Correlation (r)

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

The relationship between two variables such that a change in one variable gives in a positive or negative change in the other variable is known as correlation.

If the two variables deviate in the same direction, correlation is said to be direct or positive. That is if the increase (or) decrease in one variable gives a corresponding increase (or) decrease in the other Variable.

If the two variables deviate in the opposite direction, Correlation is said to be inverse (or) negative. That is if a increase (or) decrease in one variable gives a corresponding decrease (or) increase in the other Variable.

Example for Positive correlation is Height and weight of a group of persons.

Example for negative correlation is The Volume and pressure of a perfect gas.

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

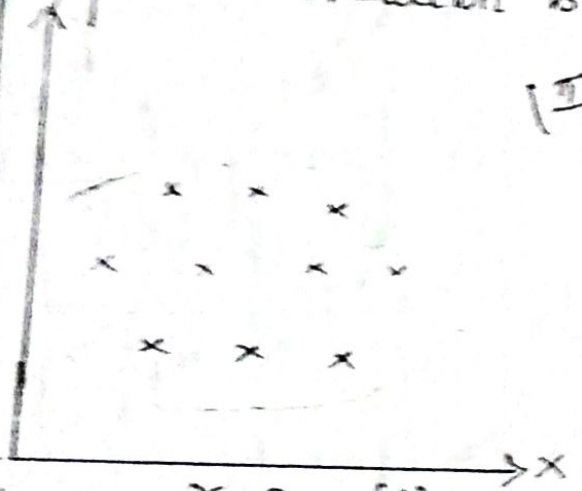
Scatter or Dot Diagram.

The values of the variables X and Y of a bivariate distribution can be plotted along the x axis and y axis. The diagram of dots so obtained is called a Scatter Diagram (or) dot diagram.

We can get a fairly good idea, whether the variables or correlated or not.

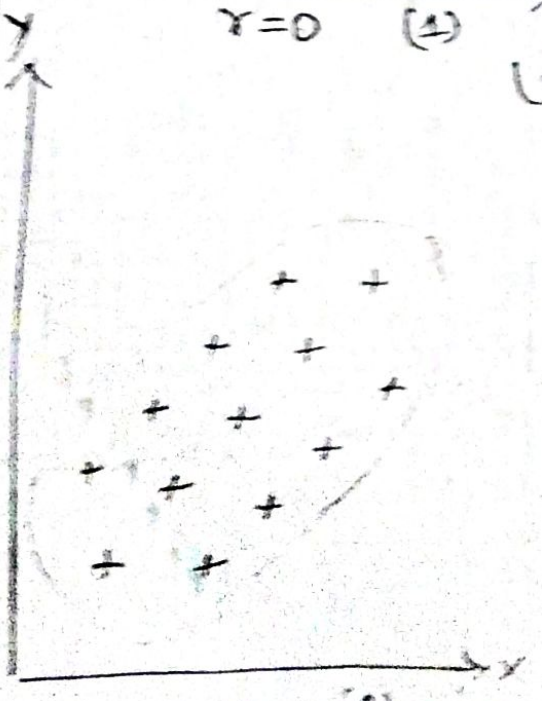
If the points are very close to other, we expect a fairly good amount of correlation between the variables.

If the points are widely scattered a poor correlation is expected.



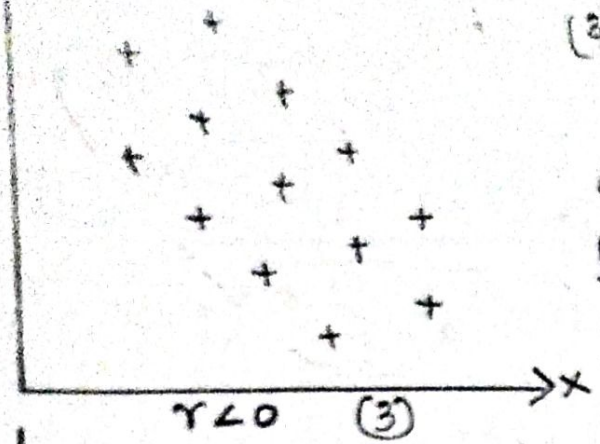
(1) This figure shows that there is no relationship between the variables.
i.e $r=0$

$r=0$ (1)



(2) This figure shows that the variables are positively correlated. Here x increases, y also increases. When x is small y is also small
i.e $r>0$

$r>0$ (2)



$r < 0$ (3)

(3) This figure shows that the variables are negatively correlated. When x is large y is small. When x is small y is large. i.e $r < 0$.

