Correlation

BS unit 5

We often encounter situations where data appears as pairs of figures relating to two variables. Such type of data are called Bi-variate data. In this type of data, one may be interested to know the relationship between the two variables. If a change in one variable is followed by a change in the other variable, they are said to be co-related and there exists correlation between them.

Types of Correlation

There are two types of Correlation

1. **Positive or Direct**: If an increase (or decrease) in one variable is followed by an increase (or decrease) in the other variable, the variables are said be directly proportional and the correlation is called Positive Correlation . In this case both variables deviate in the same direction

Ex: Advertising expenditure and sales

Heights and weights of persons

2. **Negative or Inverse**: If an increase (or decrease) in one variable is followed by an decrease (or increase) in the other variable, the variables are said be inversely proportional and the correlation is called Negative Correlation . In this case both variables deviate in opposite direction.

Ex: Price and demand of an item

- Correlation coefficient is denoted by ‘r’ and the limits are (-1 , 1)
- If the value lies between 0 to 1, correlation is +ve and if the value lies between -1 to 0, then correlation is –ve.
- If the variables are independent, r = 0.
- If r = 1, it is called perfect +ve correlation
- If r =-1, it is called perfect –ve correlation.

Scatter diagram

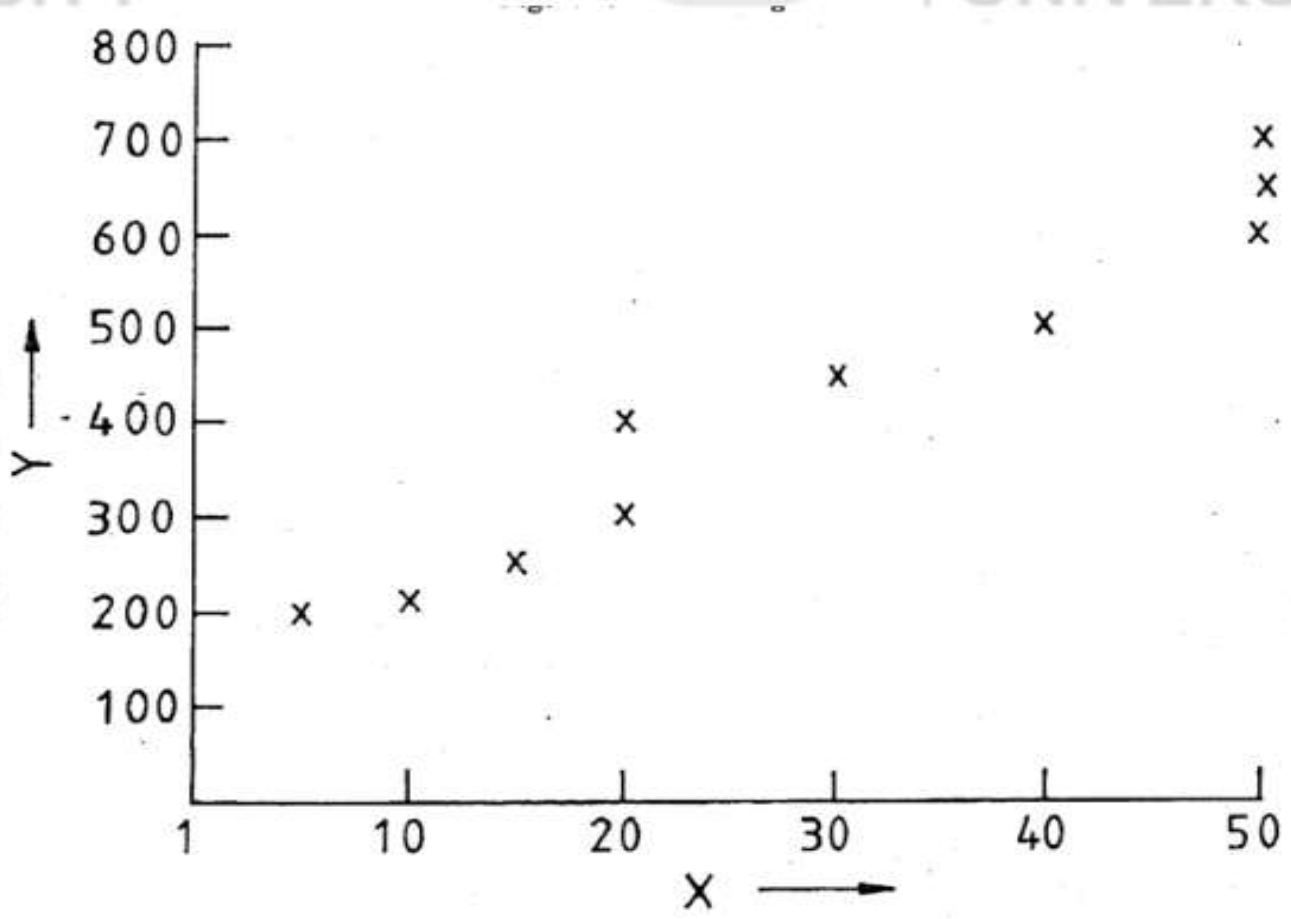
- For example, we may have figures on advertisement expenditure (X) and Sales (Y) of a firm for the last ten years, as shown in Table I. When this data is plotted on a graph as in Figure I we obtain a **scatter diagram.** A scatter diagram gives two very useful types of information. First, we can observe patterns between variables that indicate whether the variables are related. Secondly, if the variables are related we can get an idea of what kind of relationship (linear or non-linear) would describe the relationship. Correlation examines the first question of determining whether an association exists between the two variables, and if it does, to what extent.

