(a) 0.25, 30, 78(b)–0.25 –0.25, –0.25

(c) 25 0.55 0.10(d)–0.45 0.25 0.72

Sol. (b) (1) Correlation coefficient:

$$r = \frac{Cov\,(x.y)}{\sigma_x \sigma_y} = \frac{-100}{(20)(20)} = -0.25$$

(2) Both the regression Coefficients.

$$\text{Now } b_{XY} = \frac{Cov\,(X, Y)}{\sigma_y^2} = \frac{-100}{20^2} = -0.25.$$

$$\text{And } b_{YX} = \frac{Cov\,(X, Y)}{\sigma_X^2} = \frac{-100}{20^2} = -0.25.$$

Hence the correct option is (b)

Example 46. Given $\bar{x} = 50, \bar{y} = 20$, $\sigma_x = 20$, and $\sigma_y = 20$, find both the regression Coefficients.

(a) 0.25x + y – 32.5 = 0, x + 0.25y – 55 = 0

(b) 0.25x + y – 32.5 = 0, x + 0.25y – 55 = 0

(c) x + 0.25y – 55 = 0, 0.792x – 1.033y – 1.228 = 0

(d) –0.45 x + 0.25y – 55 = 0, 0.792x – 1.0y – 1.28 = 0

Sol. (b) $b_{XY} = \frac{Cov\,(X, Y)}{\sigma_y^2} = \frac{-100}{20^2} = -0.25.$

$$\text{And } b_{YX} = \frac{Cov\,(X, Y)}{\sigma_X^2} = \frac{-100}{20^2} = -0.25.$$

Regression equation y on x is $(y - \bar{y}) = b_{yx}(x - \bar{x})$ ...(i)

Therefore, $(y - 20) = -0.25(x - 50)$

$0.25x + y - 32.5 = 0$

Regression equation x on y is $(x - \bar{x}) = b_{xy}(y - \bar{y})$ .....(ii)

Therefore, $(x - 50) = -0.25(y - 20)$

$x + 0.25y - 55 = 0$

Hence the correct option is (b)

Example 47. Find the regression coefficients b_{yx} and b_{xy} of Y on X and X on Y respectively, if standard deviations of X and Y are 4 and 3 respectively, and coefficient of correlation between X and Y is 0.8

(a) 2.4, 3.8 (b) 8.2, 4.8 (c) 1.2, 2.2 (d) 3.2, 2.4

Sol. (d) Given: $\sigma_x = 4$, $\sigma_y = 3$ and $r = 0.8$

As we know that $r = \frac{Cov(x.y)}{\sigma_x \sigma_y}$

Then $0.8 = \frac{Cov(x,y)}{(4)(3)}$

Cov (x,y) = $12 \times 0.8 = 9.6$.

Therefore, $b_{XY} = \frac{Cov(X,Y)}{\sigma^2} = \frac{9.6}{3} = 3.2$

and $b_{YX} = \frac{Cov(X,Y)}{\sigma_X^2} = \frac{9.6}{4} = 2.4$

Hence the correct option is (d)

Example 48. The coefficient of correlation between the ages of husbands and wives in a community was found to be +0.8. The average of the husband's age of 25 years and that of the wives age is 22 years. Their standard deviations were 4 and 5 respectively. Find with the help of a regression equation, the expected age of husband when wife's age is 18 years.

(a) 12 (b) 13 (c) 10 (d) 11

Sol. (d) Given: $r = 0.8$, $\bar{x} = 25, \bar{y} = 22, \sigma_x = 4$ and $\sigma_y = 5$.

Let us consider the age of husband be x,

And the age of wives be y.

In the question the age of wives is independent variable and the age of husband is variable.

Therefore we have to find, y on x

As we know the standard equation of y on x is $(y - \bar{y}) = b_{yx}(x - \bar{x})$

Now, $b_{YX} = r \cdot \frac{\sigma_Y^2}{\sigma_X^2} = \frac{5^2}{4^2} = \frac{25}{18} = 1.5625$.

