CORRELATIONAND

Introduction

In the earlier chapters we have discussed the characteristics and shapes of distributions of a single variable, eg, mean, S.D. and skewness of the distributions of variables such as income, height, weight, etc. We shall now study two (or more) variables simultaneously and try to find the quantitative relationship between them. For example, the relationship between two variables like (1) income and expenditure (2) height and weight, (3) rainfall and yield of crops, (4) price and demand, etc. will be examined here. The methods of expressing the relationship between two variables are due mainly to Francis Galton and Karl Pearson.

Correlation

Correlation is a statistical measure for finding out degree (or strength) of association between two (or more) variables. By 'association' we mean the tendency of the variables to move together. Two variables X and Y are so related that movements (or variations) in one, say X, tend to be accompanied by the corresponding movements (or variations) in the other Y, then X and Y are said to be correlated. The movements may be in the same direction (i.e. either both X, Y increase or both of them decrease) or in the opposite directions (ie., one, say X, increases and the other Y decreases). Correlation is said to be positive or negative according as these movements are in the same or in the opposite directions. If Y is unaffected by any change in X, then X and Y are said to be uncorrelated.

In the words L.R. Conner:

If two or more quantities vary in sympathy so that movements in the one tend to be accompanied by corresponding movements in the other, then they are said to be correlated."

Correlation may be linear or non-linear. If the amount of variation in X bears a constant ratio to the corresponding amount of variation in Y, then correlation between X and Y is said to be linear. Otherwise it is non-linear. Correlation coefficient (r) measures the degree of linear relationship, (i.e., linear correlation) between two variables.

Determination of Correlation

Correlation between two variables may be determined by any one of the following methods:

- Scatter Diagram
- Co-variance Method or Karl Pearson's Method
- Rank Method

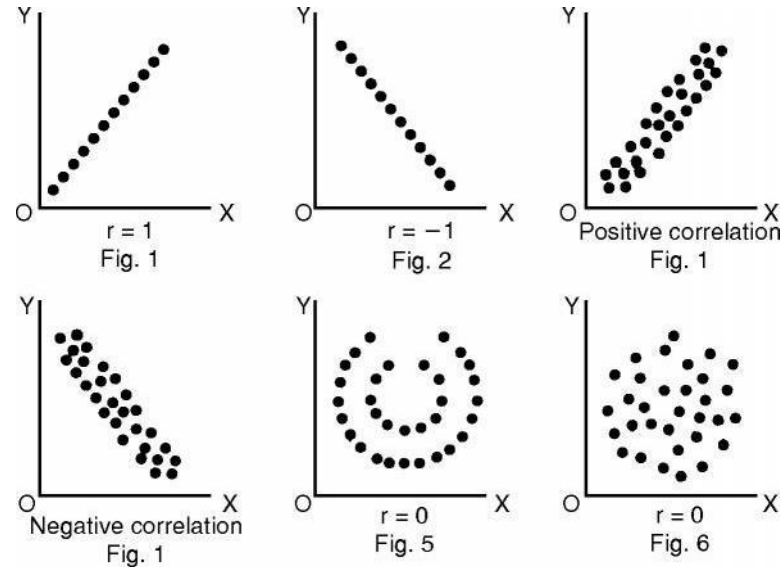
Scatter Diagram

The existence of correlation can be shown graphically by means of a *scatter diagram.* Statistical data relating to simultaneous movements (or variations) of two variables can be graphically represented-by points. One of the two variables, say X, is shown along the horizontal axis OX and the other variable Y along the vertical axis OY. All the pairs of values of X and Y are now shown by points (or dots) on the graph paper. This diagrammatic representation of bivariate data is known as scatter diagram.

The scatter diagram of these points and also the direction of the scatter reveals the nature and strength of correlation between the two variables. The following are some scatter diagrams showing different types of correlation between two variables.

In Fig. 1 and 3, the movements (or variations) of the two variables are in the same direction and the scatter diagram shows a linear path. In this case, correlation is positive or direct.