Table 1

Yearwise data on Advertisement Expenditure and Sales

Year	Advertisement Expenditure in thousand Rs. (X)	Sales in Thousand Rs. (Y)
1988	50	700
1987	50	650
1986	50	600
1985	40	500
1984	30	450
1983	20	400
1982	20	300
1981	15	250
1980	10	210
1979	5	200

Figure I: Scatter Diagram



Karl Pearson's Correlation coefficient

If X and Y are two random variables then correlation coefficient between X and Y is denoted by r and defined as

$$r = \mathrm{Corr}(x,y) = \frac{\mathrm{Cov}(x,y)}{\sqrt{V(x)V(y)}} \qquad \ldots(1)$$

$\mathrm{Corr}(x,y)$ is indication of correlation coefficient between two variables X and Y.

Where, $\mathrm{Cov}(x,y)$ the covariance between X and Y which is defined as:

$$\mathrm{Cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

and $V(x)$ the variance of X, is defined as:

$$V(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

Similarly,

$V(y)$ the variance of Y is defined by

$$V(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

where, n is number of paired observations.

Then, the correlation coefficient "r" may be defined as:

$$r = Corr(x, y) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}} \quad \ldots (2)$$

Karl Pearson's correlation coefficient r is also called product moment correlation coefficient. Expression in equation (2) can be simplified in various forms. Some of them are

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}} \quad \ldots (3)$$

or

or

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}}{\sqrt{\left\{\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2\right\}\left\{\frac{\sum_{i=1}^{n} y_i^2}{n} - \bar{y}^2\right\}}} \quad \ldots (4)$$

or

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sqrt{\left\{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right\}\left\{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right\}}} \quad \ldots (5)$$

As correlation measures the degree of linear relationship, different values of coefficient of correlation can be interpreted as below:

Value of correlation coefficient	**Correlation is**
+1	Perfect Positive Correlation
−1	Perfect Negative Correlation
0	There is no Correlation
0 - 0.25	Weak Positive Correlation
0.75 - (+1)	Strong Positive Correlation
−0.25 - 0	Weak Negative Correlation
−0.75 - (−1)	Strong Negative Correlation

Example 1: Find the correlation coefficient between advertisement expenditure and profit for the following data:

Advertisement expenditure	30	44	45	43	34	44
Profit	56	55	60	64	62	63

Solution: To find out the correlation coefficient between advertisement expenditure and profit, we have Karl Pearson's formula in many forms [(2), (3), (4), (5) and (6)] and any of them can be used. All these forms provide the same result. Let us take the form of equation (3) to solve our problem which is

$$r = \mathrm{Corr}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}}$$

Steps for calculation are as follow:

1. In columns 1 and 2, we take the values of variables X and Y respectively.
2. Find sum of the variables X and Y i.e.

$$\sum_{i=1}^{6} x_i = 240 \text{ and } \sum_{i=1}^{6} y_i = 360$$

3. Calculate arithmetic means of X and Y as

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{6} x_i}{6} = \frac{240}{6} = 40$$

and $$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{6} y_i}{6} = \frac{360}{6} = 60$$

4. In column 3, we take deviations of each observations of X from mean of X, i.e. $30 - 40 = -10$, $44 - 40 = 4$ and so on other values of the column can be obtained.
5. Similarly column 5 is prepared for variable Y i.e.

$$56 - 60 = -4,\ 55 - 60 = -5$$

and so on.

