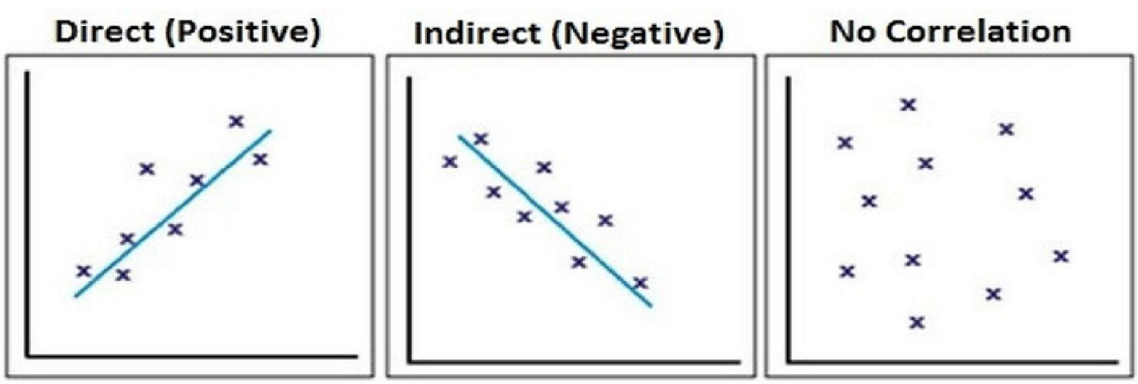
Correlation

A correlation gives the relationship between two variables say x and y. Usually, x is considered to be the independent variable also called explanatory variable and y is the dependent variable or response variable. So it is a statistical tool which measures the relationship with each other. i.e., the degree to which the variables are associated with each other such that the change in one is accompanied by the change in the other.

There are three types of correlation (i) Positive and negative correlation (ii) Simple and partial correlation (iii) Linear and non-linear correlation.

Positive and Negative correlation: Depends on the direction of change between the variables. Positive correlation is one when both the variables move in the same direction, that is when one variable increases the other on an average also increases and if one variable decreases, the other also decreases. On the other hand correlation is negative when both the variables move in the opposite direction. This shows that as one variable increases the other variable decreases and vice-versa.



Simple, Partial and multiple correlations based on the number of variables.

Simple correlation: Only when two variables are considered.

Partial correlation: When three or more variables are studied.

Multiple correlation: When three variables are studied simultaneously.

Example: To study the relationship between yield of corn per acre and the amount of organic or inorganic fertilizers used along with the rainfall, then it is a problem of multiple correlation.

Linear and non-linear correlation:

If the change in one variable bears a change in the other variable and if that is a constant ratio, then it is linear correlation Whereas the amount of change in one variable does not bear a constant ratio in the other, is called non-linear correlation.

Correlation between x and y is given by

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

$$Cov(x,y) = E((X-\mu_x)(Y-\mu_y)) \qquad where \quad \mu_x = E(X),\ \mu_y = E(Y)$$
$$= E(XY) - E(X).E(Y)$$

When X and Y are independent Cov(X,Y) = 0. If Cov(X,Y) = 0. It is not necessary that X and Y are independent.

The other formula for finding the correlation if the data X and Y are given

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

If the relationship between two variables X and Y is to be ascertained, then the following formula is used

$$r = \frac{\Sigma\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)}{\sqrt{\Sigma\left(X-\overline{X}\right)^2 \Sigma\left(Y-\overline{Y}\right)^2}} \qquad \overline{X} = mean\ of\ x \qquad and \qquad \overline{y} = mean\ of\ y$$

Properties of coefficient of correlation:

(i) r lies between ± 1

if r = 1 perfect positive correlation

if r = -1 perfect negative correlation

if r = 0 no correlation

(ii) The coefficient of correlation is independent of the origin and scale.