Then, $(y - \bar{y}) = b_{yx}(x - \bar{x})$

$(y - 22) = 1.5625(x - 25)$

$1.5625x - y = 17.0625$ ....(a)

The expected age of husband when wife's age is 18 years:

Put $y = 18$ in the 1st equation.

Then, $1.5625x - 18 = 17.0625$

$y = 11.0625 \approx 11$.

Hence the correct option is (d)

Example 49. From the following data obtain the two-regression equation and estimate the value of x when y is 130 and estimate the value of y when x is 30.

X	39	33	30	31	32	36	41	49	46	43
Y	132	134	138	129	136	131	132	135	128	125

(a) 39.328, 140.19 (b) 35.322, 120.139

(c) 89.38, 130.19 (d) 19.3228, 142.19

Sol. (a)

(X)	(Y)	$d_x = X - \bar{X}$ = X – 38	$d_y = Y - \bar{Y}$ = Y – 132	d_x^2	d_y^2	$d_x d_y$
39	132	1	0	1	0	0
33	134	-5	2	25	4	-10
30	138	-8	6	64	36	-48
31	129	-7	-3	49	9	21
32	136	-6	4	36	16	-24
36	131	-2	-1	4	1	2
41	132	3	0	9	0	0
49	135	11	3	121	9	33
46	128	8	-4	64	16	-32
43	125	5	-7	25	49	-35
$\sum X = 380$	$\sum Y = 1320$	$\sum d_x = 0$	$\sum d_y = 0$	$\sum d_x^2 = 398$	$\sum d_y^2 = 140$	$\sum d_x d_y = -93$

Here $n = 10$

Then, $\bar{X} = \frac{\Sigma X}{n} = \frac{380}{10} = 38$

And $\bar{Y} = \frac{\sum Y}{n} = \frac{1320}{10} = 132$

Now $b_{yx} = \frac{n\sum d_x d_y - \sum d_x \sum d_y}{n\sum d_x^2 - \left(\sum d_x\right)^2} = \frac{(10)(-93) - (0)(0)}{(10)(398) - (0)^2} = \frac{93}{398} = -0.234$

And $b_{xy} = \frac{n\sum d_x d_y - \sum d_x \sum d_y}{n\sum d_y^2 - \left(\sum d_y\right)^2} = \frac{(10)(-93) - (0)(0)}{(10)(140) - (0)^2} = \frac{93}{140} = -0.664$

Then the regression equation y on x is $(y - \bar{y}) = b_{yx}(x - \bar{x})$

$(y - 132) = -0.234(x - 38)$

$0.234x + y = 140.892$

$y = 140.892 - 0.234x$ .....(a)

And the regression equation is $(x - \bar{x}) = b_{xy}(y - \bar{y})$

$(x - 38) = -0.664(y - 132)$

$x + 0.664\ y = 125.648$

$x = 125.648 - 0.664y$ .....(b)

Estimate the value of x when y is 130,

Put y = 130 in 2nd equation.

Then we get, $x = 125.648 - 0.664\ (130) = 39.328.$

Estimate the value of y when x is 30

Put x = 30 in 1st equation.

$y = 140.892 - 0.234(3) = 140.19.$

Hence the correct option is (a)

Example 50. The following data give the aptitude test score and productivity indicate 10 workers selected at random

Aptitude scores (X):	60	62	65	70	72	48	53	73	65	82
Productivity index(Y):	68	60	62	80	85	40	52	62	60	81

Calculate the two-regression equation and estimate the productivity index of a worker whose test score is 92.

