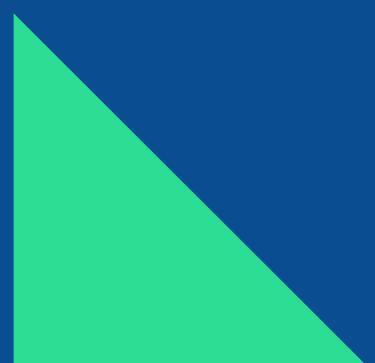




A NEW & EASY WAY TO LEARN

# AZ Documents.in

*Get All Vtu 18th Scheme  
notes for all branches here*



**THIS NOTES WAS  
DOWNLOADED FROM  
AZDOCUMENTS.IN**

[Click here to visit the website](#)

# Mat 41

## Module - 04

### Statistical Methods

Mean (Arithmetic mean) :-

If  $x_1, x_2, \dots, x_n$  be a set of  $n$  values of a variate  $x$ , the mean denoted by  $\bar{x}$  is defined as follows.

$$\bar{x} = \frac{\sum x}{n} \quad \text{or} \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

for a grouped data in the form of a frequency distribution,

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} \quad \text{or} \quad \bar{x} = \frac{\sum f x}{\sum f}$$

where  $f_i$ 's are the frequency of the classes having corresponding midpoint  $x_i$ .

Variance (V) and Standard deviation (SD) :-

If a variate  $x$  take values  $x_1, x_2, \dots, x_n$  the variance ( $V$ ) is defined as follows.

$$V = \frac{\sum (x - \bar{x})^2}{n} \quad \text{or} \quad V = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

Also for a grouped data

$$V = \frac{\sum f(x - \bar{x})^2}{\sum f} \quad \text{or} \quad V = \frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{\sum_{i=1}^n f_i}$$

Standard deviation (SD),  $\sigma = \sqrt{V}$  or  $\sigma^2 = V$

Alternative expression for  $\sigma^2$  :-

$$\text{Consider, } \sigma^2 = \frac{1}{n} \sum (x - \bar{x})^2$$

$$\begin{aligned} \sigma^2 &= \frac{1}{n} \sum [x^2 + 2x\bar{x} + (\bar{x})^2] \\ &= \frac{\sum x^2}{n} - 2(\bar{x})\bar{x} + \frac{n(\bar{x})^2}{n} \end{aligned}$$

Here,  $\frac{\sum x}{n} = \bar{x}$  &  $(\bar{x})^2$  being a constant added  
n times gives  $n(\bar{x})^2$ .

$$\text{i.e. } \sigma^2 = \frac{\sum x^2}{n} - 2(\bar{x})^2 + (\bar{x})^2$$

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

for a grouped data the expression will be of  
the form

$$\sigma^2 = \frac{\sum f x^2}{\sum f} - (\bar{x})^2$$

Example :-

- ① we shall find the mean and standard deviation of a set of observations 6, 8, 7, 5, 4, 9, 3

$$\text{Soln:- } \bar{x} = \frac{\sum x}{n},$$

$$\bar{x} = \frac{6+8+7+5+4+9+3}{7} = \frac{42}{7} = 6$$

$$\text{Thus mean } (\bar{x}) = 6$$

$$\checkmark \text{ or } \sigma^2 = \frac{\sum (x-\bar{x})^2}{n} = \frac{1}{7} \left\{ (6-6)^2 + (8-6)^2 + (7-6)^2 + (5-6)^2 + (4-6)^2 + (9-6)^2 + (3-6)^2 \right\}$$

$$\checkmark = \frac{28}{7} = 4$$

$$\therefore \sigma = \sqrt{4} = 2$$

$$\text{Alternate: } \sigma^2 = \frac{\sum x^2}{n} - (\bar{x})^2$$

$$= \frac{6^2 + 8^2 + 7^2 + 5^2 + 4^2 + 9^2 + 3^2}{7} - (6)^2$$

$$= \frac{280}{7} - 36$$

$$= 4$$

$$\text{Thus } SD = \sigma = 2.$$

- ② Let us find the mean & SD for the following grouped data

| class     | 1-10 | 11-20 | 21-30 | 31-40 | 41-50 | 51-60 |
|-----------|------|-------|-------|-------|-------|-------|
| Frequency | 3    | 16    | 26    | 31    | 16    | 8     |

