K.T.S.P. Mandal's

Hutatma Rajguru Mahavidyalaya

Rajgurunagar -410505

F.Y.B.Sc

Semester-II (CBCS 2019 Pattern)

As per the New Syllabus

Subject - ST-121: Descriptive Statistics-II

Prof. S.R.Thorat

(Head, Department of Statistics)

Chapter 1 – Correlation

1. Correlation

Many times we are intrested to study two related variables.

for example: i) Marks & Intelligent Quotient (I.Q)
ii) Demand & Supply
iii) Supply & price
iv) Rainfall & Agricultural production

Such inter-related variables are known as correlated variables.

Correlation: The extent of linear relation between the two variables is called as correlation.

Bivariate Data:

In order to determine correlation, we require data regarding two related variables. These data are known as bivariate data.
Suppose X and Y are related variables.

for example: The income of family (X) & expenditure of family (Y).
we record the values of X & Y for each of the families under study.

X	x_1	x_2	- - . x_i	- - . x_n
Y	y_1	y_2	- - . y_i	- - y_n

Here the pair (x_i, y_i) indicate the income & expenditure value of ith family.
Thus n pairs of observation are known as Bivariate Data.

Types of correlation:

1) **Positive Correlation:**

If increase in the value of one variable is associated with increase in the value of another variable or decrease in the value of one variable associated with decrease in the value of another variable, the correlation between such variables is known as positive correlation.

In positive type of correlation the changes takes place in both the variables are in same direction.

for example:-

i) Marks (x) & I.Q (Y)

ii) Electricity consumption unit (x) & Electricity bill (Y)

X	↑	↓
Y	↑	↓

ii) **Negative Correlation:**

If increase in the value of one variable is associated with decrease in the value of another variable and vice-versa. The type of correlation between such variables is negative correlation.

In negative type of correlation the changes takes place in both the variables are in opposite direction.

X	↑	↓
Y	↓	↑

Ex: supply (x) and price (Y) of commodity.

iii) No correlation:

If increase or decrease in the value of one variable is not associated with other variable, the correlation between such variables is known as no correlation.

Ex: Height of student (x) & examination score (y)

Their are various measures of correlation.

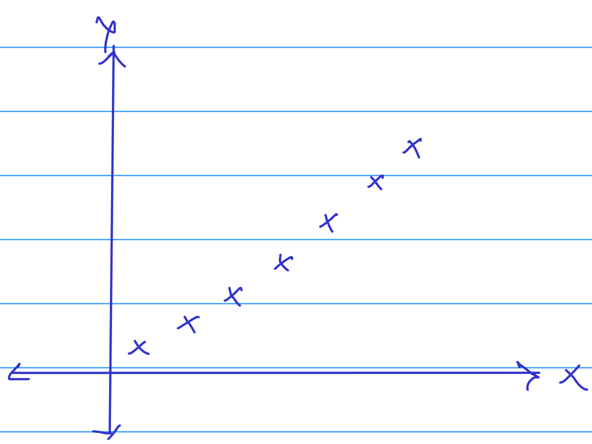
i) Scatter Diagram
ii) correlation coefficient
iii) Rank correlation coefficient.

(I) Scatter Diagram

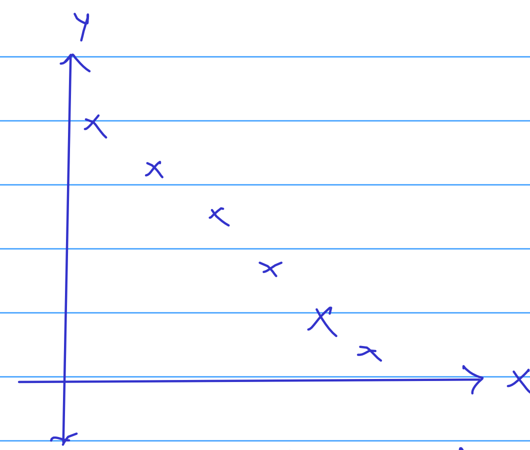
In order to visualise the correlation between two variables, the scatter diagram is used.

Let $\{(x_i, y_i) ; i = 1, 2 - - n\}$ be n pair of observations on two variables (x, y).
If these n pairs are plotted on graph paper, taking one of the variable on x-axis & other on y-axis. the diagram so obtained is known as scatter diagram.

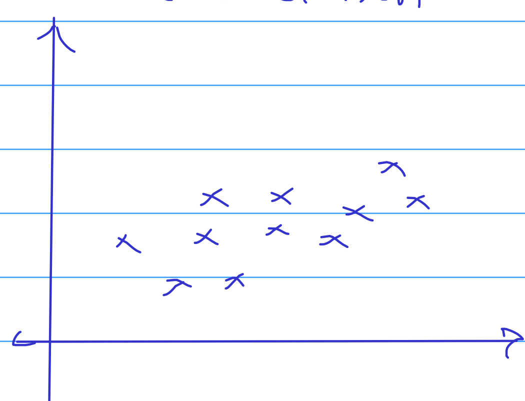
From the scatter diagram we get the general idea regarding existence of correlation and also the type of correlation.



