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.

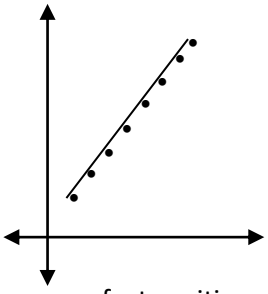
In some cases we may find correlation and yet there may not be any causal relation or we may know causal relation and yet we may not find correlation there. Such a false correlation is called **'Spurious Correlation'** .

SCATTER OR DOT DIAGRAMS:

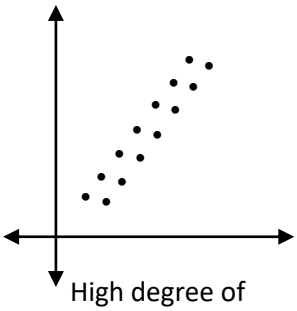
It is the simplest method of the diagrammatic representation of bivariate data.

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.

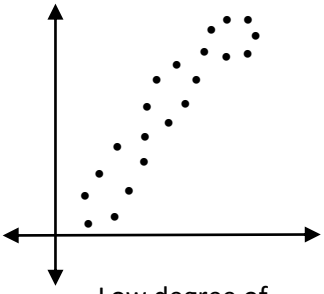
The diagram of dots so obtained is called a **dot or scatter diagram.**



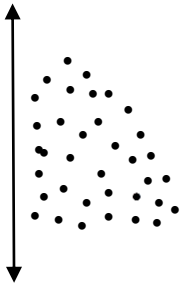
perfect positive correlation r =



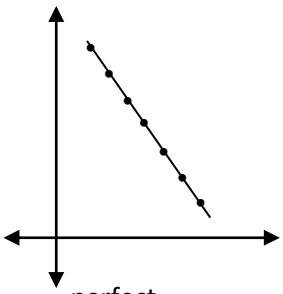
High degree of positive correlation r = +



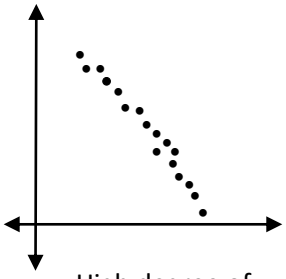
Low degree of positive



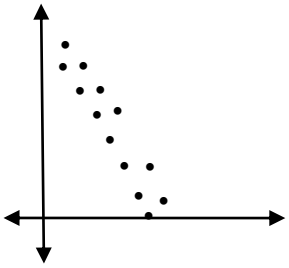
No correlation r



perfect Negative



High degree of negative correlation



low degree of negative correlation

STANDARD DEVIATION :

The root – mean square deviation denoted by s, is defined as the positive square root of the mean of the squares of the deviations from an arbitrary origin A.

Thus $s = +\sqrt{\frac{1}{N}\sum f_i(x_i - A)^2}$

When the deviations are taken from the mean $\overline{x}$, the root – mean square deviation is called the **standard deviation** and is denoted by the Greek letter $\boldsymbol{\sigma}$.

Thus $\sigma = \sqrt{\frac{1}{N}\sum f_i(x_i - \overline{x})^2}$

The square of the standard deviation σ^2 is called **variance**.

(1) $\sigma_x^2 = \frac{1}{N}\Sigma(x - \bar{x})^2$

(2) $\sigma_x^2 = \left(\frac{\Sigma x^2}{\mathrm{N}}\right) - \bar{x}^2$

(3) $\sigma_x^2 = \left(\frac{\Sigma x^2}{\mathrm{N}}\right) - \left(\frac{\Sigma \mathrm{x}}{\mathrm{N}}\right)^2$

COVARIANCE :

(1) $cov(x\,, y) = \frac{1}{N}\Sigma(x - \bar{x})(y - \bar{y})$

(2) $cov(x\,, y) = \left(\frac{\Sigma \mathrm{xy}}{\mathrm{N}}\right) - \bar{x}\bar{y}$