| class  | f   | x    | fx     | (x - $\bar{x}$ ) | $(x - \bar{x})^2$ | $f(x - \bar{x})^2$ |
|--------|-----|------|--------|------------------|-------------------|--------------------|
| 1-10   | 3   | 5.5  | 16.5   |                  | 702.25            | 2106.75            |
| 11-20  | 16  | 15.5 | 248.0  |                  | 272.25            | 4356.00            |
| 21-30  | 26  | 25.5 | 663.0  |                  | 42.25             | 1098.50            |
| 31-40  | 31  | 35.5 | 1100.5 |                  | 12.25             | 379.75             |
| 41-50  | 16  | 45.5 | 728.0  |                  | 182.25            | 2916.00            |
| 51-60  | 8   | 55.5 | 444.0  |                  | 552.25            | 4418.00            |
| Totals | 100 |      | 3200   |                  |                   | 15275              |

$$\bar{x} = \frac{\sum f x}{\sum f} = \frac{3200}{100} = 32$$

$$s^2 = \frac{\sum f (x - \bar{x})^2}{\sum f} = \frac{15275}{100} = 152.75$$

$$\therefore s = \sqrt{152.75} = 12.36$$

Curve fitting :-

\* Fitting of a straight line :  $y = ax + b$   
 consider a set 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 between the observed and estimated values of  $y$ . By the method of least squares we find parameters  $a$  &  $b$  such that the sum of squares of the residuals is minimum.

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

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

problems :-

- ① 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  |  |

The normal eqn for fitting the straight line.

$$y = ax + b \text{ are } \sum y = a \sum x + nb \quad (n=8)$$

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

| $x$ | $y$ | $xy$ | $x^2$ |
|-----|-----|------|-------|
| 1   | 1   | 1    | 1     |
| 3   | 2   | 6    | 9     |
| 4   | 4   | 16   | 16    |
| 6   | 4   | 24   | 36    |
| 8   | 5   | 40   | 64    |
| 9   | 7   | 63   | 81    |
| 11  | 8   | 88   | 121   |
| 14  | 9   | 126  | 196   |

$$\sum x = 56 \quad \sum y = 40 \quad \sum xy = 364 \quad \sum x^2 = 524$$

The normal eqns become

$$56a + 8b = 40.$$

$$524a + 56b = 364$$

$$\therefore a = 0.63 \approx 0.64, b = 0.54 \approx 0.55$$

∴ Thus by substituting these values in  
 $y = ax + b$  we obtain the eqn.

$$y = 0.64x + 0.55$$

- ② 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 |

$$\therefore y = ax + b$$

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

$$\sum xy = a \sum x^2 + b \sum x \quad (n=5)$$

| $x$ | $y$ | $xy$ | $x^2$ |
|-----|-----|------|-------|
| 5   | 16  | 80   | 25    |
| 10  | 19  | 190  | 100   |
| 15  | 23  | 345  | 225   |
| 20  | 26  | 520  | 400   |
| 25  | 30  | 750  | 625   |

$$\sum x = 75$$

$$\sum y = 114$$

$$\sum xy = 1885$$

$$\sum x^2 = 1375$$

$$\therefore 75a + 5b = 114$$

$$1375a + 75b = 1885$$

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

$$y = ax + b$$

$$y = 0.7x + 12.3$$

when  $x = 30$ , we obtain  $y = 0.7(30) + 12.3 = 33.3$

③ A simply supported beam carries a concentrated load  $P$  at its mid point. Corresponding to various values of  $P$  the maximum deflection  $y$  is measured & is given in the following table.

| $P$ | 100  | 120  | 140  | 160  | 180  | 200  |
|-----|------|------|------|------|------|------|
| $y$ | 0.45 | 0.55 | 0.60 | 0.70 | 0.80 | 0.85 |

find a law of the form  $y = a + bp$  & hence estimate  $y$  when  $P$  is 150.

