

## Correlation

There are many situations in business where we are interested to measure the relationship between two variables rather than single variable such as the income and expenditure of certain class of people, price of commodity and amount demanded volume of sales and the experience of salesman of departmental store etc. Pairs of observations of two such variables produce bi-variate distribution. It often required to know how stronger the relationship between two such variables is or it might required to know the impact of change in one variable on another variable. As an example, the following bivariate data show the ages of husbands and wives of 10 married couples.

<b>Husband</b>	36	72	37	36	51	50	47	50	37	41
<b>Wife</b>	35	67	33	35	50	46	47	42	36	41

## Correlation

Correlation is a statistical technique or tools which measure and analyses the degree or extent of linear relationship between two variables.

Correlation thus denotes the interdependence amongst variates. The degrees are expressed by a coefficient which ranges between -1 and +1. The direction of change is indicated by + or - signs.

If the **increase (decrease)** in one variable results in the corresponding **increase (decrease)** in the others i.e. if the changes are in the same directions the variables are **positively correlated**. For example, the heights and weights of a group of persons are positively correlated, advertising and sales.

If the **increase (decrease)** in one variable results in the corresponding **decrease (increase)** in the others i.e. if the changes are in the opposite directions the variables are **negatively correlated**. For example, T.V registration and cinema attendance is negatively correlated.

An absence of correlation is indicated by zero.

Correlation thus expresses the relationship through a relative measure of change and it has nothing to do with the units in which the variables are expressed.

## Uses

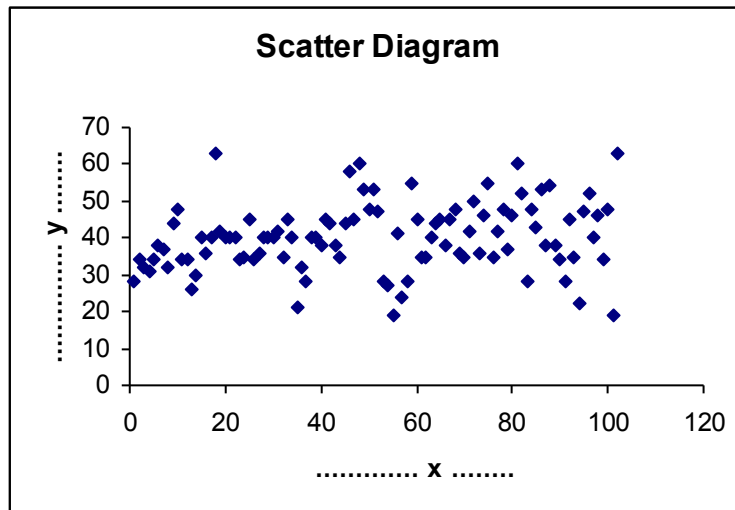
- ➡ Economic theory and business studies relationships between variables like price and quantity demanded, advertising, expenditure scales promotion measure etc. The correlation analysis helps in deriving precisely the degree and direction of such relationships.

➡ The concepts of regression are also based upon the measure of correlation.

## Scatter Diagram

Scatter diagram is a simple and attractive method of diagrammatic represent of bivariate distribution for ascertaining the nature of correlation between the variables. Thus for the bivariate distribution  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$  if the values of the variables  $X$  and  $Y$  be plotted along the  $X$ -axis and  $Y$ -axis respectively in the  $XY$  plane, the diagram of dots so obtained is known as scatter diagram.

On the other hand, a scatter plot of two variables shows the values of one variable on the  $Y$ -axis and the values of the other variable on the  $X$ -axis. Scatter plots are well suited for revealing the relationship between two variables.



## Types of Correlation

Correlation is described or classified in several different ways. Three of the most important are:

- ❖ Simple correlation
- ❖ Partial correlation and
- ❖ Multiple correlation

## Positive and negative correlation

If two variables changes in the same direction (i.e. if one increases the other also increase or if one decreases the other also decreases) then this is called a positive correlation. For example:

Positive Correlation	
X	Y
10	15
12	20
14	22
18	25
20	37

Positive Correlation	
X	Y
80	50
70	45
60	30
40	20
30	10

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

Negative Correlation	
X	Y
20	40
30	30
40	22
60	15
80	12

Negative Correlation	
X	Y
100	10
90	20
60	30
40	40
30	50

## 2. Simple, Partial and Multiple Correlation

- ◆ When only two variables are studied it is a problem of simple correlation.
- ◆ When three or more variables are studied it is a problem of either multiple or partial correlation.