6. Column 4 is the square of column 3 and column 6 is the square of column 5.
7. Column 7 is the product of column 3 and column 5.
8. Sum of each column is obtained and written at the end of column.

To find out the correlation coefficient by above formula, we require the values of $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$, $\sum_{i=1}^{n}(x_i-\bar{x})^2$ and $\sum_{i=1}^{n}(y_i-\bar{y})^2$ which are obtained by the following table:

x	y	$(x-\bar{x})$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
30	56	−10	100	−4	16	40
44	55	4	16	−5	25	−20
45	60	5	25	0	0	0
43	64	3	9	4	16	12
34	62	−6	36	2	4	−12
44	63	4	16	3	9	12
$\sum x_i = 240$	$\sum_{i=1}^{6} y_i = 360$	$\sum_{i=1}^{6}(x_i-\bar{x}) = 0$	$\sum_{i=1}^{6}(x_i-\bar{x})^2 = 202$	$\sum_{i=1}^{6}(y_i-\bar{y}) = 0$	$\sum_{i=1}^{7}(y_i-\bar{y})^2 = 70$	$\sum_{i=1}^{6}(x_i-\bar{x})(y_i-\bar{y}) = 32$

Taking the values of $\sum_{i=1}^{6}(x_i-\bar{x})(y_i-\bar{y})$, $\sum_{i=1}^{6}(x_i-\bar{x})^2$ and $\sum_{i=1}^{6}(y_i-\bar{y})^2$ from the table and substituting in the above formula we have the correlation coefficient

$$r = \mathrm{Corr}(x,y) = \frac{\sum_{i=1}^{6}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\left\{\sum_{i=1}^{6}(x_i-\bar{x})^2\right\}\left\{\sum_{i=1}^{6}(y_i-\bar{y})^2\right\}}}$$

$$r = \mathrm{Corr}(x,y) = \frac{32}{\sqrt{202\times 70}} = \frac{32}{\sqrt{14140}} = \frac{32}{118.91} = 0.27$$

Hence, the correlation coefficient between expenditure on advertisement and profit is 0.27. This indicates that the correlation between expenditure on advertisement and profit is positive and we can say that as expenditure on advertisement increases (or decreases) profit increases (or decreases). Since it lies between 0.25 and 0.5 it can be considered as week positive correlation coefficient.

Example 2: Calculate Karl Pearson's coefficient of correlation between price and demand for the following data.

Price	17	18	19	20	22	24	26	28	30
Demand	40	38	35	30	28	25	22	21	20

Solution: In Example 1, we used formula given in equation (3) in which deviations were taken from mean. When means of x and y are whole number, deviations from mean makes calculation easy. Since, in Example 1, means x and y were whole number we preferred formula given in equation (3). When means are not whole numbers calculation by formula given in equation (3) becomes cumbersome and we prefer any formula given in equation (4) or (5) or (6). Since here means of x and y are not whole number, so we are preferring formula (6)

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left\{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right\}\left\{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right\}}}$$

Let us denote price by the variable X and demand by variable Y.

To find the correlation coefficient between price i.e.X and demand Y using formula given in equation (6), we need to calculate, $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} x_i y_i$, $\sum_{i=1}^{n} x_i^2$ and $\sum_{i=1}^{n} y_i^2$ which are being obtained in the following table:

x	y	x^2	y^2	xy
17	40	289	1600	680
18	38	324	1444	684
19	35	361	1225	665
20	30	400	900	600
22	28	484	784	616
24	25	576	625	600
26	22	676	484	572
28	21	784	441	588
30	20	900	400	600
$\sum x = 204$	$\sum y = 259$	$\sum x^2 = 4794$	$\sum y^2 = 7903$	$\sum xy = 5605$