Sol: The normal equations associated with  $y = a + bp$  are as follows.

$$\sum y = na + b \sum P \quad (n=6)$$

$$\sum Py = a \sum P + b \sum P^2$$

| P   | y    | Py  | $P^2$ |
|-----|------|-----|-------|
| 100 | 0.45 | 45  | 10000 |
| 120 | 0.55 | 66  | 14400 |
| 140 | 0.60 | 84  | 19600 |
| 160 | 0.70 | 112 | 25600 |
| 180 | 0.80 | 144 | 32400 |
| 200 | 0.85 | 170 | 40000 |

$$\sum P = 900 \quad \sum y = 3.95 \quad \sum Py = 621 \quad \sum P^2 = 142000$$

$\therefore$  eqns  $\Rightarrow$

$$6a + 900b = 3.95$$

$$900a + 142000b = 621$$

$$\therefore a = 0.0476, \quad b = 0.0041$$

Thus the required Law is  $y = 0.0476 + 0.0041P$

Also when  $P = 150$ ,  $y = 0.6626 \approx 0.66$

④ Fit a straight line to the following data.

| year                    | 1961 | 1971 | 1981 | 1991 | 2001 |
|-------------------------|------|------|------|------|------|
| production<br>(in tons) | 8    | 10   | 12   | 10   | 16   |

Also find the expected production in the year 2006.

Soln:- let  $x = x - 1981$  & the line of fit with be  $y = a + bx$

$$\sum y = na + b \sum x \quad (n=5)$$

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

(4)

| $x$ | $y$ | $xy$ | $x^2$ |
|-----|-----|------|-------|
| -20 | 8   | -160 | 400   |
| -10 | 10  | -100 | 100   |
| 0   | 12  | 0    | 0     |
| 10  | 10  | 100  | 100   |
| 20  | 16  | 320  | 400   |

$$\sum x = 0, \sum y = 56, \sum xy = 160, \sum x^2 = 1000$$

The normal equations become,

$$5a = 56 \quad \text{and} \quad 1000b = 160$$

$$a = 11.2 \quad \text{and} \quad b = 0.16$$

Hence  $y = a + bx$ , with  $x = x - 1981$  becomes

$$y = 11.2 + 0.16(x - 1981)$$

Thus  $y = -305.76 + 0.16x$  is the required line of fit.

Also when  $x = 2006$ ,  $y = -305.76 + 0.16(2006)$

$$y = 15.2$$

Expected production in the year 2006 is 15.2 tons.

(5) Find the eqn of the best fitting straight line for the following data.

|    |     |    |    |   |   |   |
|----|-----|----|----|---|---|---|
| i) | $x$ | 1  | 2  | 3 | 4 | 5 |
|    | $y$ | 14 | 13 | 9 | 5 | 2 |

|     |     |   |   |    |    |    |    |
|-----|-----|---|---|----|----|----|----|
| ii) | $x$ | 0 | 1 | 2  | 3  | 4  | 5  |
|     | $y$ | 9 | 8 | 24 | 28 | 26 | 20 |

|      |     |      |      |      |      |      |      |      |
|------|-----|------|------|------|------|------|------|------|
| iii) | $x$ | 62   | 64   | 65   | 69   | 70   | 71   | 72   |
|      | $y$ | 65.7 | 66.8 | 67.2 | 69.3 | 69.8 | 70.5 | 70.9 |

|     |     |    |    |    |    |    |    |    |
|-----|-----|----|----|----|----|----|----|----|
| iv) | $x$ | 1  | 2  | 3  | 4  | 5  | 6  | 7  |
|     | $y$ | 80 | 90 | 92 | 83 | 94 | 99 | 92 |

|     |                                      |      |      |      |      |      |
|-----|--------------------------------------|------|------|------|------|------|
| (6) | Year (x)                             | 1911 | 1921 | 1931 | 1941 | 1951 |
|     | Production (y)<br>(in thousand tons) | 8    | 10   | 12   | 10   | 6    |

Soln:- Let  $x = x - 1931$  & the line of fit will be  $y = a + bx$

The normal eqns associated with  $y = a + bx$  are as follows.