In multiple correlation three or more variables are studied simultaneously. For example, when we study the relationship between the yield of rice per acre and both the amount of rainfall and the amount of fertilizers used, it is problem of multiple correlation. Similarly the relationship of plastic hardness, temperature and pressure is multivariate.

In partial correlation we recognize more than two variables. But consider only two variables to be influencing variable being kept constant. For example, in the rice problem taken above if we limit our correlation analysis of yield and rainfall to periods when a certain average daily temperature existed, it becomes a problem of partial correlation.

## Properties of the Coefficient of Correlation

The following are the important properties of the coefficient of correlation,  $r$  :

- The coefficient of correlation lies between -1 and +1,  $-1 \leq r \leq +1$ .
- It measures magnitude and direction of statistical relationship between two variables.
- It is a pure number. That is, it is independent of units of measurements of the variables.
- Correlation coefficient is symmetric function of the variables. That is  $r_{xy} = r_{yx}$ .
- The coefficient of correlation is the geometric mean of the two regression coefficients. Symbolically:  $r = \sqrt{b_{xy} \times b_{yx}}$
- If X and Y are independent variables then coefficient of correlation is zero. However, the converse is not true.
- Correlation coefficient is independent of origin and scale of measurement.

### Assumption of Simple Correlation:

The simple correlation coefficient  $r$  is based on the following assumption:

- i) The relationship between the variables is linear
- ii) Both the variables are measured on interval or ratio scales
- iii) The two variables follow bi-variate normal distribution

### Degrees of Correlation

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

**Positive Correlation:** If two variables changes in the same direction then the correlation is positive correlation. If the two variables changes in the same unit proportion, then the correlation between the two variables is **perfect positive** and in this case value of Karl Pearson the coefficient of correlation is +1. If the value of Karl Pearson's coefficient of correlation is 0.75 to less than 1 then the correlation is known as strong or high degree correlation, if the value is 0.25 to less than 0.75 then it is known as moderate degree correlation and if the value is 0 to less than 0.25 then it is known as low degree or weak correlation.

**Negative Correlation:** If two variables changes in the opposite direction then the correlation is negative correlation. If the two variables changes in the same unit proportion, then the correlation between the two variables is **perfect negative** and in this case value of Karl Pearson the coefficient of correlation is -1. If the value of Karl Pearson's coefficient of correlation is -0.75 to less than -1 then the correlation is known as strong or high degree correlation, if the value is -0.25 to less than -0.75 then it is known as moderate degree correlation and if the value is 0 to less than -0.25 then it is known as low degree or weak correlation.

**Absence of correlation:** If two series of two variables exhibit no relations between them or change in variable does not lead to a change in the other variable, then we can firmly say that there is **no correlation** or **absurd correlation** between the two variables. In such a case the coefficient of correlation is 0.

High degree, moderate degree or low degrees are the three categories of this kind of correlation. The following table reveals the effect (or degree) of coefficient or correlation.

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

**Methods of Determining Simple Correlation:** For computing or determining the value of simple correlation coefficient, we commonly use the following methods:

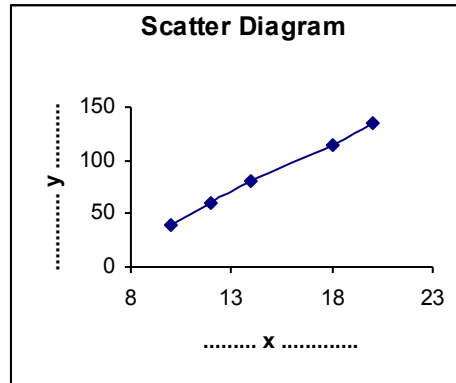
- ◆ Scatter Plot.
- ◆ Karl Pearson's coefficient of correlation.
- ◆ Spearman's Rank-correlation coefficient.
- ◆ Method of Least Squares.