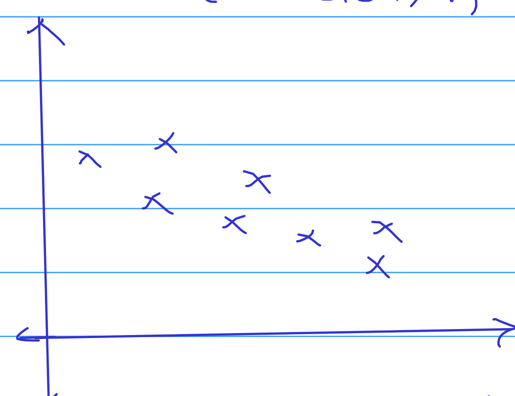
perfect positive correlation



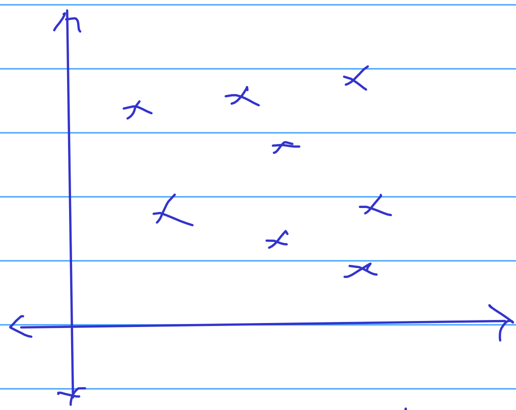
perfect Negative correlation



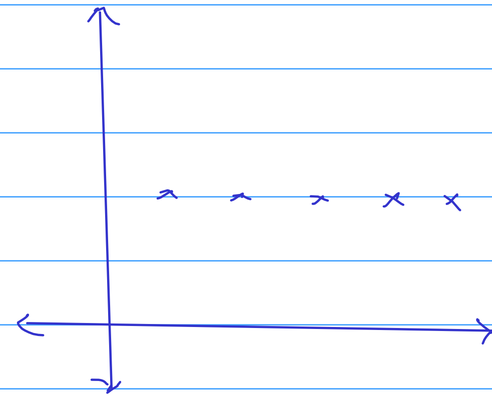
positive correlation



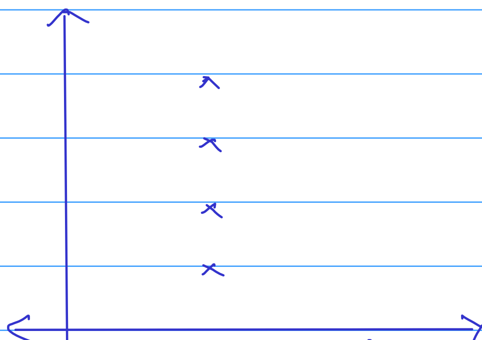
Negative correlation



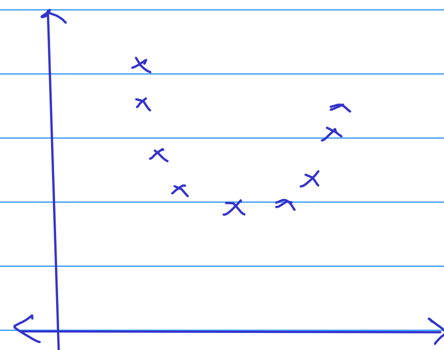
No correlation



No correlation



No correlation



Non-linear correlation

Merits of scatter Diagram:

i) It is simplest method of studing correlation
ii) It is easy to understand
iii) It is not affected due to extreme observations.

Demerits of scatter Diagram:

i) It does not give a numerical value
ii) It cannot be applied for qualitative data.

Covariance :

If $\{(x_i, y_i) : i = 1, 2, -- n\}$ be a bivariate data on (x, y), then covariance between x and y is given by,

$$\boxed{cov(x,y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}$$

Covariance measures the joint variation between the two variables.

The above formula of covariance can be simplified for computational purpose as follows.

$$cov(x,y) = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$$

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

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

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

$$= \frac{1}{n} \sum x_i y_i - 2\bar{x}\bar{y} + \bar{x}\bar{y}$$

$$\boxed{cov(x,y) = \frac{1}{n} \sum x_i y_i - \bar{x}\bar{y}}$$

Remarks:

i) $\operatorname{cov}(x, y) = \operatorname{cov}(y, x)$

ii) $\operatorname{cov}(x, x) = \operatorname{Var}(x)$

$$\left[\begin{aligned} \operatorname{cov}(x,x) &= \frac{1}{n}\sum(x-\bar{x})(x-\bar{x}) \\ &= \frac{1}{n}\sum(x-\bar{x})^2 \end{aligned}\right]$$

iii) $\operatorname{cov}(x, \text{constant}) = 0$

Properties of Covariance:

(i) $\operatorname{Cov}(x, x) = \operatorname{Var}(x)$

proof: By definition

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

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

$$\operatorname{cov}(x, x) = \operatorname{var}(x)$$

(2) Effect of change of origin:

$$\text{cov}(X-a, Y-b) = \text{cov}(X, Y)$$

where a & b are constants

Proof: Let $U = X - a \Rightarrow \bar{u} = \bar{x} - a$
$V = Y - b \Rightarrow \bar{v} = \bar{y} - b$ } (1)

Consider,

$$\text{cov}(U,V) = \frac{1}{n}\sum (u_i - \bar{u})(v_i - \bar{v})$$

$$= \frac{1}{n}\sum [x_i - a - (\bar{x} - a)][y_i - b - (\bar{y} - b)] \quad \text{from (1)}$$

$$= \frac{1}{n}\sum (x_i - \cancel{a} - \bar{x} + \cancel{a})(y_i - \cancel{b} - \bar{y} + \cancel{b})$$

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

$$\text{cov}(U,V) = \text{cov}(X,Y)$$

i.e. $\text{cov}(X-a, Y-b) = \text{cov}(X,Y)$

Remark: Similarly we can show that

$$\text{cov}(X+a, Y+b) = \text{cov}(X,Y)$$

3 Effect of change of origin & scale on Covariance

$$\operatorname{cov}\left(\frac{x-a}{h}, \frac{y-b}{k}\right) = \frac{1}{hk}\operatorname{cov}(x,y)$$

where, a, b, h & k are constants.

