CORRELATION

In a bivariate distribution, if the change in one variable affect a change in the other variable, the variables are said to be correlated.

If the two variables deviates in the same direction i.e if the increase (or decrease) in one results in a corresponding increase (or decrease) in the other, correlation is said to be **direct or positive**.
e.g. the correlation between income & expenditure is positive.

If the two variables deviate in opposite directions i.e if the increase (or decrease) in one results in a corresponding decrease (or increase) in the other, correlation is said to be **inverse or negative**.
e.g. the correlation between volume & pressure of a perfect gas or the correlation between price and demand is negative.

Correlation is said to be **perfect** if the deviation in one variable is followed by a corresponding proportional deviation in the other.

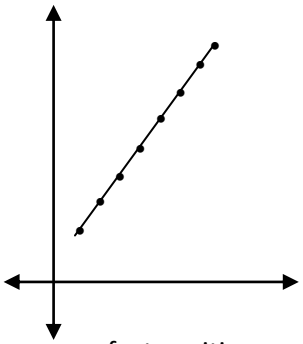
If there is no relationship between the two variables, they are said to be **independent.**

Generally, when two variables are correlated, one of them is the cause and the other is the effect. For example, we know that rainfall and production of paddy are correlated and we also know that the production of paddy is the effect and the rainfall is the cause. In some cases we may find correlation and yet there may not be any causal relation or we may know causal relation and yet we may not find correlation there. Such a false correlation is called **'Spurious Correlation'**. This happens because of the following reasons.

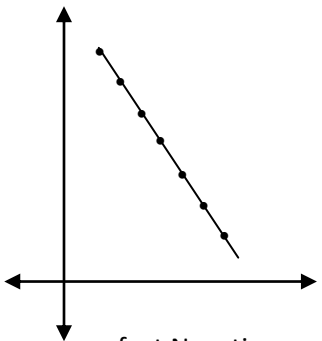
SCATTER OR DOT DIAGRAMS:

It is the simplest method of the diagrammatic representation of bivariate data. Let $(x_i, y_i), i = 1,2,3 \dots \dots n$ be a bivariate distribution. Let the values of the variables x and y be plotted along the x – axis and y – axis on a suitable scale. Then corresponding to every ordered pair, there corresponds a point or dot in the xy – plane. The diagram of dots so obtained is called a **dot or scatter diagram.**

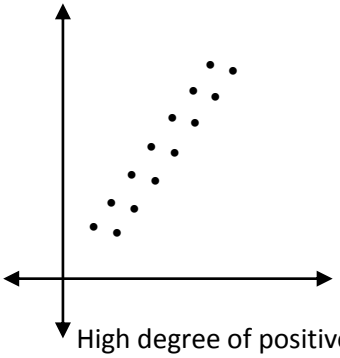
The way in which the points are scattered indicates the degree and direction of correlation. If the points are close to each other and the number of observation is not very large than a fairly good correlation is expected. If the points are widely scattered then a poor correlation is expected.



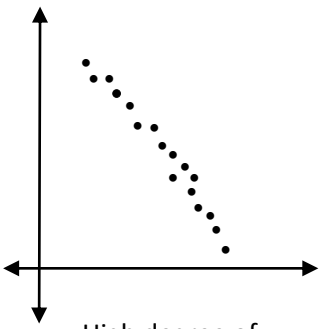
perfect positive correlation r = +1



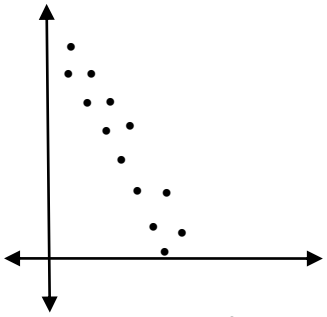
perfect Negative correlation r = -1



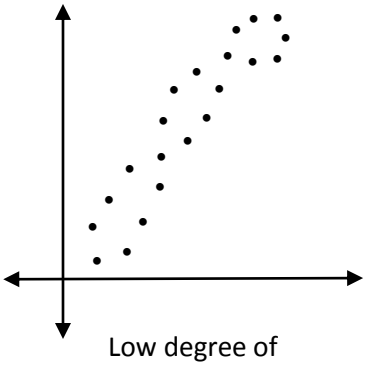
High degree of positive correlation r = +



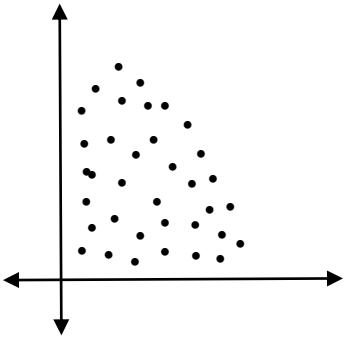
High degree of negativecorrelation r = -



low degree of negative correlation r = -



Low degree of positive correlation



No correlation r = 0

KARL PEARSON'S COEFFICIENT OF CORRELATION:

The method of scatter diagram is descriptive in nature and gives only a general idea of correlation. The most commonly used method which gives a mathematical expression for correlation is the one suggested by Karl Pearson's a British Biometrician.