### Scatter Plot (Scatter diagram or dot diagram)

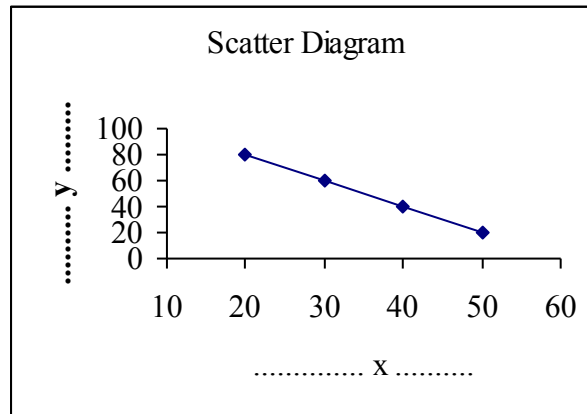
In this method the values of the two variables are plotted on a graph paper. One is taken along the horizontal ( $X$ -axis) and the other along the vertical ( $Y$ -axis). By plotting the data, we get points (dots) on the graph which are generally scattered and hence the name 'Scatter Plot'.

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

- ➡ If all points lie on a rising straight line the correlation is perfectly positive and  $r = +1$ .

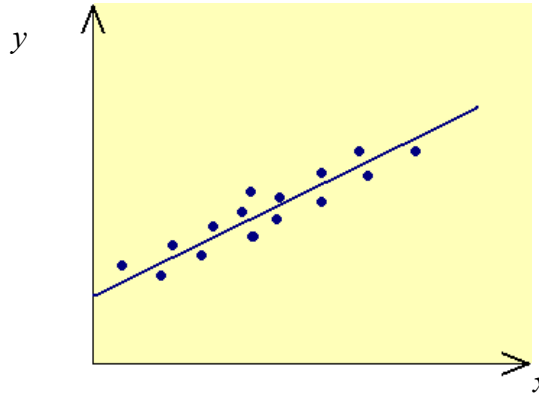


- ➡ If all points lie on a falling straight line the correlation is perfectly negative and  $r = -1$ .



- ➡ If the points lie in narrow strip, rising upwards, the correlation is high degree of positive.

Figure\_C4 Linear correlation between variables.

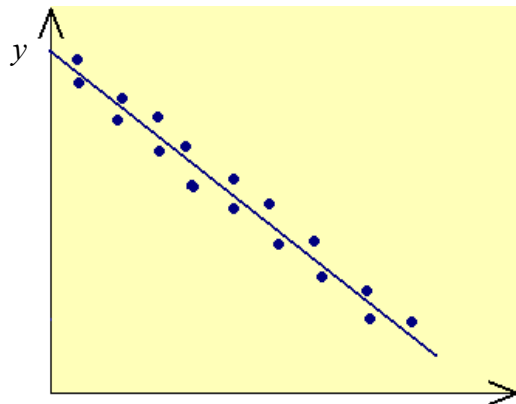


(a) Strong positive linear correlation ( $r$  is close to 1)

50

- ➡ If the points lie in a narrow strip, falling downwards, the correlation is high degree of negative.


Figure\_C6 Linear correlation between variables.

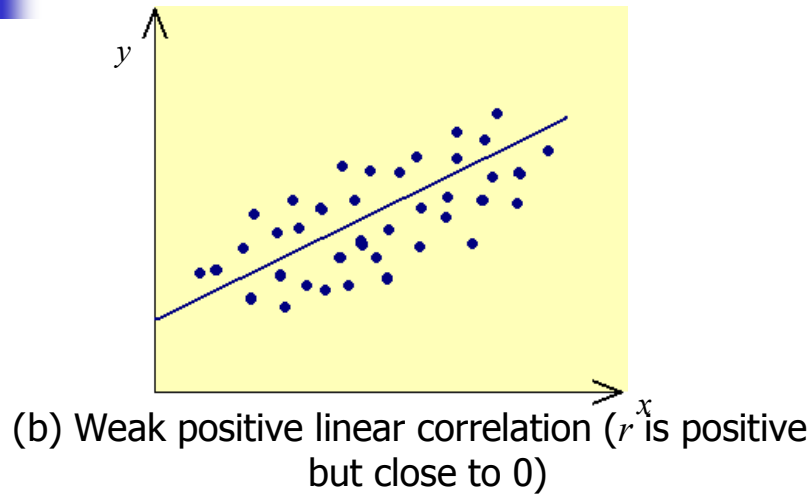


(c) Strong negative linear correlation ( $r$  is close to -1)

52


- ➡ If the points are spread widely over a broad strip, rising upwards, the correlation is low degree positive.