$$\sum y = na + b \sum x \quad (n=5)$$

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

| x   | y  | xy   | $x^2$ |
|-----|----|------|-------|
| -20 | 8  | -160 | 400   |
| -10 | 10 | -100 | 100   |
| 0   | 12 | 0    | 0     |
| 10  | 10 | 100  | 100   |
| 20  | 6  | 120  | 400   |

$$\sum x = 0 \quad \sum y = 46 \quad \sum xy = -40 \quad \sum x^2 = 1000$$

The normal eqns become,

$$46 = 5a \rightarrow ①$$

$$-40 = 1000b \rightarrow ②$$

$$\text{Eqn } ① \quad a = 9.2$$

$$\text{Eqn } ② \quad b = -0.04$$

Hence  $y = a + bx$ , with  $x = x - 1931$

$$y = 9.2 + (-0.04)(x - 1931)$$

$$= 9.2 - 0.04x + 77.24$$

Thus,  $y = 86.44 - 0.04x$  is the required line of fit.

\* fitting of a second degree parabola

$$y = ax^2 + bx + c$$

consider a set of 'n' given values  $(x, y)$  for fitting the curve  $y = ax^2 + bx + c$ . The residual  $R = y - (ax^2 + bx + c)$  is the difference b/w the observed & estimated value of  $y$ . we have to find parameters  $a, b, c$  such that the sum of the squares of the residuals is the least.

$$\sum y = a \sum x^2 + b \sum x + nc$$

$$\sum xy = a \sum x^3 + b \sum x^2 + c \sum x$$

$$\sum x^2 y = a \sum x^4 + b \sum x^3 + c \sum x^2$$

- \* ① fit a best fitting parabola  $y = ax^2 + bx + c$  for the following data:

$x \quad 1 \quad 2 \quad 3 \quad 4 \quad 5$  and hence estimate  
 $y \quad 10 \quad 12 \quad 13 \quad 16 \quad 19$  at  $x=6$ .

Soln:- The normal eqns associated with  $y = ax^2 + bx + c \rightarrow (*)$  are as follows.

$$y = ax^2 + bx + c \rightarrow ①$$

$$\sum y = a \sum x^2 + b \sum x + nc \rightarrow ②$$

$$\sum xy = a \sum x^3 + b \sum x^2 + c \sum x \rightarrow ③$$

$$\sum x^2 y = a \sum x^4 + b \sum x^3 + c \sum x^2 \rightarrow ④$$

| $x$ | $y$ | $xy$ | $x^2$ | $x^2 y$ | $x^3$ | $x^4$ |
|-----|-----|------|-------|---------|-------|-------|
| 1   | 10  | 10   | 1     | 10      | 1     |       |
| 2   | 12  | 24   | 4     | 48      | 8     | 16    |
| 3   | 13  | 39   | 9     |         |       |       |
| 4   | 16  | 64   | 16    | 256     | 64    | 256   |
| 5   | 19  | 95   | 25    | 475     | 125   | 625   |

$$\sum x = 15 \quad \sum y = 70 \quad \sum xy = 232 \quad \sum x^2 = 55 \quad \sum x^2 y = 906 \quad \sum x^3 = 225 \quad \sum x^4 = 979$$

$$\text{eqn } ① \Rightarrow 70 = 55a + 15b + 5c$$

$$\text{eqn } ② \Rightarrow 232 = 225a + 55b + 15c$$

$$\text{eqn } ③ \Rightarrow 906 = 979a + 225b + 55c$$

$$\therefore a = 0.2857 \approx 0.29, b = 0.4857 \approx 0.49, c = 9.4.$$

Thus the required second degree parabola is

$$y = 0.29x^2 + 0.49x + 9.4 \text{ also at } x=6$$

$$y = 22.78$$

- \* ② fit a parabola  $y = ax^2 + bx + c$  for the following data.
- | $x$ | 0 | 1   | 2   | 3   | 4   |
|-----|---|-----|-----|-----|-----|
| $y$ | 1 | 1.8 | 1.3 | 2.5 | 2.3 |