Karl Pearson's Coefficient of Correlation

Correlation coefficient between two random variables X and Y is denoted by $r(X,Y)$ or r_{xy} and defined as,

$$r_{xy} = \frac{\text{cov}(X,Y)}{\sigma_x \sigma_y}$$ (covariance)

$$\text{Cov}(X,Y) = \frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{n}\sum [x_i y_i - \bar{x} y_i - x_i \bar{y} + \bar{x}\bar{y}]$$

$$= \frac{1}{n}\sum x_i y_i - \bar{x}\left(\frac{1}{n}\sum y_i\right) - \left(\frac{1}{n}\sum x_i\right)\bar{y} + \bar{x}$$

$$= \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y} - \bar{x}\bar{y} + \bar{x}\bar{y}$$

$$\text{Cov}(X,Y) = \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}$$

$$\sigma_x^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$$

$$= \frac{1}{n}\sum (x_i - \bar{x})(x_i - \bar{x})$$

$$= \frac{1}{n}\sum [x_i^2 - x_i\bar{x} - \bar{x}x_i + (\bar{x})^2]$$

$$= \frac{1}{n}\sum x_i^2 - \left(\frac{1}{n}\sum x_i\right)\bar{x} - \bar{x}\left(\frac{1}{n}\sum x_i\right) + (\bar{x}$$

$$= \frac{1}{n}\sum x_i^2 - (\bar{x})(\bar{x}) - (\bar{x})(\bar{x}) + (\bar{x})^2$$

$$= \frac{1}{n}\sum x_i^2 - (\bar{x})^2 - (\bar{x})^2 + (\bar{x})^2$$

$$\sigma_x^2 = \frac{1}{n}\sum x_i^2 - (\bar{x})^2$$

Similarly $\sigma_y^2 = \frac{1}{n}\Sigma y_i^2 - (\bar{y})^2$

$$\left(\therefore r_{xy} = \frac{\frac{1}{n}\Sigma x_i y_i - \bar{x}\bar{y}}{\sqrt{\frac{1}{n}\Sigma x_i^2 - (\bar{x})^2}\sqrt{\frac{1}{n}\Sigma y_i^2 - (\bar{y})^2}}\right)$$

Correlation coefficient cannot exceed unity. Correlation Co-eff limit It always lies between -1 and $+1$.

If $r = +1$ Correlation is perfect and positive

If $r = -1$ Correlation is perfect and negative

If the variables are independent $r = 0$; but the converse is not true. (not relation)

Theorem:

Correlation Coefficient is independent of change of origin and scale.

Proof:

Let $U = \frac{X-a}{h}$; $V = \frac{Y-b}{k}$

a, b, k, h are constants. We have to prove that $r_{xy} = r_{uv}$.