(3) $cov(x\,, y) = \left(\frac{\Sigma \mathrm{xy}}{\mathrm{N}}\right) - \left(\frac{\Sigma \mathrm{x}}{\mathrm{N}}\right)\left(\frac{\Sigma \mathrm{y}}{\mathrm{N}}\right)$

KARL PEARSON'S COEFFICIENT OF CORRELATION:

Karl Pearson suggested the following coefficient of correlation to measure correlation between x and y. It is denoted by r

(1) $r = \frac{cov(x\,,y)}{\sigma_x.\sigma_y}$ (2) $r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$

(3) $r = \frac{\left(\frac{\Sigma xy}{\mathrm{N}}\right) - \left(\frac{\Sigma x}{\mathrm{N}}\right)\left(\frac{\Sigma y}{\mathrm{N}}\right)}{\sqrt{\left(\frac{\Sigma x^2}{\mathrm{N}}\right) - \left(\frac{\Sigma x}{\mathrm{N}}\right)^2}\sqrt{\left(\frac{\Sigma y^2}{N}\right) - \left(\frac{\Sigma y}{\mathrm{N}}\right)^2}}$ (4) $r = \frac{\mathrm{N}\Sigma xy - \Sigma x\Sigma y}{\sqrt{N\Sigma x^2 - (\Sigma x)^2}\sqrt{\mathrm{N}\Sigma \mathrm{y}^2 - (\Sigma \mathrm{y})^2}}$

(5) $r = \frac{\Sigma xy - N\bar{x}\bar{y}}{\sqrt{\Sigma x^2 - N\bar{x}^2}\sqrt{\Sigma \mathrm{y}^2 - N\bar{y}^2}}$

Theorem:

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

NOTE: The above theorem can also be stated as If $x = au + b \ \& \ y = cv + d$ 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

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

Theorem:

If r is coefficient of correlation than $-1 \leq r \leq 1$

EX1. Let $r_{xy} = 0.4, \text{Cov.}(x, y) = 1.6, {\sigma_y}^2 = 25$. Find σ_x

Solution: We have $r = \frac{\text{Cov.}(x,y)}{\sigma_x\sigma_y}$

But $r = 0.4, \text{Cov.}(x, y) = 1.6, \quad \sigma_y = 5$

$\therefore 0.4 = \frac{1.6}{5\sigma_x}$

$\therefore \sigma_x = \frac{1.6}{5\times 0.4} = 0.8$

EX2. 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.

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$

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

EX3. Find the coefficient of correlation between x and y for the following data.

x : 62 64 65 69 70 71 72 74

y : 126 125 139 145 165 152 180 208

x	y	xy	x^2	y^2

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

EX4. Calculate the coefficient of correlation from the following data.

$N = 10,\ \sum x = 136,\ \sum y = 243,\ \sum x^2 = 2278,\quad \sum y^2 = 6129,\ \sum xy = 3476.$

Where x, y denote the actual values.

EX5. 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.

HINT: $cov(x, y) = \frac{1}{N}\Sigma(x-\bar{x})(y-\bar{y})$

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

EX6. Calculate the correlation coefficient between x and y from the following data

$N = 10, \sum x = 140, \sum y = 150, \sum (x-10)^2 = 180,$
$\sum (y-15)^2 = 215, \sum (x-10)(y-15) = 60$

Solution: $\sum (x-10)^2 = 180$

$\sum (x^2 - 20x + 100) = 180$

$\sum x^2 - 20\sum x + 100\sum 1 = 180$

$\sum x^2 - 20(140) + 100(10) = 180$

$\sum x^2 = 1980$

$\sum (y-15)^2 = 215$

$\sum (y^2 - 30y + 225) = 215$

$\sum y^2 - 30\sum y + 225\sum 1 = 215$

$\sum y^2 - 30(150) + 225(10) = 215$

$\sum y^2 = 2465$

$\sum (x-10)(y-15) = 60$

$\sum (xy - 15x - 10y + 150) = 60$

$\sum xy - 15\sum x - 10\sum y + 150\sum 1 = 60$

$\sum xy - 15(140) - 10(150) + 150(10) = 60$

$\sum xy = 2160$

$$\therefore r = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{N(\sum x^2) - (\sum x)^2}\sqrt{N(\sum y^2) - (\sum y)^2}}$$