Soln:- The normal eqn associated  $y = ax^2 + bx + c$  are

$$\sum y = a\sum x + b\sum x^2 + c\sum x^3 \rightarrow ①$$

$$\sum xy = a\sum x^2 + b\sum x^3 + c\sum x^4 \rightarrow ②$$

$$\sum x^2 y = a\sum x^3 + b\sum x^4 + c\sum x^5 \rightarrow ③$$

| $x$ | $y$ | $xy$ | $x^2$ | $x^2 y$ | $x^3$ | $x^4$ |
|-----|-----|------|-------|---------|-------|-------|
| 0   | 1   | 0    | 0     | 0       | 0     | 0     |
| 1   | 1.8 | 1.8  | 1     | 1.8     | 1     | 1     |
| 2   | 1.3 | 2.6  | 4     | 5.2     | 8     | 16    |
| 3   | 2.5 | 7.5  | 9     | 22.5    | 27    | 81    |
| 4   | 2.3 | 9.2  | 16    | 36.8    | 64    | 256   |

$$\sum x = 10 \quad \sum y = 8.9 \quad \sum xy = 21.1 \quad \sum x^2 = 30 \quad \sum x^2 y = 66.3 \quad \sum x^3 = 100 \quad \sum x^4 = 354$$

$$\text{Eqn } ① \Rightarrow 8.9 = 5a + 10b + 30c$$

$$21.1 = 10a + 30b + 100c$$

$$\text{Eqn } ② \Rightarrow 66.3 = 30a + 100b + 354c$$

$$\text{Eqn } ③ \Rightarrow a = 1.0771, \quad b = 0.4157, \quad c = -0.0214$$

Thus the parabola of fit is

$$y = 1.0771 + 0.4157x - 0.0214x^2$$

- \* ③ fit a second degree parabola to the following data.
- | $x$ | 1.0 | 1.5 | 2.0 | 2.5 | 3.0 | 3.5 | 4.0 |
|-----|-----|-----|-----|-----|-----|-----|-----|
| $y$ | 1.1 | 1.3 | 1.6 | 2.0 | 2.7 | 3.4 | 4.1 |

- \* ④ fit a second degree parabola to the following data:

| $x$ | 0 | 1   | 2   | 3   | 4   |
|-----|---|-----|-----|-----|-----|
| $y$ | 1 | 1.8 | 1.3 | 2.5 | 6.3 |

\* fitting of a curve of the form  $y = ae^{bx}$ .<sup>⑥</sup>

consider,  $y = ae^{bx}$ . Taking logarithm (to the base e) on both sides we get.

$$\log_e y = \log(ae^{bx})$$

$$= \log_e a + \log_e e^{bx}$$

$$= \log_e a + bx \log_e e$$

$$\log_e y = \log_e a + bx \quad ①$$

$$y = A + BX \rightarrow ①$$

where  $y = \log_e y$ ,  $A = \log_e a$ ,  $B = b$ ,  $X = x$

It is evident that eqn ① is the eqn of a straight line & the associated normal eqns are as follows.

$$\sum y = nA + B \sum X \rightarrow ②$$

$$\sum xy = A \sum X + B \sum X^2 \rightarrow ③$$

Solving eqn ② & ③ we obtain 'A' & 'B'. But

$$\log_e a = A \Rightarrow a = e^A. \text{ Also } b = B.$$

Substituting these values in  $y = ae^{bx}$  we get the curve of best fit, in the required form.

\* \* fit a curve of the form  $y = ae^{bx}$  to the following data.

|   |     |     |     |      |      |
|---|-----|-----|-----|------|------|
| x | 7.7 | 100 | 185 | 239  | 285  |
| y | 2.4 | 3.4 | 7.0 | 11.1 | 19.6 |

Soln:- consider,  $y = ae^{bx} \rightarrow *$

$$y = A + bx$$

The normal eqns are as follows.

