Module - 3 :- Statistical Methods

Correlation :- The changes in one variable are associated (or) followed by changes in the other is called correlation.

The degree of relation b/w 2 variables x & y is measured by a single no. 'r' is called correlation coefficient.

* This used to find out the extent to which 2 variables are related to each other.

If 2 variables x & y are related in such a way that increase (or) decrease in one of them corresponds to increase or decrease in the other. Then variables are positively correlated. Also, if increase or decrease in one of them corresponds to decrease or increase in the other, the variables are said to be negatively correlated.

x	y
20	40
30	30
40	22
60	15
80	10

$x\uparrow$ $y\downarrow$

(-vely correlated)

x	y
20	10
30	15
40	22
60	30
80	40

($x\uparrow$ $y\uparrow$)

(+vely correlated)

If there is no relationship indicated b/w the variables, they are said to be

Ex:- 1) Income and expenditure (+vely correlated)
Height & weight $x\uparrow$ $y\uparrow$

2) Price & Demand (-vely correlated)
$x\uparrow$ $y\downarrow$ or $x\downarrow$ $y\uparrow$

3) Two coins being tossed simultaneously. (uncorrelated).

Note:- * The coefficient of correlation $r \in [-1, 1]$
(or) $-1 \le r \le 1$

* If $r = \pm 1$, there is perfect positive or negative correlation

* If $r = 0$, there is no linear correlation. but a non linear correlation may exist.

Karl-Pearson's Coefficient of correlation

The correlation coefficient r is measured by the following formula

$$r = \frac{\frac{1}{n}\sum (x-\bar{x})(y-\bar{y})}{\sigma_x \sigma_y} \rightarrow (a)$$

If $X = x - \bar{x}$ $\quad Y = y - \bar{y}$

$$\sigma_x = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n}} \qquad \sigma_y = \sqrt{\frac{\Sigma (y-\bar{y})^2}{n}}$$

$$\sigma_x = \sqrt{\frac{\Sigma x^2}{n}} \qquad \sigma_y = \sqrt{\frac{\Sigma Y^2}{n}}$$

$$\sigma_x = \frac{\sqrt{\Sigma x^2}}{\sqrt{n^2}} \qquad \sigma_y = \frac{\sqrt{\Sigma Y^2}}{\sqrt{n}}$$

$\sigma_x \sqrt{n} = \sqrt{\Sigma x^2} \rightarrow (1)$; $\sigma_y \sqrt{n} = \sqrt{\Sigma Y^2} \rightarrow (2)$

(1) & (2) in eqn (a)

$$(\sqrt{n})^2 \sigma_x \cdot \sigma_y = \sqrt{\Sigma x^2}\sqrt{\Sigma y^2}$$

$$n \sigma_x \sigma_y = \sqrt{\Sigma x^2}\sqrt{\Sigma y^2} \rightarrow (3)$$

Sub eq (3) in eq (2)

$$\therefore \boxed{r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}}}$$

Problems:-

1) Compute the coefficient of correlation of the following data.

x	1	2	3	4	5	6	7
y	9	8	10	12	11	13	14

Soln:-

x	y	$x = x - \bar{x}$	$y = y - \bar{y}$	xy	x^2	y^2
1	9	-3	-2	6	9	4
2	8	-2	-3	6	4	9
3	10	-1	-1	1	1	1
4	12	0	1	0	0	1
5	11	1	0	0	1	0
6	13	2	2	4	4	4
7	14	3	3	9	9	9
				$\Sigma xy = 26$	$\Sigma x^2 = 28$	$\Sigma y^2 = 28$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{28}{7} = 4 \qquad \bar{y} = \frac{\Sigma y}{n} = \frac{77}{7} = 11$$

$$\therefore r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}} = \frac{26}{\sqrt{28}\sqrt{28}} = \underline{0.9285}$$

2) Find the coefficient of correlation for the following data.

x	10	14	18	22	26	30
y	18	12	24	6	30	36

Sol^n -

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	XY	X^2	Y^2
10	18	-10	-3	30	100	9
14	12	-6	-9	54	36	81
18	24	-2	3	6	4	9
22	6	2	-15	30	4	225
26	30	6	9	54	36	81
30	36	10	15	150	100	225
				$\Sigma XY = 252$	$\Sigma X^2 = 280$	$\Sigma Y^2 = 630$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{120}{6} = 20$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{126}{6} = 21$$

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2}\sqrt{\Sigma Y^2}} = \frac{252}{\sqrt{280}\sqrt{630}} = 0.6$$

H.W.