$\boldsymbol{\sigma_x^2 = \frac{1}{N}\Sigma(x-\bar{x})^2 = \left(\frac{\Sigma x^2}{N}\right) - \bar{x}^2 = \left(\frac{\Sigma x^2}{N}\right) - \left(\frac{\Sigma x}{N}\right)^2}$ gives us a measure of variation in x and

$\boldsymbol{cov(x,y) = \frac{1}{N}\Sigma(x-\bar{x})(y-\bar{y}) = \left(\frac{\Sigma xy}{N}\right) - \bar{x}\bar{y} = \left(\frac{\Sigma x}{N}\right)\left(\frac{\Sigma y}{N}\right)}$ gives a measure of simultaneous variation in x and y.

Karl Pearson suggested the following coefficient of correlation to measure correlation between x and y.

It is denoted by r

$$r = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}} = \frac{\left(\frac{\Sigma xy}{N}\right) - \left(\frac{\Sigma x}{N}\right)\left(\frac{\Sigma y}{N}\right)}{\sqrt{\left(\frac{\Sigma x^2}{N}\right) - \left(\frac{\Sigma x}{N}\right)^2}\sqrt{\left(\frac{\Sigma y^2}{N}\right) - \left(\frac{\Sigma y}{N}\right)^2}} = \frac{N\Sigma xy - \Sigma x \Sigma y}{\sqrt{N\Sigma x^2 - (\Sigma x)^2}\sqrt{N\Sigma y^2 - (\Sigma y)^2}} = \frac{\Sigma xy - N\bar{x}\bar{y}}{\sqrt{\Sigma x^2 - N\bar{x}^2}\sqrt{\Sigma y^2 - N\bar{y}^2}}$$

Theorem: Correlation coefficient is independent of change of origin & change of scale.

It can also stated as If $u = \frac{x-a}{b}$ & $v = \frac{y-c}{d}$ where a, b, c, d are constants then $r_{xy} = r_{uv}$

Theorem: If $Z = ax + by$ and r is the correlation coefficient between x & y prove that

$$\sigma_z^2 = a^2\sigma_x^2 + b^2\sigma_y^2 + 2abr\sigma_x\sigma_y$$

Proof: We know that $\sigma_x^2 = \frac{\Sigma(x-\bar{x})^2}{n}$

$$\sigma_z^2 = \sigma_{ax+by}^2 = \frac{\Sigma\left((ax+by)-(\overline{ax+by})\right)^2}{n} = \frac{\Sigma\left(a(x-\bar{x})+b(y-\bar{y})\right)^2}{n} \qquad \because \overline{ax+by} = a(\bar{x}) + b(\bar{y})$$

$$= \frac{\Sigma\left(a^2(x-\bar{x})^2 + 2ab(x-\bar{x})(y-\bar{y}) + b^2(y-\bar{y})^2\right)}{n}$$

$$= a^2\left(\frac{\Sigma(x-\bar{x})^2}{n}\right) + 2ab\left(\frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}\right) + b^2\left(\frac{\Sigma(y-\bar{y})^2}{n}\right)$$

$$= a^2\sigma_x^2 + 2ab\ cov(x,y) + b^2\sigma_y^2$$

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

$\therefore \sigma_z^2 = a^2\sigma_x^2 + b^2\sigma_y^2 + 2abr\sigma_x\sigma_y$

Theorem: If r is coefficient of correlation than $-1 \le r \le 1$

Proof: r is coefficient of correlation $\quad \therefore r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$

$$\therefore r^2 = \frac{\left(\Sigma(x-\bar{x})(y-\bar{y})\right)^2}{(\Sigma(x-\bar{x})^2)(\Sigma(y-\bar{y})^2)}$$

Since by Cauchy Schwartz's Inequality $\left(\Sigma(x-\bar{x})(y-\bar{y})\right)^2 \le (\Sigma(x-\bar{x})^2)(\Sigma(y-\bar{y})^2)$

$$\therefore \frac{\left(\Sigma(x-\bar{x})(y-\bar{y})\right)^2}{(\Sigma(x-\bar{x})^2)(\Sigma(y-\bar{y})^2)} \le 1 \qquad \therefore \ r^2 \le 1 \qquad \therefore -1 \le r \le 1$$

EXERCISE

1. Prove that $\sigma_{x+y}{}^2 = \sigma_x{}^2 + \sigma_y{}^2 + 2r\sigma_x\sigma_y$ and hence deduce that $r = -\frac{\sigma_x{}^2+\sigma_y{}^2-\sigma_{x+y}{}^2}{2\sigma_x\sigma_y}$

2. Draw a scatter diagram to represent the following data.

X : 2, 4, 5, 6, 8, 11.