Proof:

Let $u = \dfrac{x-a}{h}$

$\Rightarrow u_i = \dfrac{x_i - a}{h}$ & $\bar{u} = \dfrac{\bar{x} - a}{h}$ —— (1)

$v = \dfrac{y-b}{k}$

$\Rightarrow v_i = \dfrac{y_i - b}{k}$ & $\bar{v} = \dfrac{\bar{y} - b}{k}$ —— (2)

$$\operatorname{cov}(u,v) = \frac{1}{n}\sum (u_i - \bar{u})(v_i - \bar{v})$$

$$= \frac{1}{n}\sum \left[\frac{x_i - a}{h} - \left(\frac{\bar{x}-a}{h}\right)\right]\left[\frac{y_i - b}{k} - \left(\frac{\bar{y}-b}{k}\right)\right]$$

$$= \frac{1}{n}\sum \left[\frac{x_i - \cancel{a} - \bar{x} + \cancel{a}}{h}\right]\cdot\left[\frac{y_i - \cancel{b} - \bar{y} + \cancel{b}}{k}\right]$$

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

$$= \frac{1}{hk}\,\frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})$$

$$\operatorname{cov}(u,v) = \frac{1}{hk}\operatorname{cov}(x,y)$$

i.e. $$\operatorname{cov}\left(\frac{x-a}{h}, \frac{y-b}{k}\right) = \frac{1}{hk}\operatorname{cov}(x,y)$$

Remark:

$$\operatorname{cov}(ax+b, cy+d) = ac \operatorname{cov}(x,y)$$

(4) If x, y, z are three variables then,

$$\operatorname{cov}(x+y, z) = \operatorname{cov}(x,z) + \operatorname{cov}(y,z)$$

proof: let $u = x+y$

$$\Rightarrow u_i = x_i + y_i \text{ \& } \bar{u} = \bar{x} + \bar{y} \quad \text{--- (1)}$$

consider,

$$\operatorname{cov}(x+y, z)$$

$$= \operatorname{cov}(u, z)$$

$$= \frac{1}{n}\sum (u_i - \bar{u})(z_i - \bar{z})$$

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

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

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

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

$$\operatorname{cov}(x+y, z) = \operatorname{cov}(x, z) + \operatorname{cov}(y, z)$$

---

ex ① Let $\operatorname{cov}(x, y) = 20$

then compute

i) $\operatorname{cov}(x+5, y-2)$

ii) $\operatorname{cov}(x/3, y/5)$

iii) $\operatorname{cov}\left(\frac{x-5}{7}, \frac{y+2}{3}\right)$

iv) $\operatorname{cov}(2x, 3-y)$

v) $\operatorname{cov}(-x+2, 7y-3)$

Solution: given $\operatorname{cov}(x, y) = 20$

i) $\operatorname{cov}(x+5, y-2) = \operatorname{cov}(x, y) = 20$

ii) $\operatorname{cov}(x/3, y/5) = \frac{1}{3} \cdot \frac{1}{5} \operatorname{cov}(x, y) = \frac{20}{15}$

iii) $\operatorname{cov}\left(\frac{x-5}{7}, \frac{y+2}{3}\right) = \frac{1}{7} \cdot \frac{1}{3} \operatorname{cov}(x, y) = \frac{20}{21}$

iv) $\operatorname{cov}(2x, 3-y) = 2\times(-1)\operatorname{cov}(x,y) = -2\operatorname{cov}(x,y)$

$= -2\times 20$

$\operatorname{cov}(2x, 3-y) = -40$

v) $\operatorname{cov}(-x+2, 7y-3) = (-1)(7)\operatorname{cov}(x,y)$

$= -7\operatorname{cov}(x,y)$

$= -7\times 20$

$= -140$

Ex compute covariance for the following bivariate data. also compute i) $\operatorname{cov}(-x, y)$

ii) $\operatorname{cov}(x+2, 2-y)$

iii) $\operatorname{cov}\left(x-3, \frac{3y+7}{2}\right)$

x	y
1	2
3	4
5	7
7	9
4	3

Solution :

$\operatorname{cov}(x,y) = \frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})$

$= \frac{1}{n}\sum x_iy_i - \bar{x}\bar{y} \longrightarrow (*)$

x	y	xy
1	2	2
3	4	12
5	7	35
7	9	63
4	3	12
$\Sigma x = 20$	$\Sigma y = 25$	$\Sigma xy = 124$

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{20}{5} = 4$$

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{25}{5} = 5$$

$$\text{cov}(x, y) = \frac{1}{n} \Sigma x_i y_i - \bar{x}\,\bar{y}$$

$$= \frac{1}{5}(124) - 4 \times 5$$

$$= 24.8 - 20$$

$$\underline{\text{cov}(x, y) = 4.8}$$

i) $\text{cov}(-x, y) = (-1)(1)(4.8) = -4.8$

ii) $\text{cov}(x+2, 2-y) = (1)(-1)\,\text{cov}(x, y) = -4.8$

iii) $\text{cov}\left(x-3, \frac{3y+7}{2}\right) = (1)(3/2)\,\text{cov}(x, y)$

$= 3/2 \times 4.8 = 7.2$

Karl pearson's coefficient of correlation

Let $\{(x_i, y_i);\ i = 1, 2, \ldots n\}$ is a bivariate data set on two variables x & y then the correlation coefficient between x & y is denoted by r and is defined as

$$r = corr(x,y) = \frac{cov(x,y)}{S.D(x)\ S.D(y)}$$

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

Interpretation :

i) If $r > 0$, then their is a positive correlation between two variables.

ii) If $r < 0$, then their is a negative correlation

iii) If $r = 0$, then their is no correlation between two variable

iv) If $|r| > 0.8$ we can say that their is a high correlation

v) If $0.3 < |r| < 0.8$ then is a considerable correlation

vii) If $|r| < 0.3$ then their is a negligible correlation.