3) Determine r for the following data

x	50	60	70	90	100
y	65	51	40	26	8

Ans) - $\boxed{-0.9}$

Regression :- Regression is an estimation of one independent variable in terms of other.

Equation of the regression lines

1) Straight line of the form $y = ax + b$ (x being the independent variable) is called the regression line of y on x.

$$y-\bar{y} = r\cdot\frac{\sigma_y}{\sigma_x}(x-\bar{x}) \quad \rightarrow (1)$$

where $r = \dfrac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}}$

Consider,

$$r\frac{\sigma_y}{\sigma_x} \Rightarrow \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}} \cdot \frac{\sqrt{\frac{\Sigma y^2}{n}}}{\sqrt{\frac{\Sigma y^2}{n}}}$$

$$= \frac{\Sigma xy}{\Sigma x^2}$$

$\therefore$ eq (1) becomes

$$\boxed{y-\bar{y} = \frac{\Sigma xy}{\Sigma x^2}(x-\bar{x})} \text{ or } y = \frac{\Sigma xy}{\Sigma x^2} x$$

* Here Slope of the line of regression of y on x } $= r\frac{\sigma_y}{\sigma_x}$
* regression coefficient

$\Rightarrow$ Straight line $x = ay+b$ (y being independent variable) is called the regression line of x on y.

$$(x-\bar{x}) = r\frac{\sigma_x}{\sigma_y}(y-\bar{y})$$

$$\Rightarrow \boxed{(x-\bar{x}) = \frac{\Sigma xy}{\Sigma y^2}(y-\bar{y})} \text{ or } x = \frac{\Sigma xy}{\Sigma y^2} y$$

* Slope of the line of regression x on y is } $= r\frac{\sigma_x}{\sigma_y}$
* regression coefficient

Hence correlation coefficient r is the geometric mean btw the 2 regression coefficients

$$r = \pm \sqrt{(\text{Coeff of } x)(\text{Coeff of } y)}$$

$$r = \sqrt{\left(r\frac{\sigma_y}{\sigma_x}\right) \times \left(r\frac{\sigma_x}{\sigma_y}\right)}$$

Note:- The point $(\bar{x}, \bar{y})$ lies on regression lines.

Problems:-

1) Calculate (i) regression equation of x on y and y on x from the following data.
(ii) estimate x when y is 20.

x	10	12	13	17	18
y	5	6	7	9	13

Sol:-

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	XY	X^2	Y^2
10	5	-4	-3	12	16	9
12	6	-2	-2	4	4	4
13	7	-1	-1	1	1	1
17	9	3	1	3	9	1
18	13	4	5	20	16	25
				$\Sigma XY = 40$	46	40

$$\bar{x} = \frac{\Sigma x}{n} = \frac{70}{5} = 14 \quad , \quad \bar{y} = \frac{\Sigma y}{n} = \frac{40}{5} = 8$$

(i)

a) Regression line of x on y

$$x = \frac{\Sigma XY}{\Sigma Y^2} \cdot Y$$

$$(x - \bar{x}) = \frac{\Sigma XY}{\Sigma Y^2}(y - \bar{y})$$

$$(x - 14) = \frac{40}{40}(y - 8)$$

$$x - 14 = y - 8$$

$$x = y - 8 + 14$$

$$\boxed{x = y + 6}$$ —

(b) Regression line of y on x

$$y = \frac{\Sigma xy}{\Sigma x^2} \cdot X$$

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

$$(y - 8) = \frac{40}{46}(x - 14)$$

$$y - 8 = 0.8695(x - 14)$$

$$= 0.8695x = 12.1739$$

$$y = 0.8695x - 12.1739 + 8$$

$$\boxed{y = 0.8695x - 4.1739}$$

(ii) put $y = 20$ in $x = y + 6$

$$x = 20 + 6$$

$$x = 26$$

2) Obtain the lines of regression & hence find the coefficient of correlation for the data

x	1	3	4	2	5	8	9	10	13	15
y	8	6	10	8	12	16	16	10	32	32

$$\bar{x} = \frac{\Sigma x}{n} = \frac{70}{10} = 7$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{150}{10} = 15.$$