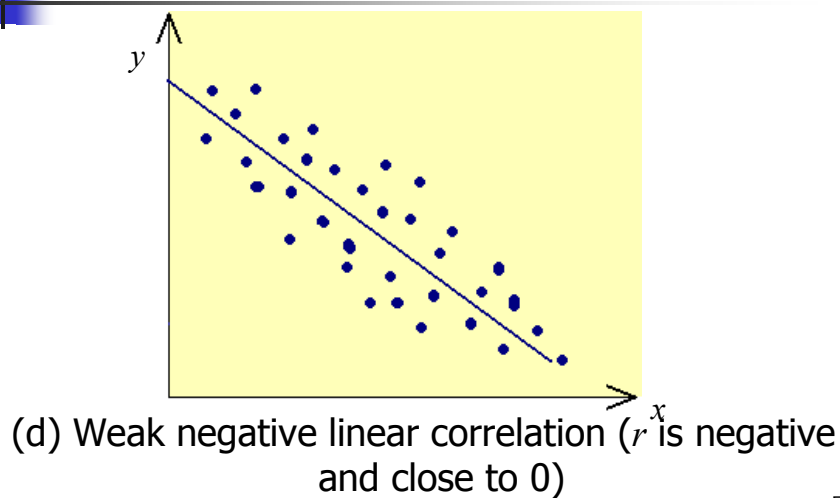
 **Figure\_C5** Linear correlation between variables.



51

- ➡ If the points are spread widely over a broad strip, falling downward, the correlation is low degree negative.

 **Figure\_C7** Linear correlation between variables.



53

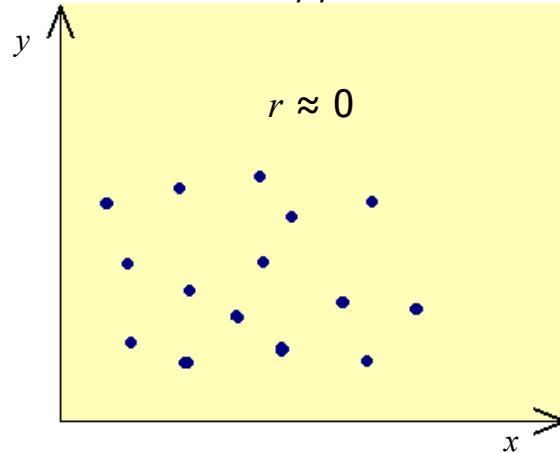


- ➡ If the points are spread (scattered) without any specific pattern, the correlation is absent. i.e.  $r = 0$ .

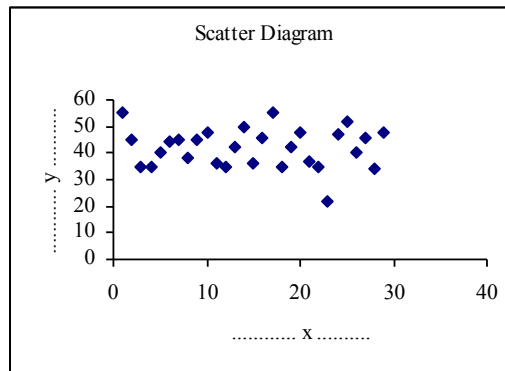


Figure\_C3 Linear correlation between two variables.

(c) No linear correlation,  $r \approx 0$



49



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

### Merits and Limitations of the Method

#### Merits

- ➡ It is simple and non-mathematical method of studying correlation between the variables. As such it can be easily understood and a rough idea can very quickly be formed as to whether or not the variables are related.

- ➡ It is not influenced by the size of extreme values whereas most of the mathematical methods of finding correlation are influenced by extreme values.
- ➡ Making a scatter diagram usually is the first step in investigating the relation ship between the variables.