$$\sum y = nA + b \sum X \rightarrow ①$$

$$\sum xy = A \sum X + B \sum X^2 \rightarrow ②$$

where  $y = \log_e y$ ,  $A = \log_e a$

$$\log mn = \log m + \log n$$

$$\log m^n = n \log m$$

$$\log e = 1$$

| $x$            | $y$  | $y = \log_e y$  | $xy$                  | $x^2$               |
|----------------|------|-----------------|-----------------------|---------------------|
| 77             | 2.4  | 0.8754          | 67.4058               | 5929                |
| 100            | 3.4  | 1.2237          | 122.37                | 10000               |
| 185            | 7.0  | 1.9459          | 359.9915              | 34225               |
| 239            | 11.1 | 2.4069          | 575.2491              | 57121               |
| 285            | 19.6 | 2.9755          | 848.0175              | 81225               |
| $\sum x = 886$ |      | $\sum y = 43.5$ | $\sum xy = 1973.0339$ | $\sum x^2 = 188500$ |
|                |      |                 |                       |                     |

$$\text{Eqn ①} \Rightarrow 9.4274 = 5A + 886b$$

$$\text{Eqn ②} \Rightarrow 1973.0339 = 886A + 188500b$$

$$A = 0.18387 \times 10^{-3}$$

$$A = \log_e a = 0.1838 \times 10^{-3} + An = 1.2017$$

$$a = e^A = e^{0.1838 \times 10^{-3}}$$

$$= e^{0.1838 \times 10^{-3}} = 1.2017$$

$$b = 9.6028 \times 10^{-3}$$

$$a = 1.2017$$

The curve of fit  $y = a e^{bx}$  is the curve of fit.

$$\text{Thus } y = (1.2017)e^{9.6028 \times 10^{-3}x}$$

② Fit an exponential curve of the form

$y = ae^{bx}$  by the method of least squares for the following data.

| No of petals  | 5   | 6  | 7  | 8 | 9 | 10 |
|---------------|-----|----|----|---|---|----|
| No of flowers | 133 | 55 | 23 | 7 | 2 | 2  |

(7)

| $x$ | $y$ | $y = \log y$ | $xy$    | $x^2$ |
|-----|-----|--------------|---------|-------|
| 5   | 133 | 4.8903       | 24.4515 | 25    |
| 6   | 55  | 4.0073       | 24.0438 | 36    |
| 7   | 23  | 3.1355       | 21.9485 | 49    |
| 8   | 7   | 1.9459       | 15.5672 | 64    |
| 9   | 2   | 0.6931       | 6.2379  | 81    |
| 10  | 2   | 0.6931       | 6.9310  | 100   |

$$\sum x = 45 \quad \sum y = 15.3652 \quad \sum xy = 99.1799 \quad \sum x^2 = 355$$

The normal eqn becomes

$$6A + 45b = 15.3652$$

$$45A + 355b = 99.1799$$

$$A = 9.4433 \quad \& \quad b = -0.9177$$

Thus the required curve of fit is,

$$y = (126.23.3) e^{-0.9177x}$$

- ③ Fit a curve of the form  $y = ax^b$  for the data
- | $x$ | 1    | 2    | 3    | 4   | 5   | 6   |
|-----|------|------|------|-----|-----|-----|
| $y$ | 2.98 | 4.26 | 5.21 | 6.1 | 6.8 | 7.5 |

- ④ Find the eqn of the best fitting curve in the form  $y = ae^{bx}$  for the data

| $x$ | 0    | 2  | 4     |
|-----|------|----|-------|
| $y$ | 5.02 | 10 | 31.62 |

## correlation and correlation co-efficient :-

co-variation of two independent magnitudes is known as correlation. If two variables  $x_i \& y_j$  are related in such a way that  $\uparrow \text{ or } \downarrow$  in one of them corresponds to  $\uparrow \text{ or } \downarrow$  in the other, we say that the variables are truly correlated. Also if increase  $\uparrow$  decrease in one of them corresponds to decrease or  $\uparrow$  in the other, the variables are said to be very correlated.

The numerical measure of correlation b/w two variables  $x_i \& y_j$  is known as Pearson's coefficient of correlation usually denoted by  $r$ .

$r$  is defined as follows.

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{n \sqrt{\sum x^2} \sqrt{\sum y^2}} \rightarrow ①$$

This can be put in an alternative form as follows. If  $x = x - \bar{x}$ ,  $y = y - \bar{y}$  we can write.