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	XY	X^2	Y^2
1	8	-6	-7	42	36	49
3	6	-4	-9	36	16	81
4	10	-3	-5	15	9	25
2	8	-5	-7	35	25	49
5	12	-2	-3	6	4	9
8	16	1	1	1	1	1
9	16	2	1	2	4	1
10	10	3	-5	15	9	25
13	32	6	17	102	36	289
15	32	8	17	136	64	289
				$= 360$	204	818

Regression line x on y

$$(x - \bar{x}) = \frac{\Sigma XY}{\Sigma Y^2}(y - \bar{y})$$

$$(x - 7) = \frac{360}{818}(y - 15)$$

$$x - 7 = 0.44y - 6.6014$$

$$x = 0.4401y + 0.3985$$

Regression line y on x

$$(y - \bar{y}) = \frac{\Sigma XY}{\Sigma X^2}(x - \bar{x})$$

$$(y - 15) = \frac{360}{204}(x - 7)$$

$$y = 1.7647x - 12.3529$$

$$y = 1.7647x + 2.6470$$

Here regression coefficient x on $y = 0.4401$

" " y on $x = 1.7647$

$$r = \sqrt{(0.4401) \times (1.7647)}$$

$$r = 0.8812$$

3) In a partially destroyed record of correlation data, the following results only are available:

Regression equations are $4x - 5y + 33 = 0$

$20x - 9y = 107$.

Calculate (i) the mean values of x & y

(ii) the coefficient of correlation b/w x & y

(iii) S.D of y.

Soln:- (i) W.K.T. $(\bar{x}, \bar{y})$ is point on the regression line

$\therefore 4\bar{x} - 5\bar{y} + 33 = 0 \Rightarrow 4\bar{x} - 5\bar{y} = -33 \rightarrow (1)$

$20\bar{x} - 9\bar{y} = 107 \rightarrow (2)$

Solve eq (1) & (2)

Consider, $4\bar{x} - 5\bar{y} = -33$

$4\bar{x} = -33 + 5\bar{y}$

$\bar{x} = \frac{-33 + 5\bar{y}}{4} \rightarrow (3)$

Sub eq (3) in eq (2)

$20\bar{x} - 9\bar{y} = 107$

$\frac{5}{20}\left(\frac{-33 + 5\bar{y}}{4}\right) - 9\bar{y} = 107$

$-165 + 25\bar{y} - 9\bar{y} = 107$

$16\bar{y} = 107 + 165$

$16\bar{y} = 272$

$\bar{y} = 17$

Sub $\bar{y} = 17$ in eq (1)

$4\bar{x} - 5\bar{y} = -33$

$4\bar{x} - 5(17) = -33$

$4\bar{x} - 85 = -33$

$$4\bar{x} = -33 + 85$$

$$4\bar{x} = 52$$

$$\bar{x} = 13$$

$\therefore \bar{x} = 13, \quad \bar{y} = 17$

(ii) Regression line y on x

$$5y = 4x + 33$$

$$y = \frac{4}{5}x + \frac{33}{5}$$

Regression line x on y

$$20x = 9y + 107$$

$$x = \frac{9}{20}y + \frac{107}{20}$$

Regression coefficient of y on $x = \frac{4}{5} = 0.8$

—11— x on $y = \frac{9}{20} = 0.45$

$$r = \sqrt{(0.8)\times(0.45)} = \underline{0.6}$$

(iii)

$$\sigma_x^2 = 9 \Rightarrow \sigma_x = 3$$

Regression coefficient x on $y = r \cdot \frac{\sigma_x}{\sigma_y} = 0.45$

$$\sigma_y = \frac{r \cdot \sigma_x}{0.45} = \frac{0.6 \times 3}{0.45} = 4$$

4) If θ is the angle b/w the 2 regression lines, S.T $\tan\theta = \left(\frac{1-r^2}{r}\right)\left(\frac{\sigma_x \cdot \sigma_y}{\sigma_x^2 + \sigma_y^2}\right)$.

Explain the significance when $r = 0$ & $r = \pm 1$.