Y : 18, 12, 10, 8, 7, 5.

Calculate the coefficient of correlation between X and Y for the above data.

3. Calculate the correlation coefficient from the following data.

X : 23, 27, 28, 29, 30, 31, 33, 35, 36, 39.

Y : 18, 22, 23, 24, 25, 26, 28, 29, 30, 32.

4. Calculate the coefficient of correlation between X and Y from the following data.

X : 3, 5, 4, 6, 2.

Y : 3, 4, 5, 2, 6.

5. Find the coefficient of correlation between height of father and height of son from the data.

Height of father : 65, 66, 67, 67, 68, 69, 71, 73.

Height of son : 67, 68, 64, 68, 72, 70, 69, 70.

6. Find the number of pairs of observations from the data: $r = 0.4, \sum xy = 108, \sigma_y = 3, \sum x^2 = 900$ where x, y are the deviation of x, y from their respective means.

7. Coefficient of correlation between two variables is 0.4. Their covariance is 12. The variance of x is 25. Find the standard deviation of y. Variance is σ_{x^2} We want σ_x

8. From the following data about 8 pairs of observations, calculate the coefficient of correlation between x &y

	x series	y series
Arithmetic Mean	68.25	68.5
S. D.	2.487	2.236

Summation of the product of the deviations of x and y from their respective means =26

9. Given : Number of pairs of observations = 10;

X series standard deviation = 22.70 Y series standard deviation = 9.592;

Summation of the products of corresponding deviations of X and Y from their respective actual

Means $= -1439$. Find r.

10. Calculate the coefficient of correlation between x and y from the following data.
$N = 10,\ \sum x = 225,\ \sum y = 189,\ \sum(x-22)^2 = 85,\ \sum(y-19)^2 = 25, \sum(x-22)(y-19) = 42,$
where x, y denote the actual values.

11. Calculate the coefficient of correlation from the data: $N = 10, \sum x = 100,\ \sum y = 150,\ \sum(x-10)^2 = 180,$ $\sum(y-15)^2 = 215,\ \sum(x-10)(y-15) = 60.$ Where x, y denote the actual values.

12. Calculate the coefficient of correlation from the following data.
$N = 10,\ \sum x = 136,\ \sum y = 243,\ \sum x^2 = 2278,\ \sum y^2 = 6129,\ \sum xy = 3476.$
Where x, y denote the actual values.

13. The variables x and y are connected by the equation $ax + by + c = 0$. Show that the correlation between them is – 1 if the sings of a and b are alike and $+1$ if they are different.

14. A sample of 25 pairs of values of x and y lead to the following results.
$\sum x = 127,\ \sum y = 100,\ \sum x^2 = 760,\ \sum y^2 = 449,\ \sum xy = 500$ Later on it was found that two pairs of values were taken as (8, 14) and (8, 6) instead of correct values (8, 12) and (6, 8). Find corrected correlation coefficient between x and y .

15. For 10 pairs of values of x and y the following values are determined

	x	y	
Mean	30.1	47.8	
S. D.	6.2	9.5	$r = 0.72$

Later on it was found that one pair of values was taken as (34, 47) instead of (43, 74). Determine the correct value of coefficient of correlation.

16. In two sets of variables x and y with 50 observations each gave the results $\bar{x} = 10,\ \bar{y} = 6,\ \sigma_x = 3,$ $\sigma_y = 2, r(x,y) = 0.3$ .But on subsequent verification it was found that one value of $x = 10$ and one value of$y = 6$ were inaccurate and were discarded. With theremaining 49 pairs of values how is the original value of 'r' affected ?

ANSWERS

2.	-0.92	**3.**	0.9948	**4.**	-0.7	**5.**	0.47
6.	N = 9	**7.**	$\sigma_y = 6$	**8.**	0.5844	**9.**	-0.66
10.	0.938	**11.**	0.305	**12.**	0.55	**13.**	-0.31
14.	0.829	**15.**	0.829	**16.**	0.3		

RANK CORRELATION:

Sometimes we have to deal with problems in which data cannot be quantitatively measured but qualitative assessment is possible. Let a group of n individuals be arranged in order of merit of two characteristic A and B. The ranks in the two characteristics, in general, different. Let $(x_i, y_i),\ i = 1,2,3 \ldots . n$ be the ranks of the n individuals in the group for the characteristics A and B respectively.