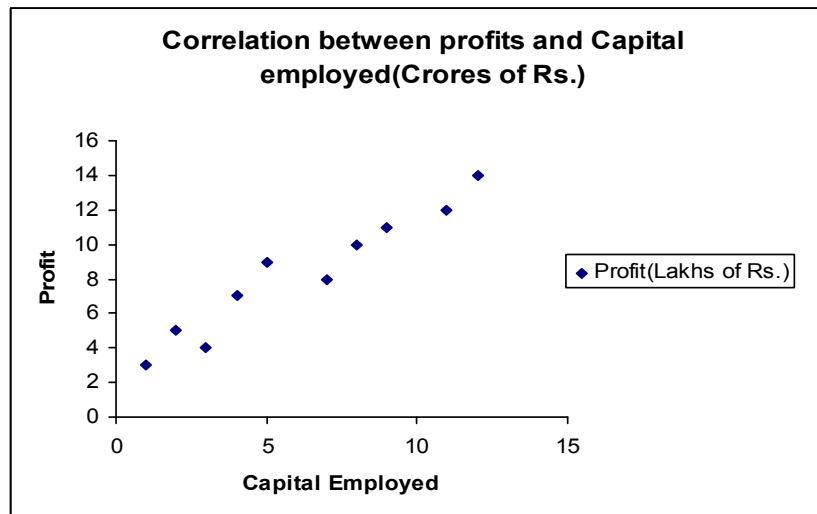
### Limitations

By applying this method we can get an idea about the direction of correlation and also whether it is high or low. But we cannot establish of correlation and also whether it is high or low. But we cannot establish the exact degree of correlation between the variables as is possible by applying the mathematical method.

**Example1:** Given the following pairs of values:

<b>Capital employed (Crores of Rs.):</b>	1	2	3	4	5	7	8	9	11	12
<b>Profit (Lakhs of Rs.)</b>	3	5	4	7	9	8	10	11	12	14

- 1) Make a scatter diagram
- 2) Do you think that there is any correlation between profits and capital employed? Is it positive? Is it high or low?



By looking at the scatter diagram we can say that the variables profits and capital employed are correlated. Further, correlation is positive because the trend to the points is upward rising from the lower left hand corner to the upper right hand corner of the diagram.

The diagram also indicate that the degree of relationship is high because the plotted points are in a narrow band which shows that it is a case of high degree of positive correlation.

### Karl Pearson's Coefficient of Correlation

Of the several mathematical methods of measuring correlation, the Karl Pearson's method, popularly known as Pearsonian coefficient of correlation, is most widely used in practice. The coefficient of correlation is denoted by the symbol  $r$ . If the two variables under study are  $X$  and  $Y$ , the following formula suggested by Karl Pearson can be used for measuring the degree of relationship.

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sqrt{\left[ \sum X^2 - \frac{(\sum X)^2}{n} \right] \left[ \sum Y^2 - \frac{(\sum Y)^2}{n} \right]}}$$

The value of the coefficient of correlation as obtained by the above formula shall always lie between  $\pm 1$ .

When  $r = +1$ , it means there is perfect positive correlation between the variables.

When  $r = -1$ , it means there is a perfect negative correlation between the variables.

When  $r = 0$ , it means there is no relationship between the variables.

**Example1:** Calculate the coefficient of correlation between the heights of father and his son for the following data.

Height of father (cm):	165	166	167	168	167	169	170	172
Height of son (cm):	167	168	165	172	168	172	169	171

**Solution:**

We know that. Correlation of coefficient

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sqrt{\left[ \sum X^2 - \frac{(\sum X)^2}{n} \right] \left[ \sum Y^2 - \frac{(\sum Y)^2}{n} \right]}}$$

Let us consider the height of father is  $X$  and height of son is  $Y$ .

By using calculator we get,

$$\begin{array}{lll} \sum X^2 = 225828 & \sum X = 1344 & n = 8 \\ \sum Y^2 = 228532 & \sum Y = 1352 & \sum XY = 227160 \end{array}$$

$$\therefore r = \frac{227160 - \frac{1344 \times 1352}{8}}{\sqrt{\left[ \left\{ 225828 - \frac{(1344)^2}{8} \right\} \left\{ 228532 - \frac{(1352)^2}{8} \right\} \right]}} = 0.603022689 = 0.603$$

**Example2:** The following data consist of observations for the weights of 10 different automobiles (in 1000 pounds) and the corresponding fuel consumptions (gallons per 100 miles).

Weight (x)	Fuel Consumption (y)
3.4	5.5
3.8	5.9
4.1	6.5
2.2	3.3
2.6	3.6
2.9	4.6
2.0	2.9
2.7	3.6
1.9	3.1
3.4	4.9

We would like to find out how y is correlated to x.