Soln:- Regression line y on x

$$(y - \bar{y}) = r \cdot \frac{\sigma_y}{\sigma_x}(x - \bar{x}) \rightarrow (1)$$

Regression line x on y

$$(x-\bar{x}) = r\frac{\sigma_x}{\sigma_y}(y-\bar{y}) \rightarrow (2)$$

Slopes of eqn (1) & (2)

$$m_1 = r\frac{\sigma_y}{\sigma_x} \qquad m_2 = \frac{\sigma_y}{r\sigma_x}$$

$$\tan\theta = \frac{m_2 - m_1}{1 + m_1 \cdot m_2}$$

$$= \frac{\frac{\sigma_y}{r\sigma_x} - \frac{r\sigma_y}{\sigma_x}}{1 + r\cdot\frac{\sigma_y}{\sigma_x}\cdot\frac{\sigma_y}{r\sigma_x}} = \frac{\frac{\sigma_y - r^2\sigma_y}{r\cdot\sigma_x}}{\frac{\sigma_x^2 + \sigma_y^2}{\sigma_x^2}}$$

$$= \frac{\left(\frac{1-r^2}{r}\right)\cdot\sigma_y}{\left(\frac{\sigma_x^2+\sigma_y^2}{\sigma_x}\right)}$$

$$\tan\theta = \left(\frac{1-r^2}{r}\right)\left(\frac{\sigma_x\cdot\sigma_y}{\sigma_x^2+\sigma_y^2}\right)$$

when $r = 0 \Rightarrow \tan\theta = \infty \Rightarrow \theta = \frac{\pi}{2}$

when the variables are linearly uncorrelated then they are $\perp^{ar}$ to each other.

when $r = \pm 1 \Rightarrow \tan\theta = 0 \Rightarrow \theta = 0$ or π

when lines of regression coincide then there is a perfect correlation b/w 2 variables.

⇒ If the tangent of the angle θ b/w the lines of regression of y on x & x on y is 0.6 and the S.D. of y is twice the S.D of x, find the coefficient of correlation b/w x & y.

Soln:- Given θ be the angle b/w the lines of regression

$\tan\theta = 0.6$ $\quad \sigma_y = 2\sigma_x$

w.k.t. $\tan\theta = \left(\frac{1-r^2}{r}\right)\left(\frac{r\cdot\sigma_y}{\sigma_x^2+\sigma_y^2}\right)$

$0.6 = \left(\frac{1-r^2}{r}\right)\left(\frac{2\sigma_x^2}{\sigma_x^2+4\sigma_x^2}\right)$

$= \left(\frac{1-r^2}{r}\right)\left(\frac{2\sigma_x^2}{5\sigma_x^2}\right)$

$0.6 = \frac{(1-r^2)}{r}\cdot\frac{2}{5}$

$3r = 2-2r^2$

$2r^2+3r-2=0$

$r = \frac{1}{2} = 0.5$ (or) $r=-2$ not possible.

$\Rightarrow r = 0.5$

H.W.

6) Find the correlation coefficient b/w x & y for the following data. Also obtain the regression lines.

x :	1	2	3	4	5	6	7	8	9	10
y :	10	12	16	28	25	36	41	49	40	50

7) In a bivariate distribution, it is found that $\sigma_x = \sigma_y$ & the acute angle b/w the lines of regression is $\tan^{-1}(3)$. Find the correlation coefficient.

Soln:- w.k.t. $\tan\theta = \left(\frac{1-r^2}{r}\right)\left(\frac{r\cdot\sigma_y}{\sigma_x^2+\sigma_y^2}\right)$

given $\theta = \tan^{-1}(3)$ $\quad \sigma_x = \sigma_y$

$\tan(\tan^{-1}(3)) = \left(\frac{1-r^2}{r}\right)\left(\frac{\sigma_x^2}{2\sigma_x^2}\right)$

$3 = \left(\frac{1-r^2}{r}\right)\left(\frac{1}{2}\right)$

$$6r = 1 - r^2$$

$$r^2 + 6r - 1 = 0$$

$$\Rightarrow r = 0.1623 \qquad r = -6.1623$$

$$\Rightarrow r = 0.\underline{1623}$$

8) With usual notation, compute $\bar{x}$, $\bar{y}$ & r from the following lines of regression.

$$2x + 3y + 1 = 0$$

$$x + 6y - 4 = 0$$

Sol:- W.K.T. $(\bar{x}, \bar{y})$ is point on the regression line

$$2\bar{x} + 3\bar{y} = -1 \quad \rightarrow (1)$$

$$\bar{x} + 6\bar{y} = 4 \quad \rightarrow (2)$$

from eq (2) $\bar{x} = 4 - 6\bar{y}$

Sub. $\bar{x}$ in eq (1)

$$2(4 - 6\bar{y}) + 3\bar{y} = -1$$

$$8 - 12\bar{y} + 3\bar{y} = -1$$

$$-9\bar{y} = -1 - 8$$

$$-9\bar{y} = -9$$

$$\bar{y} = \underline{1}$$

Sub. $\bar{y} = 1$ in eq (2)

$$\bar{x} + 6(1) = 4$$

$$\bar{x} = 4 - 6$$

$$\bar{x} = -2$$

$$\therefore \bar{x} = -2, \underline{\bar{y} = 1}$$

Regression line of y on x & x on y.