(a) 100 (b) 95.321 (c) 85.321 (d) 105

Sol. (b)

(X)	(Y)	$d_x = X - \bar{X}$ = X - 65	$d_y = Y - \bar{Y}$ = Y - 65	d_x^2	d_y^2	$d_x d_y$
60	68	-5	3	25	9	-15
62	60	-3	-5	9	25	-15
65	62	0	-3	0	9	0
70	80	5	15	25	225	75
72	85	7	20	49	400	140
48	40	-17	-25	289	625	425

53	52	-12	-13	144	169	146
73	62	8	-3	64	9	-24
65	60	0	-5	0	25	0
82	81	17	16	289	256	272
$\sum X = 650$	$\sum Y = 650$	$\sum d_x = 0$	$\sum d_y = 0$	$\sum d_x^2 = 894$	$\sum d_y^2 = 1752$	$\sum d_x d_y = 1004$

Here n=10

Then, $\bar{X} = \frac{\Sigma X}{n} = \frac{650}{10} = 65$

And $\bar{Y} = \frac{\sum Y}{n} = \frac{650}{10} = 65$

Now $b_{yx} = \frac{n\sum d_x d_y - \sum d_x \sum d_y}{n\sum d_x^2 - \left(\sum d_x\right)^2} = \frac{(10)(1004) - (0)(0)}{(10)(894) - (0)^2} = \frac{1004}{984} = 1.123$

And $b_{xy} = \frac{n\sum d_x d_y - \sum d_x \sum d_y}{n\sum d_y^2 - \left(\sum d_y\right)^2} = \frac{(10)(1004) - (0)(0)}{(10)(1752) - (0)^2} = \frac{1004}{1752} = 0.573$

Then the regression equation y on x is $(y - \bar{y}) = b_{yx}(x - \bar{x})$

$(y - 65) = 1.123(x - 65)$

$1.123x - y = 7.995$

$y = 1.123x - 7.995$ .....(a)

And the regression equation is $(x - \bar{x}) = b_{xy}(y - \bar{y})$

$1.0x - 0.573y = 27.755$

$x = 27.755 + 0.573y$....(b)

Estimate the productivity index of a worker whose test score is 92.

Put x = 92 in 1st equation.

Then we get,

$y = 1.123(92) - 7.995$

$y = 95.321.$

Hence the correct option is (b)

PROPERTIES OF REGRESSION COEFFICIENTS:

SOME IMPORTANT PROPERTIES OF REGRESSION COEFFICIENTS:

- **Property 1:** The coefficients of correlation and two regression coefficients have the same signs.
- **Property 2:** The coefficients of correlation are the geometric mean between the regression coefficients.

- **Property 3:** If one of the regression coefficients is greater than unity, the other must be less than the unity.
- **Property 4:** The two lines of regression intersect at the point $(\bar{x}, \bar{y})$ where x and y are the variables under consideration.
- **Property 5:** The regression coefficients are independent of change of origin but not for scale.

Determining the line of regression of y on x and that of x on y out of the given two regression lines:

Sometimes, it is required to find the line of regression of y on x or x on y out of the given two regression lines in such a case we below the following steps:

Step 1: Choose any one of the two regression lines as the line of regression of y on x and the other as the line of regression of x on y.

Step 2: Find two regression coefficients b_{xy} and b_{yx}

Step 3: Compute the product $b_{xy} \cdot b_{yx}$ if $b_{xy} \cdot b_{yx.} \leq 1$, then the assumption made in step 1 is correct, otherwise the assumption is wrong.

Example 51. Regression coefficient are _____.

(a) dependent of change of origin and of scale.

(b) independent of both change of origin and of scale.

(c) dependent of change of origin but not of scale.

(d) independent of change of origin but not of scale.

Sol. (d) We know, the regression coefficients are independent of the change of the origin. But, they are not independent of the change of the scale.

It means there will be no effect on the regression coefficients if any constant is subtracted from the value of x and y.

If x and y are multiplied by any constant, then the regression coefficient will change.

Hence, the correct option is (d) i.e. independent of change of origin but not of scale.