In Fig. 2 and 4, the movements of the two variables are in opposite directions and the scatter shows a linear path. In this case correlation is negative or indirect.



In Fig. 5 and 6 points (or dots) instead of showing any linear path lie around a curve or form a swarm. In this case correlation is very small and we can take r = 0.

In Fig. 1 and 2, all the points lie on a straight line. In these cases correlation is perfect and r = +1 or 1 according as the correlation is positive or negative.

Karl Pearson's Correlation Coefficient

We have remarked in the earlier section that a scatter diagram gives us only a rough idea of how the two variables, say x and y, are related. We cannot draw defensible conclusions by merely examining data from the scatter diagram. In other words, we cannot simply look at a scatter diagram variables. On the other hand, neither can we conclude that the correlation at all. We need a quantity (represented by a number), which is a measure of the extent to which x and y are related. The quantity that is used for this purpose is known as the Co-efficient of Correlation, usually denoted by r_{xy} or r. The co-efficient of correlation r_{xy} measures the degree (or extent) of relationship between the two variables x and y and is given by the following formula:

$$r_{xy} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n\,\sigma_X\,\sigma_Y} \quad \text{.... (1)}$$

where X_t and Y_i $(i = 1, 2,.., n)$ are the two sets of values of x and y respectively and $\bar{x}$, $\bar{y}$, σ_x, σ_y are respectively the corresponding means and standard deviations so that

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\Sigma X_i^2 - \bar{X}^2$$

$$\text{and } \sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{n}\Sigma Y_i^2 - \bar{Y}^2$$

The above definition of the correlation co-efficient was given by Karl Pearson in 1890 and is called *Karl Pearson's Correlation Co-efficient* after his name.

Definition

If (X_1, Y_1), (X_2, Y_2) .... (X_n, Y_n) be n pairs of observations on two variables X and Y, then the covariance of X and Y, written as cov (X,Y) is defined by

$$\text{Cov}(X,Y) = \frac{1}{n}\Sigma(X_i - \bar{X})(Y_i - \bar{Y})$$

Covariance indicates the joint variations between the two variables.

So the correlation coefficient or the coefficient of correlation (r) between X and Y is defined by

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

where σ_X, σ_Y are standard deviations of X and Y respectively.

The formula for the Correlation Coefficient r may be written in different forms.

i. If $x_i = X - \bar{X}$ and $y_i = Y - \bar{Y}$

then $$r = \frac{x_i\, y_i}{n\,\sigma_X\,\sigma_Y} \quad (1)$$

$$\therefore \text{ from (1), } r = \frac{\frac{1}{n}\Sigma x_i y_i}{\sqrt{\frac{\Sigma x_i^2}{n}} \times \sqrt{\frac{\Sigma y_i^2}{n}}} = \frac{\Sigma x_i y_i}{\sqrt{\Sigma x_i^2} \times \sqrt{\Sigma y_i^2}}$$

ii. We have

$$\text{Cov}(X, Y) = \frac{1}{n}\Sigma(X_i - \bar{X})(Y_i - \bar{Y})$$

$$\frac{1}{n}\Sigma(X_i Y_i - \bar{X}_i Y - X\bar{Y}_i + \bar{X}\bar{Y})$$

$$= \frac{\Sigma X_i Y_i}{n} - \bar{Y}\frac{\Sigma X_i}{n} - \bar{X}\frac{\Sigma Y_i}{n} + \frac{n\bar{X}\bar{Y}}{n}$$

$$\frac{\Sigma X_i Y_i}{n} - \bar{X}\bar{Y} - \bar{X}\bar{Y} + \bar{X}\bar{Y}$$

$$= \frac{\Sigma XY}{n} - \bar{X}\bar{Y} = \frac{\Sigma XY}{n} - \left(\frac{\Sigma X_i}{n}\right)\left(\frac{\Sigma Y_i}{n}\right)$$

and conclude that since more than half of the points appear to be nearly in a straight line, there is a positive or negative correlation between the