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

$$\sqrt{\sum y^2} = \sqrt{\frac{\sum (y - \bar{y})^2}{n}} = \sqrt{\frac{\sum y^2}{n}}$$

$$\therefore \sqrt{\sum x^2} \sqrt{\sum y^2} = \sqrt{\frac{\sum x^2}{n}} \sqrt{\frac{\sum y^2}{n}} = \sqrt{\sum x^2} \sqrt{\sum y^2}$$

$$\therefore n \sqrt{\sum x^2} \sqrt{\sum y^2} = \sqrt{\sum x^2} \cdot \sqrt{\sum y^2}$$

Thus eqn ① becomes

$$r = \frac{\sum xy}{\sqrt{\sum x^2} \sqrt{\sum y^2}}$$

Note:- \* The coefficient of correlation numerically does not exceed unity. i.e.  $-1 \leq r \leq +1$

\* If  $r = \pm 1$  we say that 'x & y' are perfectly correlated & if  $r = 0$  we say that 'x & y' are non correlated. (8)

Alternative formula for the correlation co-efficient  $r$ :

$$r = \frac{\sum x^2 + \sum y^2 - \sum x^2 y}{2 \sqrt{\sum x^2} \sqrt{\sum y^2}}$$

Proof: let  $z = x - y$

$$\therefore \frac{\sum z}{n} = \frac{\sum x}{n} - \frac{\sum y}{n} \quad (6) \quad \bar{z} = \bar{x} - \bar{y}$$

$$\text{Hence, } (z - \bar{z}) = (x - y) - (\bar{x} - \bar{y})$$

$$\text{i.e. } (z - \bar{z}) = (x - \bar{x}) - (y - \bar{y})$$

Squaring both sides, taking summation &  
dividing by 'n' we have,

$$\begin{aligned} \frac{\sum (z - \bar{z})^2}{n} &= \frac{\sum [(x - \bar{x}) - (y - \bar{y})]^2}{n} \\ &= \frac{\sum (x - \bar{x})^2 + \sum (y - \bar{y})^2 - 2 \sum (x - \bar{x})(y - \bar{y})}{n} \\ &\therefore r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2} \sqrt{\sum (y - \bar{y})^2}} \end{aligned}$$

$$\text{i.e. } \sum z^2 = \sum x^2 + \sum y^2 - 2 \sum x y$$

$$\text{i.e. } \sum z^2 = \sum x^2 + \sum y^2 - 2 \sum x \bar{y}$$

$$\text{Thus } r = \frac{\sum x^2 + \sum y^2 - \sum x^2 y}{2 \sqrt{\sum x^2} \sqrt{\sum y^2}}$$

Note: In general if  $z = ax + by$  we can obtain as before

$$\sum z^2 = a^2 \sum x^2 + b^2 \sum y^2 + 2ab \sum xy$$

$$\text{i.e. } \sum z^2 = a^2 \sum x^2 + b^2 \sum y^2 + 2ab \sum xy$$

## Regression :-

Regression is an estimation of one independent variable in terms of the other. If 'x' & 'y' are correlated, the best fitting straight line in the least square sense give reasonably a good relation b/w 'x' & 'y'.

The best fitting straight line of the form  $y = ax + b$  ('x' being the independent variable) is called the regression line of 'y' on 'x' &  $x = ay + b$  ('y' being the independent variable) is called the regression line of 'x' on 'y'.  $y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$ ,  $x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$

Note :- 1) The lines of regression (2)

$$y = \frac{\sum xy}{\sum x^2} (x) \text{ and } x = \frac{\sum xy}{\sum y^2} (y)$$

where  $x = x - \bar{x}$  &  $y = y - \bar{y}$ .

This form will be useful to find out the coefficient of correlation by first obtaining the lines of regression as we have deduced that

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

2) To compute the coefficient of correlation we prefer to use the formula.

$$r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{xy}^2}{2\sigma_x \sigma_y}$$

where SDs can be found by applying the formula.

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

If  $\bar{x}$  &  $\bar{y}$  are integers computation of  $r$  by the formula.

$$r = \frac{\sum xy}{\sqrt{\sum x^2} \sqrt{\sum y^2}} \text{ is convenient where } x = x - \bar{x}, y = y - \bar{y}.$$

problems:-

- ① compute the coefficient of correlation and the equation of the lines of regression for the data.