Example 52. 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 (**ICAI**)

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

Sol. (a) Given equations,

$2x + 3y = -1$ ---------- 1

$5x + 6y = -1$ ---------- 2

Multiply eq. 1 by 2, we get

$4x + 6y = -2$ --------- 3

Subtract equation 3 from 2, we get

$x = 1$

Now, substitute $x = 1$ in eq 1,

$\Rightarrow 21 + 3y = -1$

$\Rightarrow 2 + 3y = -1$

$\Rightarrow 3y = -1-2$

$\Rightarrow 3y = -3$

$\Rightarrow y = -1$

We know, the two lines of regression intersects at (x, y), then the arithmetic means of x and y are given by 1, –1.

Hence, the correct answer is option (a) i.e 1, –1.

Example 53. 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) $y = 2x + 3$ (d) $y = 2x + 5$

Sol. (c) Given, $5a + 10b = 40$ .....(i)

$10a + 25b = 95$ .....(ii)

Multiplying eq (i) by 2, we get

$10a + 20b = 80$ .....(iii)

Subtracting eq (ii) and (iii), we get

$10a + 25b = 95$

$10a + 20b = 80$

$5b = 15$

$b = 3$

from (i)

$5a + 10b = 40$

$5a + 10$ (iii) $= 40$

$5a + 30 = 40(3)$

$5a = 40 - 30$

$5a = 10$

$a = 2$

As, "a" represents the y-intercept and "b" represents the slope

So the regression line of y on x is given by $y = 2x + 3$

Hence the correct option is (c).

Example 54. 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

Sol. (b) Given, $u = 2x + 5$, $v = -3y - 6$

Regression coefficient of y on x = 2.4

We have to find the regression coefficient of v on u.

u and v can be written as

$$u = \frac{x + \frac{5}{2}}{\frac{1}{2}}$$

$$v = \frac{y + 2}{-\frac{1}{3}}$$

$$p = \frac{1}{2}, q = -\frac{1}{3}$$

$$b_{yx} = 2.4$$

$$b_{xy} = \frac{q}{p} b_{vu}$$

Substitute the values then we get

$$2.4 = -\frac{\frac{1}{3}}{\frac{1}{2}} b_{vu}$$

$$b_{vu} = -\frac{2.4 \times 3}{2} = 3.6$$

The regression coefficient of v on $u = -3.6$

Hence the correct option is (b)

Example 55. 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 and y?

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

Sol. (a) Given $r = \sqrt{b_{XY} \times b_{YX}}$ ..... (i)

Now regression line of y on x is given by:

$4y = 15 + 5x$

$$\Rightarrow y = \frac{15}{4} + \frac{5}{4}x$$

$$\Rightarrow b_{yx} = \frac{5}{4}$$

Also, coefficient of correlation $r = 0.75$

Now substituting in eq (i), we get

$$0.75 = +- \sqrt{b_{xy} \times \frac{5}{4}x}$$

Squaring both sides we get

$$(0.75)^2 = \frac{5}{4} b_{XY}$$

$$\Rightarrow 0.5625 = \frac{5}{4} b_{XY}$$

$\Rightarrow$ or $b_{XY} = 0.5625 \times \frac{4}{5}$

$\Rightarrow$ or $b_{XY} = 0.45$

Hence, the correct option is (a).

Example 56. If the regression line of y on x and of x on y are given be $2x + 3y = -1$ 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)

Sol. (a) **Given:**

$2x+3y= -1$ ..... (i)

$5x + 6y = -1$ .....(ii)

Multiplying eq (i) by 5 and eq (ii) by 2, we get:

$10x + 15y = -5$

$10x + 12y = -2$

Subtracting the above equations, we get:

$3y = -3$

$y = -1$

Substituting $y = -1$ in the eq (i), we get

$2x + 3(-1) = -1$

$2x = 2$

$x = 1$

Therefo re, $x = 1$ and $y = -1$.

Hence the correct option is (a).