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left\{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right\}\left\{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right\}}}$$

$$r = \mathrm{Corr}(x, y)$$

$$= \frac{(9\times 5605) - (204)(259)}{\sqrt{\{(9\times 4794) - (204\times 204)\}\{(9\times 7903) - (259\times 259)\}}}$$

$$r = \mathrm{Corr}(x, y) = \frac{50445 - 52836}{\sqrt{(43146 - 41616)\times(71127 - 67081)}}$$

$$r = \mathrm{Corr}(x, y) = \frac{-2391}{\sqrt{1530\times 4046}}$$

$$r = \mathrm{Corr}(x, y) = \frac{-2391}{2488.0474}$$

$$r = \mathrm{Corr}(x, y) = -0.96$$

Short-cut method

Simplification in computations can be adopted by calculating the deviations of the observations from an assumed average rather than the, actual average, and also scaling these deviations conveniently. To illustrate this short cut procedure, let us compute the correlation coefficient for the same data. We shall take U to be the deviation of X values from the assumed mean of 30 divided by 5. Similarly, V represents the deviation of Y values from the assumed mean of 400 divided by 10.

Table 3

Short cut Procedure for Calculation of Correlation Coefficient

S.No	X	y	U	V	UV	U^2	V^2
1.	50	700	4	30	120	16	900
2.	50	650	4	25	100	16	625
3.	50	600	4	20	80	16	400
4.	40	500	2	10	20	4	100
5.	30	450	0	5	0	0	25
6.	20	400	-2	0	0	4	0
7.	20	300	-2	-10	20	4	100
8.	15	250	-3	-15	45	9	225
9.	10	210	-4	-19	76	16	361
10.	5	200	-5	-20	100	25	400
Total			-2	26	561	110	3,13

$$r = \frac{\Sigma UV - \frac{\Sigma U \Sigma V}{n}}{\sqrt{\Sigma U^2 - \frac{(\Sigma U)^2}{n}} \sqrt{\Sigma V^2 - \frac{(\Sigma V)^2}{n}}}$$

$$561 - \frac{(-2)(26)}{10}$$

$$r = \frac{}{\sqrt{110 - \frac{(-2)^2}{10}} \sqrt{3136 - \frac{(26)^2}{10}}}$$

$$= \frac{566.2}{10.47 \times 55.39}$$

$$= 0.976$$

We thus obtain the same result as before.

Rank Correlation

Quite often data is available in the form of some ranking for different variables. It is common to resort to rankings on a preferential basis in areas such as food testing, competitive events (e.g. games, fashion shows, or beauty contests) and attitudinal surveys. The primary purpose of computing a correlation coefficient in such situations is to determine the extent to which the two sets of rankings are in agreement. The coefficient that is determined from these ranks is known as Spearman's rank correlation coefficient, r.

This is given by the following formula

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \qquad \ldots\ldots (18.4)$$

Here n is the number of pairs of observations and d_i is the difference in ranks for the ith observation set.

Suppose the ranks obtained by a set of ten students in a Mathematics test (variable X) and a Physics test (variable Y) are as shown below :

Rank for variable X	1	2	3	4	5	6	7	8	9	10
Rank for variable Y	3	1	4	2	6	9	8	10	5	7

To determine the rank correlation, r_s we can organise computations as shown in Table 4 :

Table 4

Determination of Spearman's Rank Correlation

Individual	Rank in Maths(X)	Rank in Physics(Y)	d =Y -X	d^2
1	**1**	3	+2	4
2	2	1	-I	1
3	3	4	+1	1
4	4	2	-2	4
5	5	6	+1	1
6	6	9	+3	9
7	7	8	+1	1
8	8	10	+2	4
9	9	5	-4	16
10	10	7	-3	9
Total				50

Using the formula (18.4) we obtain

$$r_s = 1 - \frac{6 \times 50}{10(100-1)} = 1 - 0.303 = 0.697$$

We can thus say that there is a high degree of correlation between the performance in Mathematics and Physics.

Example 1: Suppose we have ranks of 8 students of B.Sc. in Statistics and Mathematics. On the basis of rank we would like to know that to what extent the knowledge of the student in Statistics and Mathematics is related.