Now, $$r = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

$$= \frac{\frac{\Sigma XY}{n} - \left(\frac{\Sigma X_i}{n}\right)\left(\frac{\Sigma Y}{n}\right)}{\sqrt{\frac{\Sigma X^2}{n} - \left(\frac{\Sigma X_i}{n}\right)^2} \times \sqrt{\frac{\Sigma Y^2}{n} - \left(\frac{\Sigma Y}{n}\right)^2}} \qquad ...(2)$$

iii. By multiplying each term of (2) by n^2, we have

$$r = \frac{n\Sigma X_i Y_i - (\Sigma X_i)(\Sigma Y_i)}{\sqrt{n\Sigma X_i^2 - (\Sigma X_i)^2} \times \sqrt{n\Sigma Y_i^2 - (\Sigma Y_i)^2}}$$

Theorem

The correlation coefficient is independent (not affected by) of the change of origin and scale of measurement.

Proof

Let $(x_1, y_1), (x_2, y_2)$ .... (x_n, y_n) be a set of n pairs of observations.

$$r_{xy} = \frac{\frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\Sigma(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\Sigma(y_i - \bar{y})^2}} \qquad ...(1)$$

Let us transform x_i to u_i and y_i to v_i by the rules,

$$u_i = \frac{x_i - x_0}{c_1} \quad \text{and } v_i = \frac{y_i - y_0}{c_2} \qquad ...(2)$$

where x_0, y_0, c_1, c_2 are arbitrary constants.

From (2), we have

$$x_i = c_1 u_i + x_0 \quad \text{and} \quad y_i = c_2 v_i + y_0$$

$$\bar{x} = x_0 + c_1 \bar{u} \quad \text{and} \quad \bar{y} = y_0 + c_2 \bar{v}$$

where $\bar{u}$ and $\bar{v}$ are the means u_i^s and v_i^s respectively.

$$x_i - \bar{x} = c_1(u_i - \bar{u}) \quad \text{and} \quad y_i - \bar{y} = c_2(v_i - \bar{v})$$

Substituting these values in (1), we get

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n} c_1(u_i - \bar{u})c_2(v_i - \bar{v})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} c_1^2(u_i - \bar{u})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} c_2^2(v_i - \bar{v})^2}}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(u_i - \bar{u})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_i - \bar{v})^2}}$$

$$= \frac{\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})}{n\sigma_u\sigma_v} = r_{uv}$$

Here, we observe that if we change the origin and choose a new scale, the correlation co-efficient remains unchanged. Hence the proof.

Here, r_{uv} can be further simplified as

$$r_{xy} = \frac{Cov(u,v)}{\sigma_u \sigma_v}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n} u_i v_i - \bar{u}\,\bar{v}}{\sqrt{\frac{1}{n}\sum u_i^2 - \bar{u}^2}\sqrt{\frac{1}{n}\sum v_i^2 - \bar{v}^2}}$$

$$\frac{n\sum u_i v_i - \sum u_i \sum v_i}{\sqrt{n\sum u_i^2 - (\sum u_i)^2}\sqrt{n\sum v_i^2 - (\sum v_i)^2}}$$

Limits of Correlation Co-efficient

We shall now find the limits of the correlation coefficient between two variables and show that it lies between -1 and $+1$.

ie., $-1 \le r_{xy} < +1$

Proof

Let (x_1, y_1), (x_2, y_2) .... (x_n, y_n) be the given pairs of observations.