(iii) Coefficient of correlation is the geometric mean of regression coefficient that is $r = \sqrt{b_{xy} \times b_{yx}}$

1. Calculate the Karlperson's coefficient of correlation between the age and weight of the children.

Age (years)	1	2	3	4	5
Weight(Kgs)	3	4	6	7	12

Solution:

x	y	x^2	y^2	xy
1	3	1	9	3
2	4	4	16	8
3	6	9	36	18
4	7	16	49	28
5	12	25	144	60
$\Sigma x = 15$	$\Sigma y = 32$	$\Sigma x^2 = 55$	$\Sigma y^2 = 254$	$\Sigma xy = 117$

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

$$= \frac{5\times 117 - 15\times 32}{\sqrt{\left(5\times 55 - 15^2\right)\left(5\times 254 - 32^2\right)}} = \frac{585 - 480}{\sqrt{(275 - 225)(1270 - 1024)}}$$

$$= \frac{105}{\sqrt{50\times 246}} = 0.9467$$

2. Calculate the coefficient of correlation between death and birth rate for the following data:

Birth rate (x)	24	26	32	33	35	30
Death rate (y)	15	20	22	24	27	24

Solution:

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	X^2	Y^2	XY
24	15	-6	-7	36	49	42
26	20	-4	-2	16	4	8
32	22	2	0	4	0	0
33	24	3	2	9	4	6
35	27	5	5	25	25	25
30	24	0	2	0	4	0
$\Sigma x = 180$	$\Sigma y = 132$	**0**	**0**	**90**	**86**	**81**
$\bar{x} = 30$	$\bar{y} = 22$					

When $\bar{x}$ and $\bar{y}$ are not in fractions then we can use this formula

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2}} \qquad \bar{X} = mean\ of\ x \qquad and \qquad \bar{y} = mean\ of\ y$$

$$r = \frac{81}{\sqrt{90 \times 86}} = \frac{81}{\sqrt{7740}} = 0.92$$

When $\bar{x}$ and $\bar{y}$ are in fractions then the following formula is used

$$r = \frac{N\sum dx \times dy - \sum dx \times \sum dy}{\sqrt{\left(N\sum dx^2 - \left(\sum dx\right)^2\right)\left(N\sum dy^2 - \left(\sum dy\right)^2\right)}}$$

where $dx = x - A_x$, A_x is the assumed mean, $dy = y - A_y$, A_y is the assumed mean of y.

3. Calculate the coefficient of correlation between x and y series using Karl Pearson method.

x	14	12	14	16	16	17	16	15
y	13	11	10	15	15	9	14	17

Solution:

$A_x = 15$, $A_y = 14$

x	y	$dx = x - A_x = x - 15$	$dy = y - A_y = y - 14$	dx^2	dy^2	dx dy
14	13	-1	-1	1	1	1
12	11	-3	-3	9	9	9
14	10	-1	-4	1	16	4
16	15	1	1	1	1	1
16	15	1	1	1	1	1
17	9	2	-5	4	25	-10
16	14	1	0	1	0	0
15	17	0	3	0	9	0
		0	-8	18	62	6

$$r = \frac{N\sum dx \times dy - \sum dx \times \sum dy}{\sqrt{\left(N\sum dx^2 - \left(\sum dx\right)^2\right)\left(N\sum dy^2 - \left(\sum dy\right)^2\right)}}$$

$$= \frac{8 \times 6 - 0 \times (-8)}{\sqrt{(8 \times 18 - 0)(8 \times 62 - (-8)^2)}} = \frac{48}{\sqrt{144 \times (496 - 64)}}$$

$$= \frac{48}{\sqrt{62208}} = 0.192$$

Change of scale and origin:

Adding or subtracting or multiplying or dividing each term by certain constant, there is no effect on coefficient of correlation.

4. Compute the coefficient of correlation by Karl Pearson method for the following data:

x	1800	1900	2000	2100	2200	2300	2400	2500	2600
y	5	5	6	9	7	8	6	8	9

Solution: **A_x=2200, A_y=6**

x	y	dx	dx/100	dy	dx^2	dy^2	dx dy
1800	5	-400	-4	-1	16	1	4
1900	5	-300	-3	-1	9	1	3
2000	6	-200	-2	0	4	0	0
2100	9	-100	-1	3	1	9	-3
2200	7	0	0	1	0	1	0
2300	8	100	1	2	1	4	2
2400	6	200	2	0	4	0	0
2500	8	300	3	2	9	4	6
2600	9	400	4	3	16	9	12
			0	**9**	**60**	**29**	**24**

$$r = \frac{N\sum dx \times dy - \sum dx \times \sum dy}{\sqrt{\left(N\sum dx^2 - \left(\sum dx\right)^2\right)\left(N\sum dy^2 - \left(\sum dy\right)^2\right)}}$$

$$= \frac{9 \times 24 - 0 \times 9}{\sqrt{(9 \times 60)\left(9 \times 29 - (9)^2\right)}} = \frac{216}{\sqrt{97200}} = 0.69$$