Pearsonian coefficient of correlation between the ranks $x_i's$ and $y_i's$ is called the rank correlation coefficient between the characteristic A and B for that group of individuals.

The method developed by spearman is simpler than Karl Pearson's method. Since it depends upon ranks of the items and actual values of the items are not required. Hence, this can be used to study correlation even when actual values are not known. For instance we can study correlation between intelligence and honesty by this method.

If d_i denotes the difference in ranks of the i^{th} individuals then $d_i = x_i - y_i$ then we can prove $r = 1 - \frac{6\Sigma d_i^2}{(n^3-n)}$

This is called spearman's formula for Rank correlation. It may be denoted by R.

NOTE: $\Sigma d_i = \Sigma(x_i - y_i) = \Sigma x_i - \Sigma y_i = 0$ always. This serves as a check on calculations.

Ex. Calculate the rank correlation coefficient from the following data

Rank in English	1	3	7	5	4	6	2	10	9	8
Rank in Statistics	3	1	4	5	6	9	7	8	10	2

Solution: Calculation of R

Students No.	Rank in English x_i	Rank in Mathematics y_i	${d_i}^2 = (x_i - y_i)^2$
1	1	3	4
2	3	1	4
3	7	4	9
4	5	5	0
5	4	6	4
6	6	9	9
7	2	7	25
8	10	8	4
9	9	10	1
10	8	2	36
$N = 10$			$\sum {d_i}^2 = 96$

Now, $R = 1 - \frac{6\sum {d_i}^2}{n^3-n} = 1 - \frac{6\times 96}{990} = 1 - \frac{96}{165} = 0.42$

EQUAL RANKS:

In some cases it may happen that there is a tie between two or more members i.e they have equal values and hence equal ranks. In such cases we divide the rank among equal members.

For instance, **(i)** If two items have 4th rank we divide the 4th& next rank 5th between them equally and give $\frac{4+5}{2} = 4.5$th rank to each of them.

(ii) If three items have the same 4th rank, we give each of them $\frac{4+5+6}{3} = 5th$ rank.

After assigning ranks in this way an adjustment is necessary. If m is the number of items having equal ranks then the factor $\frac{1}{12}(m^3 - m)$ is added to Σd_i^2. If there are more than one cases of this type, this factor is added corresponding to each case. Then, $R = 1 - \frac{6\left[\Sigma d_i^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \ldots\ldots\ldots\right]}{n^3-n}$

Ex. Obtain the rank correlation coefficient from the following data

X	10	12	18	18	15	40
Y	12	18	25	25	50	25

Solution:

X	Rank x_i	Y	Rank y_i	$d_i = x_i - y_i$	$d_i^2 = (x_i - y_i)^2$
10	1	12	1	0	0.00
12	2	18	2	0	0.00
18	4.5	25	4	0.5	0.25
18	4.5	25	4	0.5	0.25
15	3	50	6	-3	9.00
40	6	25	4	2	4.00
$N = 6$				$\sum d_i = 0$	$\Sigma d_i^2 = 13.50$

There are two items in X series having equal values at the rank 4. Each is given the rank 4.5.

Similarly, there are three items in Y series at the rank 3. Each of them is given the rank 4.

$$\therefore R = 1 - \frac{6\left[\Sigma d_i^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2)\right]}{n^3 - n}$$

Since, $\Sigma d_i^2 = 13.50,\ m_1 = 2,\ m_2 = 3,\ N = 6$

$$R = 1 - \frac{6\left[13.50 + \frac{1}{12}(8-2) + \frac{1}{12}(27-3)\right]}{216-6} = 1 - 0.4571 = 0.5429$$

EXERCISE

1. Calculate the rank correlation coefficient from the following data.

X : 12 17 22 27 32.

Y : 113 119 117 115 121 .

2. Compute Spearman's rank correlation coefficient from the following data.

X : 85, 74, 85, 50, 65, 78, 74, 60, 74, 90.

Y : 78, 91, 78, 58, 60, 72, 80, 55, 68, 70.

3. The coefficient of rank correlation between marks in Physics and Chemistry obtained by a group of students is 0.8. If the sum of the squares of differences in ranks in 33, find the number of pairs of students.

4. The coefficient of rank correlation of the marks obtained by 10 students in Physics and Chemistry was found to be 0.5. It was later discovered that the difference in rank in the two subjects obtained by one of the students was wrongly taken as 3 instead of 7. Find the correct coefficient of rank correlation.

ANSWERS

1. 0.6 **2.** 0.54 **3.** 10 **4.** 0.25757