Ex Compute the correlation coefficient for the following data.

x	1	3	5	7	4
y	2	4	7	9	3

Solution :

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

where, $cov(x,y) = \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}$

$$\sigma^2_x = \frac{1}{n}\sum x_i^2 - \bar{x}^2$$

$$\sigma^2_y = \frac{1}{n}\sum y_i^2 - \bar{y}^2$$

x	y	x^2	y^2	xy
1	2	1	4	2
3	4	9	16	12
5	7	25	49	35
7	9	49	81	63
4	3	16	09	12
20	25	100	159	124

$\bar{x} = \frac{\sum x}{n} = \frac{20}{5} = 4$ | $\bar{y} = \frac{\sum y}{n} = \frac{25}{5} = 5$

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

$$= \frac{100}{5} - (4)^2$$

$$= 20 - 16$$

$$\underline{\sigma_x^2 = 4}$$

$$\underline{\underline{\sigma_x = 2}}$$

$$\sigma_y^2 = \frac{1}{n}\sum y_i^2 - \bar{y}^2$$

$$= \frac{159}{5} - (5)^2$$

$$= 31.8 - 25$$

$$\underline{\sigma_y^2 = 6.8}$$

$$\sigma_y = 2.6077$$

$$\operatorname{cov}(x,y) = \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}$$

$$= \frac{124}{5} - (4 \times 5)$$

$$= 24.8 - 20$$

$$\operatorname{cov}(x,y) = \underline{4.8}$$

$$r = \operatorname{corr}(x,y) = \frac{\operatorname{cov}(x,y)}{\sigma_x \cdot \sigma_y}$$

$$= \frac{4.8}{2 \times 2.6077}$$

$$= \frac{4.8}{5.2154}$$

$$\underline{r = 0.9204}$$

There is a high positive correlation between x & y

Properties of Correlation Coefficient :

(1) Correlation coefficient does not change in magnitude under the change of origin & scale more precisely.

$$\text{corr}\left(\frac{X-a}{h}, \frac{Y-b}{K}\right) = \begin{cases} \text{corr}(X,Y) \ ; & \text{if } h \,\&\, K \text{ have same algebraic sign} \\ -\text{corr}(X,Y) \ ; & \text{if } h \,\&\, K \text{ have opposite algebraic sign.} \end{cases}$$

Proof :

Let $U = \dfrac{X-a}{h}$ & $V = \dfrac{Y-b}{K}$

$$\sigma^2_u = \frac{1}{h^2}\sigma^2_x \qquad\qquad \sigma^2_v = \frac{1}{K^2}\sigma^2_y$$

$$\sigma_u = \frac{1}{|h|}\sigma_x \qquad\qquad \sigma_v = \frac{1}{|K|}\sigma_y \qquad (1)$$

$$\text{cov}(U,V) = \text{cov}\left(\frac{X-a}{h}, \frac{Y-b}{K}\right)$$

$$\text{cov}(U,V) = \frac{1}{hK}\text{cov}(X,Y) \qquad (2)$$

consider,

$$\operatorname{corr}(U,V) = \frac{\operatorname{cov}(U,V)}{\sigma_U \cdot \sigma_V}$$

$$= \frac{\frac{1}{hk}\operatorname{cov}(X,Y)}{\frac{1}{|h|}\sigma_x \frac{1}{|k|}\sigma_y}$$

$$= \frac{|hk|}{hk} \frac{\operatorname{cov}(X,Y)}{\sigma_x \sigma_y}$$

$$\operatorname{corr}(U,V) = \frac{|hk|}{hk} \operatorname{corr}(X,Y) \quad \longrightarrow (3)$$

$$|x| = \begin{cases} x & \text{if } x > 0 \\ -x & \text{if } x < 0 \end{cases}$$

consider

$$|hk| = \begin{cases} hk \Rightarrow & hk > 0 \\ & \text{iff } h > 0 \,\&\, k > 0 \text{ or } \\ & h < 0 \,\&\, k < 0 \\ -hk \Rightarrow & hk < 0 \\ & \text{iff } h > 0 \,\&\, k < 0 \text{ or } \\ & h < 0 \,\&\, k > 0 \end{cases}$$

$$\longrightarrow (4)$$

equation (B) becomes.

$$\mathrm{corr}(u,v) = \frac{|hk|}{hk}\,\mathrm{corr}(x,y)$$

$$= \begin{cases} \dfrac{hk}{hk}\,\mathrm{corr}(x,y); & \text{if } h\,\&\,k \text{ have same algebraic sign} \\[2ex] \dfrac{-hk}{hk}\,\mathrm{corr}(x,y); & \text{if } h\,\&\,k \text{ have opposite signs} \end{cases}$$

using (4)

$$\mathrm{corr}(u,v) = \begin{cases} \mathrm{corr}(x,y) & \text{if } h\,\&\,k \text{ have same algebraic sign} \\ -\mathrm{corr}(x,y) & \text{if } h\,\&\,k \text{ have opposite algebraic sign} \end{cases}$$