PRACTICE QUESTIONS (PART E)

1. The regression equation are $2x + 3y + 1 = 0$ and $5x + 6y + 1 = 0$, then Mean of x and y respectively are **(Dec 2022)**

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

2. The correlation coefficient between x and y is $\frac{-1}{2}$.The value of $\frac{-1}{8}$ Find b_{yx}.

   (a) – 2 (b) – 4 (c) 0 (d) 2

3. 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

4. The equations of the two lines of regression are $4x + 3y + 7 = 0$ and $3x + 4y + 8 = 0$. Find the correlation coefficient between x and y. (Dec 2022)

   (a) –0.75 (b) 0.25 (c) –0.92 (d) 1.25

Answer Key

1. (c) **2.** (a) **3.** (a) **4.** (a)

Example 57. What is a spurious correlation? **(ICAI)**

(a) It is a bad relation between two variables.

(b) It has very low correlation between the two variables.

(c) It is the correlation between two variables having no causal relation.

(d) It is a negative correlation.

Sol. (c) Spurious correlation refers to the correlation between two variables that do not have any causal relationship between them. It is a statistical term used to describe a situation where two variables appear to be related to each other, but actually, they are not.

Hence the correct option is (c)

Example 58. Scatter diagram helps us to **(ICAI)**

(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)

Sol. (d) Scatter plot (scatter graph, scatter chart) uses dots to represent values for two different numeric variables.

Scatter diagram is a tool for analyzing the relationship between two variables, i.e., whether variables are correlated or not.

Also, it will help to identify the types of correlation

And from a diagram we can observe that the relation is linear or non linear.

Hence the correct option is (d)

Example 59. When r = 1, all the points in a scatter diagram would lie **(ICAI)**

(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)

Sol. (d) Perfectly positive correlation if all the points on the scatter diagram are on a straight line with positive slope then the correlation is said to be perfectly positive, that is r = +1.

Perfectly negative correlation if all the points of the scatter diagram fall on a straight line with negative slope in the correlation is said to be perfectly negative or r = −1.

Hence the correct option is (d)

Example 60. What are the limits of the correlation coefficient? **(ICAI)**

(a) No limit(b)−1 and 1, excluding the limits

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

Sol. (d) Either two variables can be completely related i.e. 1 or we can say positive correlation or the two variables are completely opposite i.e. negatively correlated –1 situation case.

Hence the correct option is (d)

Example 61. If the coefficient of correlation between two variables is –0.9, then the coefficient of determination is (**ICAI**)

(a) 0.9 (b) 0.81 (c) 0.1 (d) 0.19

Sol. (b) Coefficient of Correlation and Coefficient of Determination

The coefficient of correlation is a statistical measure that indicates the degree of association between two variables. It ranges from –1 to +1.

Coefficient of Correlation = –0.9

The formula for the coefficient of determination, denoted by r^2 where r is Coefficient of Correlation.

The coefficient of determination is $(-0.9)^2 = 0.81$

Hence the correct option is (b)

Example 62. If the coefficient of correlation between two variables is 0.7, then the percentage of variation unaccounted for is (**ICAI**)

(a) 70% (b) 30% (c) 51% (d) 49%

Sol. (d) The coefficient of correlation between two variables is 0.7,

Then the percentage of variation unaccounted = $r^2 \times 100 = 0.49(100) = 49\%$.

Hence the correct option is (d)

Example 63. What are the limits of the two regression coefficients? (**ICAI**)

(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

Sol. (d) The product of the regression coefficients must be numerically less than unity to avoid problems with multicollinearity. Multicollinearity occurs when there is a high correlation between two or more independent variables in a multiple regression model. This can lead to unstable and unreliable estimates of the regression coefficients. To avoid this problem, the product of the regression coefficients must be less than one.

Hence the correct option is (d)

PRACTICE QUESTIONS (PART F)

1. If the regression line of Y on X is given by Y = X + 2 and Karl Pearson's coefficient is 0.5 then $\frac{\sigma_y^2}{\sigma_x^2} = ?$

(a) 3 (b) 2 (c) 4 (d) None

2. Given the following series:

x:	10	13	12	15	8	15
y:	12	16	18	16	7	18

The rank correlation coefficient r = ?