Then $$r_{xy} = \frac{\frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum (y_i - \bar{y})^2}}$$

We put

$$X_i = x_i - \bar{x}, \qquad Y_i = y_i - \bar{y}$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \quad ...(1)$$

Similarly $$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 \quad ...(2)$$

and $$r_{xy} = \frac{\sum_{i=1}^{n} X_i Y_i}{n\sigma_x \sigma_y} \quad ...(3)$$

Now we have

$$\sum_{i=1}^{n}\left(\frac{X_i}{\sigma_x} \pm \frac{Y_i}{\sigma_y}\right)^2 = \frac{\sum_{i=1}^{n} X_i^2}{\sigma_x^2} + \frac{\sum_{i=1}^{n} Y_i^2}{\sigma_y^2} + \frac{2\sum_{i=1}^{n} X_i Y_i}{\sigma_x \sigma_y}$$

$$= \frac{n\sigma_x^2}{\sigma_x^2} + \frac{n\sigma_y^2}{\sigma_y^2} \pm 2nr_{xy} \quad \text{using (1), (2), (3).}$$

$$2n \pm 2n\, r_{xy} = 2n(1 \pm r_{xy})$$

Left hand side of the above identity is the sum of the squares of n numbers and hence it is positive or zero.

Hence, $1 \pm r_{xy} \ge 0$ or, $r_{xy} \le 1$ an d $r_{xy} \ge -1$

or $-1 \le r_{xy} \le +1$

ie., the correlation co-efficient lies between 1 and + 1. Hence the proof.

Note:

If $r_{xy} = 1$, we say that there is perfect positive correlation between x and y.

If $r_{xy} =$ 1, we say that there is perfect negative correlation between x and y.

If $r_{xy} = 0$, we say that there is no correlation between the two variables, i.e., the two variables are uncorrelated.

If $r_{xy} > 0$, we say that the correlation between x and y is positive (direct).

If $r_{xy} < 0$, we say that the correlation between x and y is negative (indirect).

X	Y	x = X $\overline{X}$	y = Y $\overline{Y}$	x^2	y^2	xy
1	6	3	4	9	16	12
2	8	2	2	4	4	4
3	11	1	1	1	1	1
4	9	0	1	0	1	0
5	12	1	2	1	4	2
6	10	2	0	4	0	0
7	14	3	4	9	16	12
28	70			28	42	29

$$\overline{X} = \frac{\Sigma X}{n} = \frac{28}{7} = 4 \text{ and } \overline{Y} = \frac{\Sigma Y}{n} = \frac{70}{7} = 10$$

Karl Pearson's coefficient of correlation (r) is given by

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

$$= \frac{29}{\sqrt{28}\sqrt{42}} = \mathbf{0.8457}$$

Example 9

Karl Pearson's coefficient of correlation between two variables X and Y is 0.28 their covariance is +7.6. If the variance of X is 9, find the standard deviation of Y-series.

Solution

Karl Pearson's coefficient of correlation r is given by

$$r = \frac{\text{cov}(X, Y)}{\sigma x\, \sigma y}$$

Here r =0.28, Cov (X, Y) = 7.6 and $\sigma x^2 = 9$; $\sigma x = 3$.

Using (1) $0.28 = \frac{7.6}{\ }$

or, $0.84\, \sigma y = 7.6, or$ $\sigma y = \frac{7.6}{0.84} = \frac{760}{84}$

= **9.048**

Example 10

Calculate Pearson's coefficient of correlation between advertisement cost and sales as per the data given below:

Advt cost in '000 Rs:	39	65	62	90	82	75	25	98	36	78
Sales in lakh Rs:	47	53	58	86	62	68	60	91	51	84

Solution

Karl Pearson's coefficient of correlation (r) is given by

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

X	Y	$x = X - \overline{X}$	$y = Y - \overline{Y}$	x^2	y^2	xy
39	47	26	19	676	361	494
65	53	0	13	0	169	0
62	58	3	8	9	64	24
90	86	25	20	625	400	500
82	62	17	4	289	16	68
75	68	10	2	100	4	20
25	60	40	6	1600	36	240
98	91	33	25	1089	625	825

Example 8

Find the coefficient of correlation from the following data:

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

36	51	29	15	841	225	435
78	84	13	18	169	324	234
650	660	0	0	5398	2224	2704

$$\overline{X} = \frac{\Sigma X}{n} = \frac{650}{10} = 65 \; ; \; \overline{Y} = \frac{\Sigma Y}{n} = \frac{660}{10} = 66$$

$$r = \frac{2704}{\sqrt{5398} \times \sqrt{2224}} = \mathbf{0.78}$$

Example 11

Calculate Pearson's coefficient of correlation from the following taking 100 and 50 as the assumed average of X and Y respectively:

X:	104	111	104	114	118	117	105	108	106	100	104	105
Y:	57	55	47	45	45	50	64	63	66	62	69	61

Solution

X	Y	u = X 100	v = Y 50	u^2	v^2	uv
104	57	4	7	16	49	28
111	55	11	5	121	25	55
104	47	4	3	16	9	12
114	45	14	5	196	25	70
118	45	18	5	324	25	90
117	50	17	0	289	0	0
105	64	5	14	25	196	70

$$\Sigma X Y = 508 - (6 \times 14 + 8 \times 6) + (8 \times 12 + 6 \times 8)$$

108	63	8	13	64	169	104
106	66	6	16	36	256	96
100	62	0	12	0	144	0
104	69	4	19	16	361	76
105	61	5	11	25	121	55
		96	84	1128	1380	312

$$r = \frac{\Sigma u_i v_i - \Sigma u_i \Sigma v_i}{\sqrt{n \Sigma u_i^2 - (\Sigma u_i)^2}\sqrt{n \Sigma v_i^2 - (\Sigma v_i)^2}}$$

$$= \frac{12 \times 312 - 96 \times 84}{\sqrt{12 \times 1128 - (96)^2}\sqrt{12 \times 1380 - (84)^2}}$$

0.67

Example 12

A computer while calculating the correlation coefficient between two variables X and Y from 25 pairs of observations obtained the following results: $n = 25$, $\Sigma X = 125$, $\Sigma Y = 100$, $\Sigma X^2 = 650$, $\Sigma Y^2 = 460$ and $XY = 508$. It was, however, discovered at the time of checking that two pairs of observations were not correctly copied. They were taken as (6, 14) and (8, 6), while the correct values were (8, 12) and (6, 8). Prove that the correct value of the correlation coefficient should be 2/3.

Solution

When the two incorrect pairs of observations are replaced by the correctpairs, the revised results for the whole series are:

ΣX = 125 (Sum of two incorrect values of X) + (Sum of two correct values of X)

= 125 (6 + 8) + (8 + 6) = 125

Similarly

ΣY = 100 (14 + 6) + (12 + 8) = 100

ΣX^2 = 650 $(6^2 + 8^2) + (8^2 + 6^2) = 650$

ΣY^2 = 460 $(14^2 + 6^2) + (12^2 + 8^2)$

= 460 232 + 208 = 436 and

= 508   132 + 144 = 520 ;

Correct value of the correlation co efficient is

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

$$= \frac{25 \times 520 - 125 \times 100}{\sqrt{25 \times 650 - 125^2} \sqrt{25 \times 436 - 100^2}}$$

2/3

Rank Correlation Coefficient

Simple correlation coefficient (or product-moment correlation coefficient) is based on the magnitudes of the variables. But in many situations it is not possible to find the magnitude of the variable at all. For example, we cannot measure beauty or intelligence quantitatively. In this case, it is possible to rank the individuals in some order. Rank correlation is based on the rank or the order and not on the magnitude of the variable. It is more suitable if the individuals (or variables) can be arranged in order of merit or proficiency. If the ranks assigned to individuals range from 1 to n, then the Karl Pearson's correlation coefficient between two series of ranks is called Rank correlation coefficient. Edward Spearman's formula for Rank correlation coefficient (R) is given by.

$$R = 1 - \frac{6 \Sigma d^2}{n(n^2-1)} \quad or \quad 1 - \frac{6 \Sigma d^2}{(n^3 - n)}$$

where d is the difference between the ranks of the two series and n is the number of individuals in each series.