Remark: $\mathrm{corr}(-x,-y) = \mathrm{corr}(x,y)$

$\mathrm{corr}(x,-y) = -\mathrm{corr}(x,y)$

$\mathrm{corr}(x,-y) = -\mathrm{corr}(x,y)$

Ex: Suppose $\mathrm{Corr}(x,y) = 0.5$ the compute

i) $\mathrm{corr}(x+3,\ y-2)$

ii) $\mathrm{corr}\left(\frac{x-2}{3},\ \frac{y+2}{7}\right)$

iii) $\mathrm{Corr}\left(\frac{x}{7},\ \frac{-3+y}{9}\right)$

iv) $\mathrm{corr}\left(\frac{7-x}{13},\ \frac{y+2}{9}\right)$

v) $\mathrm{corr}(5x-3,\ 7y+3)$

vi) $\mathrm{corr}(-3x,\ 5+y)$

vii) $\mathrm{corr}\left(\frac{x}{3},\ -2y\right)$

Vii) $\mathrm{corr}\left(\frac{1}{3}+x,\ y-\frac{1}{3}\right)$

Solution:

i) $\mathrm{corr}(x+3,\ y-2) = \mathrm{corr}(x,y) = 0.5$

ii) $\mathrm{corr}\left(\frac{x-2}{3},\ \frac{y+2}{7}\right) = \mathrm{corr}(x,y) = 0.5$

iii) $\mathrm{corr}\left(\frac{x}{7},\ \frac{-3+y}{9}\right) = \mathrm{Corr}(x,y) = 0.5$

iv) $\mathrm{corr}\left(\frac{7-x}{13},\ \frac{y+2}{9}\right) = -\mathrm{corr}(x,y) = -0.5$

v) $\mathrm{corr}(5x-3,\ 7y+3) = \mathrm{corr}(x,y) = 0.5$

vi) $\mathrm{corr}(-3x,\ 5+y) = -\mathrm{corr}(x,y) = -0.5$

vii) $\mathrm{corr}\left(\frac{x}{3},\ -2y\right) = -\mathrm{corr}(x,y) = -0.5$

viii) $\mathrm{corr}\left(\frac{1}{3}+x,\ y-\frac{1}{3}\right) = \mathrm{corr}(x,y) = 0.5$

ii) $\text{corr}(x, x) = 1$

Proof:

$$\text{corr}(x,x) = \frac{\text{cov}(x,x)}{\sigma_x \cdot \sigma_x}$$

$$= \frac{\text{var}(x)}{\sigma^2_x}$$

$$= \frac{\text{var}(x)}{\text{var}(x)}$$

$$\text{corr}(x,x) = 1$$

iii) Correlation coefficient 'r' always lie between -1 & 1

i.e $-1 \leq r \leq 1$

Proof:

we prove this property using the cauchy schwartz inequality which state that;

if $\{(a_i, b_i) ; i = 1, 2, \ldots n\}$ are real numbers

then,

$$\left(\Sigma a_i b_i\right)^2 \leq \Sigma a_i^2 \times \Sigma b_i^2 \quad \text{--- (1)}$$

let,

$a_i = (x_i - \bar{x})$

$b_i = (y_i - \bar{y})$

$$\left(\sum (x_i-\bar{x})(y_i-\bar{y})\right)^2 \leq \sum (x_i-\bar{x})^2 \sum (y_i-\bar{y})^2$$

dividing both sides by n^2

$$\left(\frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{n}\right)^2 \leq \frac{1}{n}\sum (x_i-\bar{x})^2 \times \frac{1}{n}\sum (y_i-\bar{y})^2$$

$$\left(\text{cor}(x,y)\right)^2 \leq \text{var}(x) \cdot \text{var}(y)$$

$$\left(\text{cor}(x,y)\right)^2 \leq \sigma_x^2 \cdot \sigma_y^2$$

$$\left(\frac{\text{cor}(x,y)}{\sigma_x \sigma_y}\right)^2 \leq 1$$

$$\left(\text{corr}(x,y)\right)^2 \leq 1$$

$$\varepsilon^2 \leq 1$$

i.e. $\quad \underline{-1 \leq \varepsilon \leq 1}$

Ex (1) find the correlation coefficient between x & y given that,

$n = 25$, $\Sigma x = 75$, $\Sigma y = 100$, $\Sigma x^2 = 250$, $\Sigma y^2 = 500$ & $\Sigma xy = 325$

solution :

$\bar{x} = \frac{\Sigma x}{n}$

$= \frac{75}{25}$

$\bar{x} = 3$

$\bar{y} = \frac{\Sigma y}{n}$

$= \frac{100}{25}$

$\bar{y} = 4$

$\sigma^2_x = \frac{1}{n}\Sigma x^2 - \bar{x}^2$

$= \frac{250}{25} - (3)^2$

$= 10 - 9$

$\sigma^2_x = 1$

$\sigma_x = 1$

$\sigma^2_y = \frac{1}{n}\Sigma y^2 - \bar{y}^2$

$= \frac{500}{25} - (4)^2$

$= 20 - 16$

$\sigma^2_y = 4$

$\sigma_y = 2$

$cov(x,y) = \frac{\Sigma xy}{n} - \bar{x}\bar{y}$

$= \frac{325}{25} - (3 \times 4)$

$= 13 - 12$

$\underline{cov(x,y) = 1}$

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

$$= \frac{1}{1 \times 2}$$

$= 1/2$

$\underline{r = 0.5}$

There is a considerable positive correlation between x & y.