$$2x+3y+1=0$$

$$3y=-1-2x$$

$$y=-\frac{1}{3}-\frac{2}{3}x \quad (\text{or}) \quad -\frac{2}{3}x-\frac{1}{3}$$

&

$$x+6y-4=0$$

$$x=-6y+4$$

regression coefficients are $-\frac{2}{3}$ & -6

$$\therefore r=\sqrt{\left(\frac{-2}{3}\right)(-6)} \Rightarrow \pm 2$$

$r \in [-1, 1]$ Not possible.

Regression line of x on y & y on x.

$$2x+3y+1=0$$

$$2x=-1-3y$$

$$x=\frac{-3y}{2}-\frac{1}{2}$$

&

$$x+6y-4=0$$

$$6y=4-x$$

$$y=-\frac{x}{6}+\frac{4}{6}$$

Regression coefficients are $-\frac{3}{2}$ & $-\frac{1}{6}$

$$\therefore r=\sqrt{\left(-\frac{3}{2}\right)\left(-\frac{1}{6}\right)}=\sqrt{\frac{1}{4}}$$

$$r=\pm 1/2=\pm 0.5$$

since $\frac{3}{2}$ & $\frac{1}{6}$ are negative

$$r=-0.5$$

$$\therefore \bar{x}=-2, \quad \bar{y}=1 \quad r=-0.5$$

9) From the following eqn of the lines of regression compute the mean of x ($\bar{x}$), mean of y ($\bar{y}$) & correlation coefficient r.

(i) $y = x$, $4x - y - 3 = 0$ Ans - $\bar{x} = 1 = \bar{y}$, $r = 0.5$

(ii) $y = 0.516x + 33.73$,

$x = 0.512y + 32.52$

Ans $\bar{x} = 67.67$

$\bar{y} = 68.65$

$r = +0.51$

11/ Rank Correlation Coefficient

The coefficient of correlation in respect of the ranks of some two characteristics of an individual or an observation is called the rank correlation coefficient. It is denoted by ρ (rho). ρ is computed by

$$\rho = 1 - \frac{6\Sigma(x-y)^2}{n(n^2-1)} \quad \text{(or)} \quad 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

where $d^2 = x - y$

Note:- (1) If the rankings of x, y are entirely in the same order like for example,

$x: 1, 2, 3, 4, 5$; $y = 1, 2, 3, 4, 5$ then

$\Sigma d^2 = \Sigma(x-y)^2 = 0$. This will give us $\rho = +1$ and is called perfect direct correlation.

If the ranking of x and y are entirely in the opposite order like for example,

$x: 1, 2, 3, 4, 5$; $y: 5, 4, 3, 2, 1$

then $\Sigma d^2 = \Sigma(x-y)^2 = (-4)^2 + (-2)^2 + (0)^2 + (2)^2 + (4)^2$

$= 40$

$\therefore \rho = 1 - \frac{6(40)}{5(5^2-1)} = -1$ and is called perfect inverse correlation.

(2) ρ in case of repeated ranks

If two or more magnitudes are repeated,

then ... will be a tie for a particular rank. In such cases we assign the average rank to all those magnitudes and use the correction factor $\frac{m(m^2-1)}{12}$ along with Σd^2 in the formula for ρ

where m denotes no of times a magnitude repeated.

The correction factor must be added every time the 'tie' occurs for a particular rank.

$$\therefore \rho = 1 - \frac{6\left[\Sigma d^2 + \frac{m(m^2-1)}{12} + \cdots\right]}{n(n^2-1)}$$

Ex:- If there is a tie for the second rank in two magnitudes, the rank $\frac{2+3}{2} = 2.5$ is assigned to both the magnitudes, thereby wasting both 2nd & 3rd ranks ($m = 2$).

Again, suppose there is a tie for 5th rank among 3 magnitudes, we take the average of the 3 ranks in a row starting from 5.

i.e. $\frac{5+6+7}{3} = 6$ $(m = 3)$.