$$= \frac{10(2160) - (140)(150)}{\sqrt{10(1980) - (140)^2}\sqrt{10(2465) - (150)^2}}$$

$$= 0.9149$$

EX7. A computer while calculating the correlation coefficient between two variables x and y obtained the following constants $N = 25, \sum x = 125, \sum y = 100, \sum x^2 = 650, \sum y^2 = 460, \sum xy = 508.$

It was later discovered that it had recorded two pairs as

X	6	8
Y	14	6

while the correct values were

X	8	6
Y	12	8

Calculate correct correlation of coefficient.

Solution: We have $r = \frac{\sum xy - N\bar{x}\bar{y}}{\sqrt{\sum x^2 - N\bar{x}^2}\sqrt{\sum y^2 - N\bar{y}^2}}$

$$\text{Correct } \sum x = \text{Incorrect } \sum x - (\text{Sum of incorrect } x) + (\text{Sum of correct } x)$$
$$= 125 - (6 + 8) + (8 + 6) = 125$$
$$\therefore \text{Correct } \bar{x} = 125/25 = 5$$
$$\text{Correct } \sum y = 100 - (14 + 6) + (12 + 8) = 100$$
$$\therefore \text{Correct } \bar{y} = 100/25 = 4$$
$$\text{Correct } \sum x^2 = 650 - (6^2 + 8^2) + (8^2 + 6^2) = 650$$
$$\text{Correct } \sum y^2 = 460 - (14^2 + 6^2) + (12^2 + 8^2) = 436$$
$$\text{Correct } \sum xy = 508 - (6 \times 14 + 8 \times 6) +$$
$$(8 \times 12 + 6 \times 8) = 520$$
$$\therefore r = \frac{520 - 25 \times 5 \times 4}{\sqrt{650 - 25(5)^2}\sqrt{436 - 25(4)^2}} = \frac{20}{\sqrt{25}\sqrt{36}} = \frac{20}{5 \times 6} = \frac{2}{3}$$

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.

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

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.

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

X : 12 17 22 27 32.

Y : 113 119 117 115 121 .

Solution:

X	Rank x_i	Y	Rank y_i	$d_i = x_i - y_i$	$d_i^2 = (x_i - y_i)^2$
12	1	113	1	0	0
17	2	119	4	-2	4
22	3	117	3	0	0
27	4	115	2	2	4
32	5	121	5	0	0
$N = 5$				$\sum d_i = 0$	$\Sigma d_i^2 = 8$

$$R = 1 - \frac{6\Sigma d_i^2}{(n^3 - n)} = 1 - \frac{6 \times 8}{120} = \frac{72}{120} = 0.6$$

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,

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

Ex.9 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, $\Sigma d_i^2 = 13.50,\ m_1 = 2,\ m_2 = 3,\ N = 6$

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

10. 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.

11. 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 ranks 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

Solution: Since $R = 1 - \frac{6\sum {d_i}^2}{n^3 - n}$ and $r = 0.5, n = 10$

$0.5 = 1 - \frac{6\sum {d_i}^2}{1000-10} \qquad \therefore \sum {d_i}^2 = \frac{495}{6}$

$\therefore$ Corrected $\sum {d_i}^2$

$= \text{Incorrect } \sum {d_i}^2 - (\text{Incorrect rank diff.})^2 + (\text{Correct rank diff.})^2$

$= \frac{495}{6} - 3^2 + 7^2 = \frac{735}{6}$

$\therefore$ Corrects $R = 1 - \frac{6\times(735/6)}{990} = 1 - \frac{735}{990} = 0.26$