② find the correlation coefficient between x & y Given that

$n = 8$, $\Sigma(X - \bar{X})^2 = 36$, $\Sigma(Y - \bar{Y})^2 = 44$, $\Sigma(x - \bar{x})(y - \bar{y}) = 24$

Solution:

$var(x) = 1/n \, \Sigma(x - \bar{x})^2 = 1/8 \times 36 = 4.5$

$\sigma_x = \sqrt{var(x)} = \sqrt{4.5} = 2.1213$

$var(y) = 1/n \, \Sigma(y - \bar{y})^2 = 1/8 \times 44 = 5.5$

$\sigma_y = \sqrt{var(y)} = \sqrt{5.5} = 2.3452$

$$\operatorname{cov}(x,y) = \frac{1}{n} \sum (x-\bar{x})(y-\bar{y})$$

$$= \frac{1}{8} \times 24$$

$$\operatorname{cov}(x,y) = 3$$

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

$$= \frac{3}{2.1213 \times 2.3452}$$

$$= \frac{3}{4.9749}$$

$$r = 0.6030$$

There is a considerable positive correlation between x & y

(3) Compute the value of r & interpret it.

$n = 100$, $\bar{x} = 62$, $\bar{y} = 53$, $\sigma_x = 10$, $\sigma_y = 12$

$\sum (x-\bar{x})(y-\bar{y}) = 8000$

<u>solution</u> :

$$\mathrm{cov}(x,y) = \frac{1}{n} \sum (x-\bar{x})(y-\bar{y})$$

$$= \frac{1}{100}\ 8000$$

$$\mathrm{cov}(x,y) = 80$$

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

$$= \frac{80}{10 \times 12}$$

$$= \frac{80}{120}$$

$$= 2/3$$

$$r = 0.6667$$

There is a considerable positive correlation between x and y.

④ Compute correlation coefficient between x & y given that :

$n = 100$, $\Sigma(x-35) = 25$, $\Sigma(y-19) = 68$

$\Sigma(x-35)^2 = 167$, $\Sigma(y-19)^2 = 162$

$\Sigma(x-35)(y-19) = 130$

Solution :   Let

$u = x - 35$ | $v = y - 19$

given, $\Sigma(x-35) = 25$ | $\Sigma(y-19) = 68$

$\Rightarrow \Sigma u = 25$ | $\Sigma v = 68$

$\Sigma(x-35)^2 = 167$ | $\Sigma(y-19)^2 = 162$

$\Rightarrow \Sigma u^2 = 167$ | $\Sigma v^2 = 162$

$\Sigma(x-35)(y-19) = 130$

$\Sigma uv = 130$

$\bar{u} = \frac{\Sigma u}{n} = \frac{25}{100}$ | $\bar{v} = \frac{\Sigma v}{n} = \frac{68}{100}$

$\bar{u} = 0.25$ | $\bar{v} = 0.68$

$var(u) = \frac{\Sigma u^2}{n} - \bar{u}^2$ | $var(v) = \frac{\Sigma v^2}{n} - \bar{v}^2$

$= \frac{167}{100} - (0.25)^2$ | $= \frac{162}{100} - (0.68)^2$

$= 1.67 - 0.0625$

$var(u) = 1.6075$

$\sigma_u = \sqrt{var(u)}$

$= \sqrt{1.6075}$

$\sigma_u = 1.2678$

$= 1.62 - 0.4624$

$var(v) = 1.1576$

$\sigma_v = \sqrt{var(v)}$

$= \sqrt{1.1576}$

$\sigma_v = 1.0759$

$cov(u,v) = \frac{1}{n} \sum uv - \bar{u}\bar{v}$

$= \frac{1}{100} 130 - (0.25 \times 0.68)$

$= 1.3 - 0.17$

$cov(u,v) = 1.13$

$corr(u,v) = \frac{cov(u,v)}{\sigma_u . \sigma_v}$

$= \frac{1.13}{1.2678 \times 1.0759}$

$= \frac{1.13}{1.3640}$

$corr(u,v) = 0.8284$

by using the change of origin property

$corr(u,v) = corr(x,y) = 0.8284$

i.e $corr(x-35, y-19) = corr(x,y) = 0.8284$

Variance of a Linear combination:

suppose x and y are two variables with means $\bar{x}$ & $\bar{y}$ and variances σ_x^2 & σ_y^2 respectively.

Let $ax+by$ be a a linear Combination between the two variables x & y (where a & b are constants)

then,

$$var(ax+by) = a^2 var(x) + b^2 var(y) + 2ab\, cov(x,y)$$

alternately, $var(ax+by) = a^2 var(x) + b^2 var(y) + 2ab\, r\, \sigma_x \sigma_y$

proof : Let $\left.\begin{array}{l} u = ax+by \Rightarrow \bar{u} = a\bar{x}+b\bar{y} \\ u_i = ax_i + by_i \end{array}\right\} - (1)$

consider,

$$var(ax+by) = var(u)$$

$$= \frac{1}{n}\sum (u_i - \bar{u})^2$$

$$= \frac{1}{n}\sum \left[ax_i + by_i - (a\bar{x}+b\bar{y})\right]^2 \quad \text{by eqn } (1)$$

$[A+B]^2 = A^2+B^2+2AB$

$$= \frac{1}{n}\sum \left[ax_i + by_i - a\bar{x} - b\bar{y}\right]^2$$

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

$$= \frac{1}{n}\sum \left[a^2(x_i-\bar{x})^2 + b^2(y_i-\bar{y})^2 + 2ab(x_i-\bar{x})(y_i-\bar{y})\right]$$

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

$$+ 2ab \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$\boxed{var(ax+by) = a^2 var(x) + b^2 var(y) + 2ab\, cov(x,y)} \quad \text{--- (2)}$$