**Solution:** We know that. Correlation of coefficient

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sqrt{\left[ \left\{ \sum X^2 - \frac{(\sum X)^2}{n} \right\} \left\{ \sum Y^2 - \frac{(\sum Y)^2}{n} \right\} \right]}}$$

By using calculator we get,

$$\sum X^2 = 89.29 \quad \sum X = 29 \quad n = 8$$

$$\sum Y^2 = 207.31 \quad \sum Y = 43.9 \quad \sum XY = 135.8$$

$$\therefore r = \frac{135.8 - \frac{29 \times 43.9}{10}}{\sqrt{\left\{ 89.29 - \frac{(29)^2}{10} \right\} \left\{ 207.31 - \frac{(43.9)^2}{10} \right\}}} = 0.976629971 = 0.976$$

**Example3:** Suppose that we took 7 mice and measured their body weight and their length from nose to tail. We obtained the following results and want to know if there is any relationship between the measured variables. [To keep the calculations simple, we will use small numbers]

Mouse	Units of weight (X)	Units of length (Y)
1	1	2
2	4	5
3	3	8
4	4	12
5	8	14
6	9	19
7	8	22

**Solution:** We know that. Correlation of coefficient

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

By using calculator we get,

$$\begin{array}{lll} \sum X^2 = 251 & \sum X = 37 & n = 7 \\ \sum Y^2 = 1278 & \sum Y = 82 & \sum XY = 553 \end{array}$$

$$\therefore r = \frac{553 - \frac{37 \times 82}{7}}{\sqrt{\left\{251 - \frac{(37)^2}{7}\right\} \left\{1278 - \frac{(82)^2}{7}\right\}}} = 0.901441541 = 0.90$$

**Example4:** The data below are the heights (cm) and weights (Kg) of 20 female students taking STAT 201. Calculate the coefficient of correlation between the heights and weights of female students of the following data.

SL	fht	fw
1	167	60
2	164	65
3	170	64
4	163	47
5	152	46
6	160	57
7	170	57
8	160	55
9	157	55
10	170	65
11	150	50
12	156	46
13	168	60
14	159	55
15	160	50
16	172	69
17	175	56
18	169	56
19	169	72
20	156	56

### Solution

We know that. Correlation of coefficient

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

Let us consider fht is denoted by  $X$  and fwt is denoted by  $Y$ .

By using calculator we get,

$$\begin{array}{lll} \sum X^2 = 534615 & \sum X = 3267 & n = 20 \\ \sum Y^2 = 66113 & \sum Y = 1141 & \sum XY = 187045 \end{array}$$

$$\therefore r = \frac{187045 - \frac{3267 \times 1141}{20}}{\sqrt{\left[ \left\{ 534615 - \frac{(3267)^2}{20} \right\} \left\{ 66113 - \frac{(1141)^2}{20} \right\} \right]}} = 0.673318089 = 0.673$$

**Example:** An instructor in a statistics course set a final examination and also required the students to do a data analysis project. For a random sample of 10 students, the scores obtained are shown in the table. Find the simple correlation between the examination and project scores and comment on your result.

Examination	181	162	174	178	193	169	172	183	190	184
Project	176	171	169	176	187	162	180	175	192	179

### (3)Spearman's Rank Correlation

The association between two series of rank is called rank correlation. The method of ascertaining the coefficient of correlation by ranks was devised by Charles Edwards Spearman in 1904. This method is especially useful in case when the actual magnitudes or item values are not given and simply their ranks in the series are known. Spearman's rank correlation coefficient, usually denoted by  $\rho$  (Rho) is given by the formula:

$$\rho = 1 - \frac{6 \sum d_i^2}{n^3 - n} = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Where,  $d$  stands for the difference between the pair of ranks and  $n$  the number of paired observations.

The value of Spearman's rank correlation coefficient ranges between -1 and +1. When  $\rho$  is +1, the concordance between rankings is perfect and the ranks are in the same direction. When  $\rho$  is -1, there is also perfect concordance between rankings but the ranks are in opposite direction.

In rank correlation we may have two types of problems:

A. Where actual ranks are given.

B. Where ranks are not given.

### A. Where Actual Ranks are given

Where Actual Ranks are given the steps required for computing rank correlation are:

- Take the difference of the two ranks i.e  $(R_1 - R_2)$  and denote these differences by  $d$ .
- Square these differences and obtain the total  $\sum d_i^2$
- Apply the formula

$$\rho = 1 - \frac{6 \sum d_i^2}{n^3 - n}$$

#### Example1:

Two managers are asked to rank a group of employees in order of potential for eventually becoming top managers. The rankings are as follows:

Employee	Ranked by manager I	Ranked by Manager II
A	10	9
B	2	4
C	1	2
D	4	3
E	3	1
F	6	5
G	5	6
H	8	8
I	7	7
J	9	10

Compute the coefficient of rank correlation and comment on the value.