(a) $1-\dfrac{6\sum d^2+\sum_{i=1}^{2}\dfrac{m_1\left(m_1^2-1\right)}{12}}{n\left(n^2-1\right)}$ (b) $1-\dfrac{6\left[\sum_{1=i}^{2}\dfrac{m_1\left(m_1^2-1\right)}{12}\right]}{n\left(n^2-1\right)}$

(c) $1-6\sum d^2+\sum_{i=1}^{2}\dfrac{\dfrac{m_t^2\left(m_1^2-1\right)}{12}}{n\left(n^2-1\right)}$ (d) $1-6\sum d^2+\sum_{i=1}^{2}\dfrac{\dfrac{m_1\left(m_1^2-1\right)}{12}}{n\left(n^2-1\right)}$

3. Determine spearman's rank correlation coefficient from the given data $\sum d^2 = 30, n = 10$:

(a) r = 0.82 (b) r = 0.32 (c) r = 0.40 (d) None of the above

4. Which of the following statements is not true about scatter diagrams?

(a) It finds the type of correlation

(b) It helps to identify whether variables are correlated or not

(c) It determines the linear or non-linear correlation

(d) It finds the numerical value of correlation coefficient

5. Given that

X	-3	-3/2	0	3/2	3
Y	9	9/4	0	9/4	9

Then Karl pearson's coefficient of correlation is

(a) Positive (b) Zero (c) Negative (d) None

6. A.M of regression coefficient is

(a) Equal to r (b) Greater than or equal to r

(c) Half of r (d) None

7. The regression coefficient is independent of the change of

(a) Scale (b) Origin

(c) Scale and origin both (d) None of these

8. If the correlation coefficient between the variables X and Y is 0.5, then the correlation coefficient between the variables 2x – 4 and 3 - 2y is

(a) 1 (b) 0.5 (c) –0.5 (d) 0

9. If the two regression lines are $3x = y$ and $8y = 6x$, then the value of correlation coefficient is

(a) 0.5 (b) -0.5 (c) 0.75 (d) -0.80

10. If $r = 0.6$ then the coefficient of non – determination will be:

(a) 0.40 (b) -0.60 (c) 0.36 (d) 0.64

11. The correlation coefficient (r) is the ____ of the two regression coefficient (b_{yx} and b_{xy})

(a) AM (b) GM (c) HM (d) Median

12. If there is a constant increase in a series, then the corresponding graph will be

(a) Convex curve (b) Concave curve

(c) Parabola (d) Straight line from the left to the right

13. If the plotted points are scatter diagram are evenly distribution, then the correlation is

(a) Zero (b) Negative (c) Positive (d) (a) or (b)

14. The coefficient of determination is defined by the formula

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

(c) both (a) and (b) (d) None

15. In the 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 SD (b) Coefficient of regression

(c) Coefficient of correlation (d) None

16. Correlation coefficient is _____ of units of measurement.

(a) Dependent (b) Independent (c) Both (d) None

17 In case speed of an automobile and the distance required to stop the car after applying brakes correlation is

(a) Positive (b) Negative (c) Zero (d) None

18 A relationship $r^2 = 1 - \frac{500}{300}$ is not possible

(a) True (b) False (c) Both (d) None

19. Rank correlation coefficient lies between

(a) 0 to 1 (b) −1 to +1 inclusive of these value

(c) −1 to 0 (d) Both

19. The two line of regression intersect at the point

(a) Mean (b) Mode (c) Median (d) None of these

21. If the two line of regression are $x + 2y - 5 = 0$ and $2x + 3y - 8 = 0$, then the regression line of y on x is:

(a) $x + 2y - 5 = 0$ (b) $2x + 3y - 8 = 0$

(c) $x + 2y = 0$ (d) $2x + 3y = 0$

22. If the sum of squares of the differences of ranks, given by two judges X and Y of 10 students is 28, what is the value of the rank correlation coefficient?