Problems:-

1) Ten competitors in a beauty contest are ranked by 2 judges in the following order. Compute the coefficient of rank correlation.

I	1	6	5	3	10	2	4	9	7	8
II	6	4	9	8	1	2	3	10	5	7

Soln:- we have,

$$S = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

For the given data, $n=10$

we have $\Sigma d^2 = (1-6)^2 + (6-4)^2 + (5-9)^2 + (3-8)^2 + (10-1)^2 + (2-2)^2 + (4-3)^2 + (9-10)^2 + (7-5)^2 + (8-7)^2$

i.e. $\Sigma d^2 = 25+4+16+25+81+0+1+1+4+1$

$= 158$

Hence $S = 1 - \frac{6(158)}{10(10^2-1)}$

$$\boxed{S = 0.042}$$

2) Ten competitors in music contest are ranked by 3 judges A, B, C in the following order. Use the rank correlation coefficient to decide which pair of judges have the nearest approach to common taste of music.

A	1	6	5	10	3	2	4	9	7	8
B	3	5	8	4	7	10	2	1	6	9
C	6	4	9	8	1	2	3	10	5	7

A	B	d_{AB}	d_{BC}	d_{CA}
1	3	6	4	9	25
6	5	4	1	1	4
5	8	9	9	1	16
10	4	8	36	16	4
3	7	1	16	36	4
2	10	2	64	64	0
4	2	3	4	1	1
9	1	10	64	81	1
7	6	5	1	1	4
8	9	7	1	4	1
			$\Sigma d^2_{AB} = 200$	$\Sigma d^2_{BC} = 214$	$\Sigma d^2_{CA} = 60$

we have, $$\rho = 1 - \frac{6\,\Sigma d^2}{n(n^2-1)} \quad n = 10$$

Now, $$\rho_{AB} = 1 - \frac{6(200)}{10(10^2-1)} = -0.21$$

$$\rho_{BC} = 1 - \frac{6(214)}{10(10^2-1)} = -0.297$$

$$\rho_{CA} = 1 - \frac{6(60)}{10(10^2-1)} = +0.636$$

Here ρ_{AB} & ρ_{BC} are negative which means their tastes (A & B ; B & C) are opposite.

But ρ_{CA} is positive and is nearer to 1 (perfect correlation).

Thus judes C & A have the nearest approach to common taste of music.

3) Ten students got the following percentage of marks in 2 subjects x & y. Compute their rank correlation coefficient.

Marks in x	78	36	98	25	75	82	90	62	65
Marks in y	84	51	91	60	68	62	86	58	53
								39	
								47	

Soln:-

Marks in x	Rank (x)	Marks in y	Rank (y)	$d=(x-y)$	$d^2=(x-y)^2$
78	4	84	3	1	1
36	9	51	9	0	0
98	1	91	1	0	0
25	10	60	6	4	16
75	5	68	4	1	1
82	3	62	5	-2	4
90	2	86	2	0	0
62	7	58	7	0	0
65	6	53	8	-2	4
39	8	39	10	-2	4
					$\Sigma d^2 = 30$

We have, $\rho = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$ $\quad n = 10$

$$\rho = 1 - \frac{6(30)}{10(10^2-1)} = 0.8181 \simeq 0.82$$

$\therefore \rho = 0.82$

4) Compute the rank correlation coefficient for the following ranks of ten candidates in 2 subjects.

Sub-1	4	8	7	6	5	9	10	3	2	1
Sub-2	6	7	8	1	5	10	9	2	3	4

Ans:- 0.73

5) Compute the rank correlation coefficient for the following data.

x	68	64	75	50	64	80	75	40	55	64
y	62	58	68	45	81	60	68	48	50	70

Sol:-

x	Rank(x)	y	Rank(y)	$d = x - y$	d^2
68	4	62	5	-1	1
64	6	58	7	-1	1
75	2.5	68	3.5	-1	1
50	9	45		-1	1
64	6	81	1	5	25
80	1	60	6	-5	25
75	2.5	68	3.5	-1	1
40	10	48	9	1	1
55	8	50	8	0	0
64	6	70	2	4	16

$\frac{2+3}{2} = 2.5$ (m=2)

$\frac{5+6+7}{3} = \frac{18}{3} = 6$ (m=3)

$\frac{3+4}{2} = \frac{7}{2} = 3.5$ (m=2)

$$\rho = 1 - \frac{6\left[\Sigma d^2 + \frac{m(m^2-1)}{12} + \cdots\right]}{n(n^2-1)}, \quad n=10$$

$$= 1 - \frac{6\left[72 + \frac{2(2^2-1)}{12} + \frac{3(3^2-1)}{12} + \frac{2(2^2-1)}{12}\right]}{10(10^2-1)} = 0.545$$

$\therefore \rho = 0.545$

Curve Fitting

The method of finding a specific relation $y = f(x)$ for the data to satisfy as accurately as possible and such an eqn is called the best fitting eqn or the curve of best fit.