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

$$\Rightarrow cov(x,y) = r \sigma_x \sigma_y$$

put in equation (2) above

$$\boxed{var(ax+by) = a^2 var(x) + b^2 var(y) + 2ab\, r \sigma_x \sigma_y}$$

__Remark:__

i) for $a=1, b=1$

$$var(x+y) = var(x) + var(y) + 2 r \sigma_x \sigma_y$$

ii) for $a=1, b=-1$

$$var(x-y) = var(x) + var(y) - 2 r \sigma_x \sigma_y$$

iii) If x and y are uncorrelated (independent)

i.e. if $r = 0$

$$\left.\begin{aligned} Var(X+Y) &= Var(X) + Var(Y) \\ Var(X-Y) &= Var(X) + Var(Y) \end{aligned}\right\}$$

that means

$Var(X+Y) = Var(X-Y)$ iff $\underline{\underline{r = 0}}$

Ex ① Let $Var(X) = 1$, $Var(Y) = 4$, $Cov(X,Y) = 2$

compute

i) $corr(X,Y)$

ii) $Var(X+Y)$

iii) $Var(X-Y)$

iv) $Var(2X+3Y)$

v) $Var(-X-2Y)$

Solution :

i) $r = corr(X,Y) = \dfrac{Cov(X,Y)}{\sigma_x \sigma_y}$

$= \dfrac{2}{1 \times 2}$

$\underline{r = 1}$

ii) $var(x+y) = var(x) + var(y) + 2\varepsilon\sigma_x\sigma_y$

$= 1 + 4 + 2(1)1\cdot2$

$= 1+4+4$

$\underline{var(x+y) = 9}$

iii) $var(x-y) = var(x) + var(y) - 2\varepsilon\sigma_x\sigma_y$

$= 1+4-2(1)1\cdot2$

$= 1+4-4$

$\underline{var(x-y) = 1}$

$var(ax+by) = a^2\,var(x) + b^2\,var(y) + 2\varepsilon ab\sigma_x\sigma_y$

iv) $var(2x+3y)$

$= 2^2\,var(x) + 3^2\,var(y) + 2(1)\,2\cdot3\cdot1\cdot2$

$= 4(1) + 9(4) + 24$

$= 4+36+24$

$= 64$

v) $var(-x-2y)$

$= (-1)^2\,var(x) + (-2)^2\,var(y) + 2(1)(-1)(-2)\,1\cdot2$

$= 1 + 4(4) + 8$

$= 1+16+8$

$= 25$

(2) If x and y are uncorrelated variables with $\sigma^2_x = k$, $\sigma^2_y = 2$ find the value of k such that

$$Var(3x - y) = 25$$

Solution : Given x & y are uncorrelated

i.e $r = 0$

$$Var(3x - y) = 25$$

$$3^2 var(x) + (-1)^2 var(y) + 2(0)3(-1)\sigma_x \sigma_y = 25$$

$$9(k) + 1\cdot(2) + 0 = 25$$

$$9k + 2 = 25$$

$$9k = 23$$

$$k = 23/9$$

(3) Suppose x & y are variables with variance σ^2_x & σ^2_y respectively. r is the correlation coefficient between them then show that,

$$r = \frac{\sigma^2_x + \sigma^2_y - var(x-y)}{2\sigma_x \sigma_y}$$

proof : we know that,

$$var(x-y) = \sigma^2_x + \sigma^2_y - 2r\sigma_x\sigma_y$$

$$2r\sigma_x\sigma_y = \sigma^2_x + \sigma^2_y - var(x-y)$$

$$r = \frac{\sigma^2_x + \sigma^2_y - var(x-y)}{2\sigma_x\sigma_y}$$

Given that ; $n = 10$, $\bar{x} = \bar{y} = 35$

$\sigma^2 x = 2500$, $\sigma^2 y = 2209$ & $\Sigma(x-y)^2 = 13590$

Compute r.

Solution : $u = x - y$ $\quad \bar{u} = \bar{x} - \bar{y} = 35 - 35 = 0$

$$var(x-y) = Var(u) = \frac{\Sigma u^2}{n} - \bar{u}^2$$

$$= \frac{\Sigma (x-y)^2}{n} - 0$$

$$= \frac{13590}{10} - 0$$

$$Var(x-y) = 1359$$

$$var(x-y) = \sigma^2 x + \sigma^2 y - 2 r \sigma x \sigma y$$

$$1359 = 2500 + 2209 - 2(r)(50)(47)$$

$$1359 = 4709 - 4700 r$$

$$4700 r = 4709 - 1359$$

$$4700 r = 3350$$

$$r = \frac{3350}{4700} = 0.7128$$

$$r = 0.7128$$

Suppose x, y & z are uncorrelated variables with the same standard deviation σ. obtain correlation coefficient between $x+y$ & $y+z$

Solution : Let $u = x+y$ & $v = y+z$

$$\text{corr}(u,v) = \frac{\text{cov}(u,v)}{\sigma_u \cdot \sigma_v} \quad \text{———} \circledast$$

$$\text{cov}(u,v) = \text{cov}(x+y,\ y+z)$$

$$= \text{cov}(x,y) + \text{cov}(x,z) + \text{cov}(y,y) + \text{cov}(y,z)$$

$$= 0 + 0 + \text{var}(y) + 0$$

since x, y & z are uncorrelated

i.e $\text{corr}(x,y) = \text{corr}(x,z) = \text{corr}(y,z) = 0$

$\Rightarrow \text{cov}(x,y) = \text{cov}(x,z) = \text{cov}(y,z) = 0$

$$= \sigma^2_y$$

$$\underline{\text{cov}(u,v) = \sigma^2}$$

given $\sigma_x = \sigma_y = \sigma_z = \sigma$

$$\text{var}(u) = \text{var}(x+y)$$

$$= \text{var}(x) + \text{var}(y) + 2\varepsilon\sigma_x\sigma_y$$

$$= \sigma^2_x + \sigma^2_y + 2(0)\sigma_x\sigma_y$$

since x & y are uncorrelated

$$= \sigma^2 + \sigma^2$$

since $\sigma_x = \sigma_y = \sigma_z = \sigma$

$$\mathrm{var}(u) = 2\sigma^2$$

$$\mathrm{var}(v) = \mathrm{var}(y+z)$$
$$= \mathrm{var}(y) + \mathrm{var}(z) + 2\varepsilon\, \sigma_y \sigma_z$$
$$= \sigma_y^2 + \sigma_z^2 + 2(0)\, \sigma_y \sigma_z$$
$$= \sigma^2 + \sigma^2 + 0$$