**Solution:**

#### Calculation of Rank Correlation Coefficient

Employee	Ranked by manager I ( $R_1$ )	Ranked by Manager II ( $R_2$ )	
----------	----------------------------------	-----------------------------------	--



			$d^2 = (R_1 - R_2)^2$
A	10	9	
B	2	4	By using
C	1	2	Calculator
D	4	3	
E	3	1	
F	6	5	
G	5	6	
H	8	8	
I	7	7	
J	9	10	
Total			$\sum d_i^2 = 14$

We know that,

$$\rho = 1 - \frac{6 \sum d_i^2}{n^3 - n} = 1 - \frac{6 \times 14}{10^3 - 10} = 0.915$$

Thus we find that there is a high degree of positive correlation in the ranks assigned by the two managers.

### B. Where Ranks are not given

When we are given the actual data and not the ranks it will be necessary to assign the ranks. Ranks can be assigned by taking either the highest value as 1 or the lowest value as 1. But whether we start with the lowest value or the highest value we must follow the same method in case of all the variables.

#### Example1:

Calculate the rank correlation coefficient for the following data of marks of 2 tests given to candidates for a clerical job:

<b>Preliminary test</b>	92	89	87	86	83	77	71	63	53	50
<b>Final test</b>	86	83	91	77	68	85	52	82	37	57

**Solutions:**

#### Calculation of Rank Correlation Coefficient

Preliminary test	$R_1$	Final test	$R_2$	$d^2 = (R_1 - R_2)^2$
92	10	86	9	
89	9	83	7	
87	8	91	10	

86	7	77	5	By using
83	6	68	4	Calculator
77	5	85	8	
71	4	52	2	
63	3	82	6	
53	2	37	1	
50	1	57	3	
Total				$\sum d_i^2 = 44$

We know that,

$$\rho = 1 - \frac{6 \sum d_i^2}{n^3 - n} = 1 - \frac{6 \times 44}{10^3 - 10} = 1 - 0.267 = 0.733$$

Thus there is a high degree of positive correlation between preliminary and final test.

### Merits and Limitations of the Rank Method

#### Merits

- This method is simpler to understand and easier to apply compared to the Karl Pearson's method.
- Rank correlation can be safely used in case of linear and curvilinear relationship between two variables x and y. But simple correlation coefficient measures only the strength of linear relationship between two variables.
- Where the data are of a qualitative nature like honesty, efficiency, intelligence etc., this method can be used with great advantage. It also applied for quantitative data but simple correlation is only applicable for quantitative data.
- This is the only method that can be used where we are given the ranks and not the actual data.
- Even where actual data are given rank method can be applied for ascertaining rough degree of correlation.

#### Limitations:

- This method cannot be used for finding out correlation in a grouped frequency distribution.
- Where the number of observations exceed 30 the calculations becomes quite tedious and require a lot of time. Therefore this method should not applied where

n exceeding 30 unless we are given the ranks and not the actual values of the variable.

Example: The scores of eight students of business statistics and business mathematics in an examination are given below:

Marks in business statistics	52	60	50	54	55	58	48	70
Marks in business math	48	51	68	55	60	53	47	62

Compute the coefficient of rank correlation and comment on your result.

#### (4) Method of Least Squares

For finding out correlation by the coefficient method of least squares we have to calculate the values of two regression coefficients that of  $x$  on  $y$  and  $y$  on  $x$ . The correlation coefficient is the square root of the product of two regression coefficients. Symbolically,

$$r = \sqrt{b_{xy} \times b_{yx}}$$

#### Coefficient of Determination

One very convenient and useful way of interpreting the value of coefficient of correlation between two variables is to use the square of coefficient of correlation, which is called coefficient of determination. The coefficient of determination thus equals  $r^2$ .

\*\*\* If the value of  $r = 0.9$ ,  $r^2$  will be 0.81 and this would mean that 81% of the variation in the dependent variable has been explained by the independent variable.