Rank in Statistics	1	2	3	4	5	6	7	8
Rank in Mathematics	2	4	1	5	3	8	7	6

Solution: Spearman's rank correlation coefficient formula is

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

Let us denote the rank of students in Statistics by R_x and rank in Mathematics by R_y. For the calculation of rank correlation coefficient we have to find $\sum_{i=1}^{n} d_i^2$ which is obtained through the following table:

Rank in Statistics (R_x)	Rank in Mathematics (R_y)	Difference of Ranks ($d_i = R_x - R_y$)	d_i^2
1	2	−1	1
2	4	−2	4
3	1	2	4
4	5	−1	1
5	3	2	4
6	8	−2	4
7	7	0	0
8	6	2	4
			$\sum d_i^2 = 22$

Here, n = number of paired observations = 8

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 22}{8 \times 63} = 1 - \frac{132}{504} = \frac{372}{504} = 0.74$$

Without repeated observations

Example 3: Calculate rank correlation coefficient from the following data:

x	78	89	97	69	59	79	68
y	125	137	156	112	107	136	124

Solution: We have some calculation in the following table:

x	y	Rank of x (R_x)	Rank of y (R_y)	$d = R_x - R_y$	d^2
78	125	4	4	0	0
89	137	2	2	0	0
97	156	1	1	0	0
69	112	5	6	−1	1
59	107	7	7	0	0
79	136	3	3	0	0
68	124	6	5	1	1
					$\sum_{i=1}^{n} d_i^2 = 2$

Spearman's Rank correlation formula is

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

$$r_s = 1 - \frac{6 \times 2}{7(49-1)} = 1 - \frac{12}{7 \times 48}$$

$$= 1 - \frac{1}{28} = \frac{27}{28} = 0.96$$

Rank correlation with repeated observations

If two or more individuals have same value, in this case common ranks are assigned to the repeated items. This common rank is the average of ranks they would have received if there were no repetition. For example we have a series 50, 70, 80, 80, 85, 90 then 1st rank is assigned to 90 because it is the biggest value then 2nd to 85, now there is a repetition of 80 twice. Since both values are same so the same rank will be assigned which would be average of the ranks that we would have assigned if there were no repetition. Thus, both 80 will receive the average of 3 and 4 i.e. (Average of 3 & 4 i.e. (3 + 4) / 2= 3.5) 3.5 then 5th rank is given to 70 and 6th rank to 50. Thus, the series and ranks of

Series	50	70	80	80	85	90
Ranks	6	5	3.5	3.5	2	1

In the above example 80 was repeated twice. It may also happen that two or more values are repeated twice or more than that.

For example, in the following series there is a repetition of 80 and 110. You observe the values, assign ranks and check with following.

Series	50	70	80	90	80	120	110	110	110	100
Ranks	10	9	7.5	6	7.5	1	3	3	3	5

When there is a repetition of ranks, a correction factor $\frac{m(m^2-1)}{12}$ is added to $\sum d^2$ in the Spearman's rank correlation coefficient formula, where m is the number of times a rank is repeated. It is very important to know that this correction factor is added for every repetition of rank in both characters.

Example 4: Calculate rank correlation coefficient from the following data:

Expenditure on advertisement	10	15	14	25	14	14	20	22
Profit	6	25	12	18	25	40	10	7

Solution: Let us denote the expenditure on advertisement by x and profit by y

x	**Rank of x (R_x)**	**y**	**Rank of y (R_y)**	**$d = R_x - R_y$**	**d^2**
10	8	6	8	0	0
15	4	25	2.5	1.5	2.25
14	6	12	5	1	1
25	1	18	4	−3	9
14	6	25	2.5	3.5	12.25
14	6	40	1	5	25
20	3	10	6	−3	9
22	2	7	7	−5	25
					$\sum d^2 = 83.50$

Here rank 6 is repeated three times in rank of x and rank 2.5 is repeated twice in rank of y, so the correction factor is

$$\frac{3(3^2-1)}{12}+\frac{2(2^2-1)}{12}$$

Hence rank correlation coefficient is

$$r_s = 1-\frac{6\left\{83.50+\frac{3(3^2-1)}{12}+\frac{2(2^2-1)}{12}\right\}}{8(64-1)}$$

$$r_s = 1-\frac{6\left\{83.50+\frac{3\times 8}{12}+\frac{2\times 3}{12}\right\}}{8X\ 63}$$

$$r_s = 1-\frac{6(83.50+2.50)}{504}$$

$$r_s = 1-\frac{516}{504}$$

$$r_s = 1-1.024 = -0.024$$

There is a negative association between expenditure on advertisement and profit.