$$U = \frac{X-a}{h}$$

$$uh = x - a$$

$$x = a + uh$$

$E(x) = a + h.E(u)$ $\quad x - E(x) = a + uh - a - h\,E(u)$

$x - E(x) = h[u - E(u)]$ $\quad x - E(x) = h[u - E(u)]$

$$V = \frac{y-b}{k}$$

$$Vk = y - b$$

$$y = b + k \cdot v$$

$$E(y) = b + k \cdot E(v)$$

$$y - E(y) = k[v - E(v)]$$

$$Cov(x,y) = E\left[\{x - E(x)\}\{y - E(y)\}\right]$$

$$= E\left[h\{u - E(u)\}\, k\{v - E(v)\}\right]$$

$$= E\left[hk\{u - E(u)\}\{v - E(v)\}\right]$$

$$= hk \cdot E\left[\{u - E(u)\}\{v - E(v)\}\right]$$

$$= hk \cdot cov(u,v).$$

$$\sigma_x^2 = E\left[x - E(x)\right]^2$$

$$= E\left[h\{u - E(u)\}\right]^2$$

$$= E\left[h^2\{u - E(u)\}^2\right]$$

$$= h^2 \cdot E\left[u - E(u)\right]^2$$

$$\sigma_x^2 = h^2 \cdot \sigma_u^2$$

$$\sigma_x = h \cdot \sigma_u$$

$$\sigma_y^2 = E\left[y - E(y)\right]^2$$

$$= E\left[k\{v - E(v)\}\right]^2$$

$$= E\left[k^2\{v - E(v)\}^2\right]$$

$$= k^2 E\left[v - E(v)\right]^2$$

$$= k^2 \cdot \sigma_v^2$$

$$\sigma_y = k \cdot \sigma_v$$

$$\sigma y^2 = E\left[Y - E(y)\right]^2$$

$$= k$$

Correlation Coefficient of X and Y is,

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

$$= \frac{hk.\ Cov(u,v)}{h\sigma_u \cdot k\sigma_v}$$

$$= \frac{Cov(u,v)}{\sigma_u \sigma_v}$$

$$= r_{uv}.$$

Hence $r_{xy} = r_{uv}$)

Theorem:

(Two independent Variables are uncorrelated

Proof (4" expectation)

$$(Cov(X,Y) = E[\{X - E(X)\}\{Y - E(Y)\}])$$

$$= E[XY - X.E(Y) - E(X).Y + E(X)E(Y)]$$

$$= E(XY) - E(X).E(Y) - E(X).E(Y) + E(X).E(Y)$$

$$= E(XY) - E(X).E(Y)$$

If X and Y are independent $E(XY) = E(X)E(Y)$

$\therefore Cov(X,Y) = E(X).E(Y) - E(X).E(Y) = 0.$

Hence $r_{xy} = \frac{Cov(X,Y)}{\sigma_x \sigma_y} = 0.$

i.e The two independent Variables are uncorrelated.

2M

Note: Two independent Variables are uncorrelated, but the converse is not true, i.e Two uncorrelated variables may not be independent.

Probable Error of Correlation Coefficient.

If 'r' is the correlation coefficient then the standard error $S.E = \frac{1-r^2}{\sqrt{n}}$

Probable error (P.E) $= 0.6745 \times S.E$

$= 0.6745 \times \left\{\frac{1-r^2}{\sqrt{n}}\right\}$

Probable error is helpful in testing the reliability of an observed correlation coefficient.

If $r < P.E$ correlation is not at all significant.

If $r > 6$ P.E it is definitely significant.

Probable error also helps us to find the limits within which the correlation coefficient canbe expected to vary.

The limits are $r \pm P.E.$

Limits for Correlation Coefficient

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

$$r_{xy} = \frac{\frac{1}{n}\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\left\{\frac{1}{n}\Sigma(x_i-\bar{x})^2\right\}\left\{\frac{1}{n}\Sigma(y_i-\bar{y})^2\right\}}}$$

Squaring both sides,

$$r_{xy}^2 = \frac{\frac{1}{n^2}\left[\Sigma(x_i-\bar{x})(y_i-\bar{y})\right]^2}{\frac{1}{n}\Sigma(x_i-\bar{x})^2 \cdot \frac{1}{n}\Sigma(y_i-\bar{y})^2}$$

Let $a_i = (x_i-\bar{x})$ and $b_i = (y_i-\bar{y})$

$$r_{xy}^2 = \frac{(\Sigma(a_i b_i)^2}{(\Sigma a_i^2)(\Sigma b_i^2)}$$

Schwartz inequality states that, if a_i, b_i $(i=1,2\ldots n)$ are real quantities then,

$$(\Sigma a_i b_i)^2 \leq (\Sigma a_i^2)(\Sigma b_i^2)$$

the sign of equality holding iff,

$$\frac{a_1}{b_1} = \frac{a_2}{b_2} = \ldots\ldots = \frac{a_n}{b_n}$$

Using this we get,

$$r_{xy}^2 \leq \frac{(\Sigma a_i^2)(\Sigma b_i^2)}{(\Sigma a_i^2)(\Sigma b_i^2)}$$

$$\leq 1.$$

$$|r_{xy}| \leq 1$$

$$\Rightarrow -1 \leq r_{xy} \leq 1$$

i.e Correlation Coefficient cannot exceed unity

Hence, Correlation Coefficient lies in between -1 and $+1$.