+ Fitting of a straight line $y = ax + b$

Consider sets of n given values (x, y) for fitting the straight line $y = ax + b$ where a & b are parameters to be determined.

The residual $R = y - (ax + b)$ is the difference b/w the observed and estimated values of y.

By the method of least squares we find parameters a and b such that the sum of squares of the residuals is minimum.

Let $S = \sum_1^n R^2$

ie $S = \sum_1^n [y - (ax + b)]^2$

Treating S as a function of 2 parameters a & b the necessary conditions for S to be minimum are

$$\frac{\partial S}{\partial a} = 0 \quad \& \quad \frac{\partial S}{\partial b} = 0$$

ie. $$2 \sum_1^n [y - (ax + b)](-x) = 0$$

and $$2 \sum_1^n [y - (ax + b)](-1) = 0$$

Dividing both the eqns by 2, we have

$$-\sum_1^n xy + \sum_1^n ax^2 + \sum_1^n bx = 0$$

$$-\sum_1^n y + \sum_1^n ax + \sum_1^n b = 0$$

But $\sum_1^n b = b + b + \cdots + n \text{ times}$

$= nb$

Hence $a\Sigma x^2 + b\Sigma x = \Sigma xy$

$a\Sigma x + nb = \Sigma y$

These eqns are called normal eqns. for fitting the straight line $y = ax + b$.

Problems

1) Fit a curve of the form $y = ax + b$ to following table.

x	1	2	3	4	5
y	14	13	9	5	2

Soln:-

x	y	xy	x^2
1	14	14	1
2	13	26	4
3	9	27	9
4	5	20	16
5	2	10	25
=15	=43	=97	55

$n = 5$

normal eqns are

$a\Sigma x + nb = \Sigma y$

$a\Sigma x^2 + b\Sigma x = \Sigma xy$

$a(15) + 5b = 43$

$a(55) + 15b = 97$

$15a + 5b = 43 \rightarrow (1)$

$55a + 15b = 97 \rightarrow (2)$

from eq ①

$$15a = 43 - 5b$$

$$a = \frac{43}{15} - \frac{5b}{15}$$

Sub a in eq ②

$$55a + 15b = 97$$

$$55\left(\frac{43}{15} - \frac{5b}{15}\right) + 15b = 97$$

$$11\left(\frac{43 - 5b}{3}\right) + 15b = 97$$

$$473 - 55b + 45b = 291$$

$$-10b = 291 - 473$$

$$-10b = -182$$

$$b = 18.2$$

Now Sub $b = 18.2$ in eq ①

$$15a + 5b = 43$$

$$15a + 5(18.2) = 43$$

$$15a + 91 = 43$$

$$15a = -48$$

$$a = -3.2$$

$\therefore a = -3.2$, $b = 18.2$

So, $y = ax + b$

$\boxed{y = (-3.2)x + 18.2}$ is a curve of best fit.

2) Fit the straight line $y = ax + b$ to the following data

x	-5	-3	-1	0	1	2	4
y	0.4	0.1	-0.2	-0.3	-0.3	0.1	0.4

Soln) -

x	y	xy	x^2
-5	0.4	-2	25
-3	0.1	-0.3	9
-1	-0.2	+0.2	1
0	-0.3	0	0
1	-0.3	-0.3	1
2	0.1	0.2	4
4	0.4	1.6	16
-2	+0.2	-0.6	56

$$n = 7$$

Normal eqns are

$$a\Sigma x + nb = \Sigma y$$

$$a\Sigma x^2 + b\Sigma x = \Sigma xy$$

$$a(-2) + (7)(b) = -0.2$$

$$\Rightarrow -2a + 7b = -0.2 \rightarrow (1)$$

$$a(56) + b(-2) = -0.6$$

$$56a - 2b = -0.6 \rightarrow (2)$$

$$-2a + 7b = +0.2 \rightarrow (1)$$

$$56a - 2b = -0.6 \rightarrow (2)$$

from eq (1) $\quad -2a = +0.2 - 7b$

$$a = -\frac{0.2}{2} + \frac{7}{2}b$$

Sub a in eq (2)