$$\mathrm{var}(v) = 2\sigma^2$$

Hence,

$$\varepsilon(u,v) = \frac{\mathrm{cov}(u,v)}{\sigma_u \sigma_v}$$
$$= \frac{\sigma^2}{\sqrt{2}\sigma \times \sqrt{2}\sigma}$$
$$= \frac{\cancel{\sigma^2}}{2\cancel{\sigma^2}}$$
$$= \frac{1}{2}$$
$$\varepsilon(u,v) = 0.5$$

i.e. $\varepsilon(x+y,\ y+z) = 0.5$

Merits of correlation coefficient:

1. It determines a single value which gives the extent of linear relationship between two variables & the type of correlation.

2. It depends on all observation.

Demerits of correlation coefficient:

1. It cannot be computed for qualitative data.

2. It is affected by extreme observation

3. It measures only the linear relationship.

Spearman's rank correlation coefficient:

Ranking : ordered arrangement of units according to the merit they posses is called ranking.

Rank: The number indicating the position in ranking is known as rank.

Tie: If two or more observations have the same value then we can say that the tie is occured in ranking.
In this case we allot a common rank these observations or units.
This rank is average ranks which will be allotted if these observations are slightly different

Rank correlation coefficient:

Case (1) without ties

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

where, n: number of pairs of observation.

$d_i = x_i - y_i$ where x_i & y_i's are the ranks to x & y series respectively

(1) Six entries in a music contest were rated by two judges x & y as follows

Ranks of x	5	6	4	3	2	1
Ranks of y	6	2	1	3	4	5

Compute spearman's rank correlation coefficient between x & y

Solution :

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

Ranks of x_i	Rank of y_i	$d_i = x_i - y_i$	d_i^2
5	6	-1	1
6	2	4	16
4	1	3	9
3	3	0	0
2	4	-2	4
1	5	-4	16

$\sum d_i^2 = 46$

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

$$= 1 - \frac{6(46)}{6(36-1)} = 1 - \frac{276}{210} = 1 - 1.3142$$

$$R = -0.3142$$

(2) calculate spearman's rank correlation coefficient between the following marks given by two judges in series of eight one-act plays in a drama competition.

one act play No.	1	2	3	4	5	6	7	8
Marks by Judge A	81	72	60	33	29	11	56	42
Marks by Judge B	75	56	42	15	30	20	60	80

Solution :

one act play No.	Marks by Judge A	Marks by Judge B	x_i = Rank for series A	y_i = Rank for series B	$d_i = x_i - y_i$	d_i^2
1	81	75	1	2	-1	1
2	72	56	2	4	-2	4
3	60	42	3	5	-2	4
4	33	15	6	8	-2	4
5	29	30	7	6	1	1
6	11	20	8	7	1	1
7	56	60	4	3	1	1
8	42	80	5	1	4	16
						32

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

$$= 1 - \frac{6(32)}{8(64-1)} = 1 - \frac{192}{504}$$

$$R = 0.6190$$

Rank correlation coefficient:

case (ii) with ties

Ex 1 Obtain the rank correlation coefficient for the following data:

A	68	64	75	50	64	80	75	40	55	64
B	62	58	68	45	81	60	68	48	50	70

Solution:

A	B	X = Rank of A	Y = Rank of B	X^2	Y^2	XY
68	62	4	5	16	25	20
64	58	6	7	36	49	42
75	68	2.5	3.5	6.25	12.25	8.75
50	45	9	10	81	100	90
64	81	6	1	36	1	6
80	60	1	6	1	36	6
75	68	2.5	3.5	6.25	12.25	8.75
40	48	10	9	100	81	90
55	50	8	8	64	64	64
64	70	6	2	36	04	12
		55	55	382.5	384.5	347.5

2, 3

5, 6, 7

6

3, 4

$$\bar{x} = \frac{\Sigma x}{n} = \frac{55}{10} = 5.5$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{55}{10} = 5.5$$

$$\sigma_x^2 = \frac{\Sigma x^2}{n} - \bar{x}^2$$

$$= \frac{382.5}{10} - (5.5)^2$$

$$\sigma_x^2 = 8$$

$$\sigma_y^2 = \frac{\Sigma y^2}{n} - \bar{y}^2$$

$$= \frac{384.5}{10} - (5.5)^2$$

$$\sigma_y^2 = 8.2$$

$\sigma_x = 2.8284$ | $\sigma_y = 2.8636$

$$cov(x,y) = \frac{\Sigma xy}{n} - \bar{x}\bar{y}$$

$$= \frac{347.5}{10} - (5.5 \times 5.5)$$

$$= 34.75 - 30.25$$

$$cov(x,y) = 4.5$$

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

$$= \frac{4.5}{2.8284 \times 2.8636}$$

$$= \frac{4.5}{8.0994}$$

$$R = corr(x,y) = 0.5556$$

Note: Spearmans rank correlation coefficient with ties is simply correlation coefficient between the ranks.