Rank Correlation

(The ranks 1, 2, 3 .... n are given to the n Students for their marks in Mathematics and English.)

Let X denote the ranks for Mathematics and Y denote the ranks for English.

Let (x_i, y_i) $i = 1, 2 \ldots n$ be their ranks in Mathematics and English.) Pearsonian Coefficient of correlation between the ranks x_i's and y_i's is called the Rank Correlation.

Assume that no two students get the same rank in either Mathematics or in English.

X and Y takes the values 1, 2 .... n.

X :	1	2	3	4	5 ...	n
Y :	2	5	4	3	n	9

$$\therefore \boxed{\bar{x} = \bar{y}} = \frac{1}{n}(1+2+3+\ldots\ldots+n)$$

$$= \frac{1}{n}\left[\frac{n(n+1)}{2}\right]$$

$$\boxed{\bar{x} = \bar{y} = \frac{n+1}{2}}$$

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

$$= \frac{1}{n}\left[1^2 + 2^2 + \ldots\ldots + n^2\right] - \left(\frac{n+1}{2}\right)^2$$

$$= \frac{1}{n}\left[\frac{n(n+1)(2n+1)}{6}\right] - \frac{(n+1)^2}{4}$$

$$= \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} \Rightarrow \frac{2(2n+1) - 3(n+1)}{12}$$

$$= \frac{(n+1)}{12}\left[2(2n+1) - 3(n+1)\right]$$

$$\sigma_x^2 = \frac{(n+1)}{12}\left[4n+2-3n-3\right]$$

$$= \frac{(n+1)}{12}\left[n-1\right]$$

$$= \frac{(n^2-1)}{12}$$

$x_i \neq y_i$; $\therefore \sigma_x^2 = \sigma_y^2 = \frac{(n^2-1)}{12}$

$z = x - y$

$d_i = (x_i - y_i)$

Add and Subtract $\bar{x}$ $\quad d_i = x_i - y_i + \bar{x} - \bar{x}$

$d_i = (x_i - \bar{x} - y_i + \bar{x})$ $\quad [\bar{x} = \bar{x}]$

$= (x_i - \bar{x}) - (y_i - \bar{y})$ taking Σ and squre both side

$$\Sigma d_i^2 = \Sigma\left[(x_i - \bar{x}) - (y_i - \bar{y})\right]^2$$

$$= \Sigma\left[(x_i - \bar{x})^2 + (y_i - \bar{y})^2 - 2(x_i - \bar{x})(y_i - \bar{y})\right]$$

Dividing by 'n'

$$\frac{1}{n}\Sigma d_i^2 = \frac{1}{n}\Sigma\left[(x_i - \bar{x})^2 + (y_i - \bar{y})^2 - 2(x_i - \bar{x})(y_i - \bar{y})\right]$$

$$= \frac{1}{n}\Sigma(x_i - \bar{x})^2 + \frac{1}{n}\Sigma(y_i - \bar{y})^2 - \frac{2}{n}\Sigma(x_i - \bar{x})(y_i -$$

$$= \sigma_x^2 + \sigma_y^2 - 2.\,\mathrm{Cov}(x,y)$$

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

$$\therefore \frac{1}{n}\Sigma d_i^2 = \sigma_x^2 + \sigma_y^2 - 2r.\sigma_x\sigma_y$$

$\sigma_x = \sigma_y$ $\therefore \frac{1}{n}\Sigma d_i^2 = \sigma_x^2 + \sigma_x^2 - 2r.\sigma_x\sigma_x$

$$= 2\sigma_x^2 - 2r.\sigma_x^2$$

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

$$\frac{\Sigma d_i^2}{n \cdot 2\sigma_x^2} = 1-r$$

$$r = 1 - \left[\frac{\Sigma d_i^2}{2 \cdot n \cdot \sigma_x^2}\right]$$

$$= 1 - \left[\frac{\Sigma d_i^2}{2 \cdot n \cdot \left(\frac{n^2-1}{12}\right)}\right]$$

$$= 1 - \left[\frac{\Sigma d_i^2}{\frac{n(n^2-1)}{6}}\right]$$

$$= 1 - \left[\frac{6\Sigma d_i^2}{n(n^2-1)}\right].$$

This is Spearman's formula for Rank Correlation Coefficient.