$$56a - 2b = -0.6$$

$$56\left(-\frac{0.2}{2} + \frac{7}{2}b\right) - 2b = -0.6$$

$$28\left(\frac{-0.2+7b}{2}\right) - 2b = -0.6$$

$-5.6 + 196b - 2b = -0.6$

$194b = -0.6 + 5.6$

$194b = 5$

$b = 0.025$

3) Fit a straight line $y = ax + b$ for the following data.

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

Sol:-

x	y	xy	x^2
1	1	1	1
3	2	6	9
4	4	16	16
6	4	24	36
8	5	40	64
9	7	56	81
11	8	88	121
14	9	126	196
$\Sigma x = 56$	$\Sigma y = 40$	$\Sigma xy = 364$	$\Sigma x^2 = 524$

The normal eqns are

$\Sigma y = a\Sigma x + nb \quad (n = 8)$

$\Sigma xy = a\Sigma x^2 + b\Sigma x$

The normal eqns become

$$56a + 8b = 40$$

$$524a + 56b = 364$$

By solving, we have

$a = 0.636 \approx 0.64$, $b = 0.545 \approx 0.55$ (0.5114)

$\therefore y = ax + b$

$$\boxed{y = 0.64x + 0.55}$$

4) Find the eqn of the best fitting straight line for the following data & hence estimate the value of the dependent variable corresponding to the value 30 of the independent variable.

x	5	10	15	20	25
y	16	19	23	26	30

Ans! -

$$a = 0.7, \quad b = 12.3$$

$$y = (0.7)x + 12.3$$

5) Fit a straight line in the least square sense for the following data.

x	50	70	100	120
y	12	15	21	25

Ans! -

$$a = 0.187 \approx 0.19, \quad b = 2.27 \approx 2.28$$

$$\boxed{y = (0.19)x + (2.28)}$$

Fitting of a curve of the form $y = ax^b$

Consider, $y = ax^b$

Take log on b.s.

we get,

$$\log_e y = \log(ax^b)$$

$$\log_e y = \log_e a + \log_e x^b$$

$$\log_e y = \log_e a + b \log_e x$$

$$Y = A + BX \quad \rightarrow (1)$$

where $Y = \log_e y$, $A = \log a$, $B = b$,

$X = \log_e x$

The normal eqns associated with (1) are as follows

$$\Sigma Y = nA + B\Sigma X \quad \rightarrow (2)$$

$$\Sigma XY = A\Sigma X + B\Sigma X^2 \quad \rightarrow (3)$$

Solving eqs (2) & (3), we obtain A, B.

But $\log_e a = A$

$\Rightarrow a = e^A$

Substituting the values of a, b in $y = ax^b$ we get the curve of best fit in the required form.

Remark:- We can similarly fit the curves $y = ae^{bx}$, $y = ab^x$ etc.

1) Fit a least square geometric curve $y = ax^b$ from the following data.

x	1	2	3	4	5
y	0.5	2	4.5	8	12.5

Consider $y = ax^b$

Take log on b.s.

$$\log y = \log a + b \log x$$

$$Y = A + bX$$

where $Y = \log y$, $A = \log a$, $X = \log x$

x	y	$X = \log x$	$Y = \log y$	XY	X^2
1	0.5	0	-0.6931	0	0
2	2	0.6931	0.6931	0.4803	0.4803
3	4.5	1.0986	1.5040	1.6522	1.2069
4	8	1.3862	2.0794	2.8824	1.9215
5	12.5	1.6094	2.5257	4.0648	2.5901
		4.7873	6.1091	9.0797	6.1988

normal eqns

$$nA + b\Sigma X = \Sigma Y$$

$$A\Sigma X + b\Sigma X^2 = \Sigma XY$$

Here $n = 5$

$$5A + 4.7873b = 6.1091$$

$$4.7873A + 6.1988b = 9.0797$$

$$A = -0.6932 \quad ; \quad b = 2$$

$\log a = A$

$$\log a = -0.6932$$

$$\Rightarrow a = e^{-0.6932} = 0.4999$$

$$\approx 0.5$$

$$y = ax^b$$

$y = (0.5)x^2$ curve of best fit