(a) 0.725 (b) 0.650 (c) 0.750 (d) 0.873

23. The ranks of five cars given by two critics are:

Car	Critic 1	Critic 2
A	2	1
B	1	3
C	4	2
D	5	4
E	3	5

What is the rank correlation coefficient between the ranks assigned by the two critics?

(a) 1 (b) 0.3 (c) -1 (d) 0.5

24. For 12 pairs of observations, the number of concurrent deviations was found to be 3. What is the value of the coefficient of concurrent deviation?

(a) $\sqrt{\frac{5}{11}}$ (b) $\frac{5}{11}$ (c) $-\sqrt{\frac{5}{11}}$ (d) None of these

25. The sum of squares of rank differences in the ages of 12 individuals is 64. Calculate the coefficient of rank correlation.

(a) 0.78 (b) 0.73 (c) 0.87 (d) 0.80

26. If the two line of regression are $2Y - X = 35$ and $10X - Y = 70$. The regression line of Y on X is

(a) $2Y - X = 35$ (b) $10X - Y = 70$

(c) Any of the two line (d)None of the two line

27 Consider the regression line of y on x as $y = 20 + 0.4(x - 15)$ and the regression line of x on y as $x = 30 + 0.9y - 40$, what is the correlation coefficient between x and y?

(a) 0.36 (b) 0.4

(c) 0.6 (d) None of the above

28. If the two regression lines are $3X = Y$ and $8Y = 6X$, then the value of correlation coefficient is

(a) -0.5 (b) 0.5 (c) 0.75 (d) -0.80

Answer Key									
1. (c)	2. (b)	3. (a)	4. (b)	5. (b)	6. (b)	7. (b)	8. (c)	9. (a)	10. (d)
11. (b)	12. (d)	13. (a)	14. (c)	5. (c)	16. (b)	17. (a)	18. (a)	19. (b)	20. (a)
21. (a)	22. (d)	23. (b)	24. (c)	25. (a)	26. (a)	27. (C)	28. (b)		

SUMMARY

- There are four ways to find Correlation :
- Scatter Diagram :
- A scatter diagram is a graphical presentation of bivariate data {(Xi, Yi): i = 1, 2, ... n} on two quantitative variables X and Y that allows us to show two variables together.
- Perfect Positive Correlation : If the points of the scatter diagram fall on a straight line and have a positive(upward) slope, then the correlation is said to be perfectly positive;
- Positive Correlation: When the points of the scatter diagram cluster around a straight line (upward slope from left to right), then the correlation is said to be positive.
- Perfect Negative Correlation : If the points of the scatter diagram fall on a straight line and have a negative(downward) slope, then the correlation is said to be perfectly negative
- Negative Correlation: When the points of the scatter diagram cluster around a straight line (downward/negative slope), then the correlation is said to be negative.
- No Correlation: When the points of the scatter diagram are scattered in a haphazard manner, then there is zero or no correlation.
- Correlation coefficient (r) :

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

Degree of Correlation	Positive	Negative
Perfect correlation	+1	-1
Very high degree of correlation	+ 0.9 to + 1	- 0.9 to – 1
Fairly high degree of correlation	+ 0.75 to + 0.9	- 0.75 to - 0.9
Moderate degree of correlation	+ 0.50 to + 0.75	- 0.50 to - 0.75
Low degree of correlation	+ 0.25 to + 0.50	- 0.25 to - 0.5
Very low degree of correlation	0 to + 0.25	- 0.25 to 0
No correlation	0	0

REGRESSION:

- Linear regression establishes the linear relationship between two variables based on a line of best fit. Linear regression is thus graphically depicted using a straight line with the slope defining how the change in one variable impacts a change in the other.
- The y-intercept of a linear regression relationship represents the value of one variable when the value of the other is zero
- Formulated by : y = mx + b
- where m is the slope and b is the intercept
- $r = \pm\sqrt{b_{yx} \times b_{xy}}$
- $b_{yx} = \frac{q}{p} \times b_{vu}$ where $u = \frac{x-a}{p}$ and $v = \frac{y-c}{q}$
- Coefficient of determination : r^2