We use the symbol 'ρ' for Rank Correlation Coefficient.

$$\therefore \boxed{\rho = 1 - \left[\frac{6\Sigma d_i^2}{n(n^2-1)}\right]}$$

Repeated Ranks.

2m [If any two or more students get the same rank, then the Spearman's formula for calculating the rank correlation coefficient breaks down.]

In this case common ranks are given to the repeated items. This common rank is the average of the ranks which these items would have assumed and the next item will get the rank next to the ranks already given.

As a result of this we add the correction factor $\frac{m(m^2-1)}{12}$ to Σd^2, where 'm' is the number of items having the same rank. This correction factor is tobe added for each repeated Value.

$$\rho = 1 - \left\{ \frac{6(\Sigma d_i^2 + \text{correction factor})}{n(n^2-1)} \right\}$$

Limits for Rank Correlation Coefficient.

(Spearman's Rank Correlation coefficient is,

$$\rho = 1 - \left\{ \frac{6\Sigma d_i^2}{n(n^2-1)} \right\}$$

[If $x_i = y_i$, then $d_i = 0$. $\therefore \rho = +1$

i.e If the ranks are equal then the maximum value of ρ is +1]

If the ranks are in the opposite direction, then $\rho = -1$.

x :	1	2	3	.....	n
y :	n	(n-1)	n-2	.....	1

$$\Sigma d^2 = (1-n)^2 + \{2-(n-1)\}^2 + \{3-(n-2)\}^2 \quad d : x - y$$
$$+ \ldots\ldots + (n-1)^2$$
$$= (1-n)^2 + (3-n)^2 + (5-n)^2 + \ldots\ldots + (n-1)^2$$
$$= (n-1)^2 + (n-3)^2 + (n-5)^2 \ldots\ldots \text{upto } 'n' \text{ terms.}$$
$$\Sigma d^2 = \sum_{r=1}^{n} [n - (2r-1)]^2 \quad \text{sum of the natural number} = \frac{n(n+1)}{2}, \quad \text{1st number} = (n+1)^2$$
$$= \sum_{r=1}^{n} [(n+1) - 2r]^2$$
$$= \sum_{r=1}^{n} [(n+1)^2 - 4r(n+1) + 4r^2]$$
$$= n(n+1)^2 - 4\sum_{r=1}^{n} r(n+1) + 4\Sigma r^2$$
$$= n(n+1)^2 - 4\left\{\frac{n(n+1)}{2}\right\}(n+1) + 4\left[\frac{n(n+1)(2n+1)}{6}\right]$$
$$= n(n+1)^2 - 2n(n+1)^2 + \frac{2}{3}\{n(n+1)(2n+1)\}$$
$$= n(n+1)\left[(n+1) - 2(n+1) + \frac{2}{3}(2n+1)\right]$$
$$= n(n+1)\left[n+1-2n-2 + \frac{2}{3}(2n+1)\right]$$
$$= n(n+1)\left[-n-1 + \frac{2}{3}(2n+1)\right]$$
$$= n(n+1)\left[\frac{-3n-3+4n+2}{3}\right]$$
$$\Sigma d^2 = n(n+1)\left[\frac{(n-1)}{3}\right] = \frac{n(n^2-1)}{3}$$

$$\therefore \rho = 1 - \left\{\frac{6\Sigma d^2}{n(n^2-1)}\right\}$$

$$= 1 - \left\{\frac{6\, n(n^2-1)}{3\, n(n^2-1)}\right\}$$

$$= 1 - (6/3)$$

$$= 1 - 2$$

$$= -1.$$

∴ The limits for rank correlation coefficient is $-1 \leq \rho \leq 1$.)