Derivation of Spearman's Formula for Rank Correlation Coefficient

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

Proof:

Let $(x_1, y_1), (x_2, y_2), \ldots. (x_n, y_n)$ be the ranks of n individuals in two characters (or series) Edward Spearman's Rank correlation coefficient R is the product-moment correlation coefficient between these ranks and, therefore, we can write.

$$R = \frac{Cov(x, y)}{\sigma_x \sigma_y} \qquad \ldots(1)$$

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

But the ranks of n individuals are the natural numbers 1, 2,.... n arranged in some order depending on the qualities of the individuals.

$x_1, x_2, \ldots, x_n$ are the numbers 1, 2... n in some order.

$$\therefore \Sigma x = 1 + 2 + \ldots. + n = \frac{n(n+1)}{2} \text{ and}$$

$$\Sigma x^2 = 1^2 + 2^2 + \ldots. + n^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{n+1}{2}, \quad \frac{\Sigma x^2}{n} = \frac{(n+1)(2n+1)}{6}$$

$$\therefore \sigma_x^2 = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2 = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}$$

$$= \left(\frac{n+1}{12}\right)(4n + 2 - 3n - 3) = \frac{n^2 - 1}{12}$$

similarly,

$$\bar{y} = \frac{n+1}{2} \text{ and } \sigma_y^2 = \frac{n^2-1}{12}$$

Let $d_i = x_i - y_i$; then $d_i = (x_i - \bar{x}) - (y_i - \bar{y})$ [$\bar{x} = \bar{y}$]

Calculate the rank correlation coefficient.

$$\therefore \frac{\Sigma d_i^2}{n} = \frac{\Sigma\{(x_i-\bar{x})-(y_i-\bar{y})\}^2}{n}$$

$$= \frac{\Sigma(x_i-\bar{x})^2}{n} + \frac{\Sigma(y_i-\bar{y})^2}{n} - \frac{2\,\Sigma(x_i-\bar{x})(y_i-\bar{y})}{n}$$

$$\sigma_X^2 + \sigma_Y^2 - 2\operatorname{cov}(x, y)$$

or, 2cov (x, y) $= \frac{n^2-1}{12} + \frac{n^2-1}{12} - \frac{\Sigma d_i^2}{n} = \frac{2(n^2-1)}{12} - \frac{\Sigma d_i^2}{n}$

or, cov (x, y) $= \frac{n^2-1}{12} - \frac{\Sigma d_i^2}{2n}$

Hence, from (1), we get

$$R = \left(\frac{n^2-1}{12} - \frac{\Sigma d^2}{2n}\right) \Big/ \left(\frac{n^2-1}{12}\right)$$

$$= \mathbf{1 - \frac{6\,\Sigma d^2}{n(n^2-1)}} \quad \text{[omitting i]}$$

Example 13

Student (Roll No.)	1	2	3	4	5	6	7	8	9	10
Marks in Maths.	78	36	98	25	75	82	90	62	65	69
Marks in Stat.	84	51	91	60	68	62	86	58	53	47

Solution

In Mathematics, Student with Roll No. 3 gets the highest mark 98 and is ranked 1; Roll No. 7 securing 90 marks has rank 2 and so on. Similarly, we can find the ranks of students in statistics.

Roll No.	Mathematics Marks	Rank (x)	Statistics Marks	Rank (y)	Rank Diff. d = x − y	d^2
1	78	4	84	3	1	1
2	36	9	51	9	0	0
3	98	1	91	1	0	0
4	25	10	60	6	4	16
5	75	5	68	4	1	1
6	82	3	62	5	2	4
7	90	2	86	2	0	0
8	62	7	58	7	0	0
9	65	6	53	8	2	4
10	39	8	47	10	2	4
Total						$30 = \Sigma d^2$

Applying Edward Spearman’s formula:

$$R = 1 - \frac{6\,\Sigma d^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 30}{10(10^2-1)} = 1 - \frac{18}{99}$$

$$= 1 - \frac{2}{11} = \frac{9}{11} = \mathbf{0.82}$$