$x \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7$

$y \quad 9 \quad 8 \quad 10 \quad 12 \quad 11 \quad 13 \quad 14$

Soln:- we have  $\gamma = \frac{5x^2 + 5y^2 - 5\bar{x}\bar{y}}{2\sqrt{5x} \sqrt{5y}} \rightarrow ①$

| $x$           | $y$           | $z = x-y$      | $x^2$            | $y^2$            | $z^2$            |
|---------------|---------------|----------------|------------------|------------------|------------------|
| 1             | 9             | -8             | 1                | 81               | 64               |
| 2             | 8             | -6             | 4                | 64               | 36               |
| 3             | 10            | -7             | 9                | 100              | 49               |
| 4             | 12            | -8             | 16               | 144              | 64               |
| 5             | 11            | -6             | 25               | 121              | 36               |
| 6             | 13            | -7             | 36               | 169              | 49               |
| 7             | 14            | -7             | 49               | 196              | 49               |
| $\sum x = 28$ | $\sum y = 77$ | $\sum z = -49$ | $\sum x^2 = 140$ | $\sum y^2 = 875$ | $\sum z^2 = 347$ |

$$\bar{x} = \frac{\sum x}{n}, \quad \bar{y} = \frac{\sum y}{n}, \quad \bar{z} = \frac{\sum z}{n}$$

$$\bar{x} = \frac{28}{7} = 4, \quad \bar{y} = \frac{77}{7} = 11, \quad \bar{z} = \frac{-49}{7} = -7$$

$$5x^2 = \frac{\sum x^2}{n} - (\bar{x})^2, \quad 5y^2 = \frac{\sum y^2}{n} - (\bar{y})^2, \quad 5z^2 = \frac{\sum z^2}{n} - (\bar{z})^2$$

$$= \frac{140}{7} - (4)^2 = 4, \quad = \frac{875}{7} - (11)^2 = 4, \quad = \frac{347}{7} - (-7)^2 = 0.5$$

$$\text{we have, } 5x^2 = 4, \quad 5y^2 = 4, \quad 5z^2 = 5x^2 - 5x \cdot y = 0.57$$

$$\text{Eqn } ① \Rightarrow \gamma = \frac{4+4-0.57}{2\sqrt{4}\sqrt{4}} = 0.92875 \approx 0.93,$$

Thus  $\gamma = 0.93$

The lines of regression are given by

$$y - \bar{y} = \gamma \frac{5y}{5x} (x - \bar{x}) \quad | \quad x - \bar{x} = \gamma \frac{5x}{5y} (y - \bar{y})$$

$$y - 11 = \frac{(0.93) \cdot 2}{2} (x - 4), \quad x - 4 = \frac{(0.93) \cdot 2}{2} (y - 11)$$

$$y - 11 = 0.93(x - 4), \quad x - 4 = 0.93(y - 11)$$

Thus  $y = 0.93x + 7.28$  &  $x = 0.93y - 6.23$  are the lines of regression.

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

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

Soln:- Here  $\bar{x} = 4$ ,  $\bar{y} = 11$

$$\therefore x = x - \bar{x}, y = y - \bar{y}$$

$$= x - 4, y = y - 11$$

| $x$ | $y$ | $x$ | $y$ | $xy$ | $x^2$ | $y^2$ |
|-----|-----|-----|-----|------|-------|-------|
| 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   | 4   | 4    | 4     | 16    |
| 7   | 14  | 3   | 3   | 9    | 9     | 9     |

$$\sum xy = 26, \sum x^2 = 28, \sum y^2 = 28$$

We shall consider regression lines in the form.

$$y = \frac{\sum xy}{\sum x^2} \cdot x \quad \text{and} \quad x = \frac{\sum xy}{\sum y^2} \cdot y$$

$$\text{i.e } y - 11 = \frac{26}{28} (x - 4), \quad x - 4 = \frac{26}{28} (y - 11)$$

$$y - 11 = 0.93(x - 4), \quad x - 4 = 0.93(y - 11)$$

$$y = 0.93x + 7.28, \quad x = 0.93y - 6.23$$

These are the regression lines and we compute ' $r$ ' as the geometric mean of the regression coefficients.

$$\text{i.e } r = \sqrt{(\text{coeff of } x)(\text{coeff of } y)} = \sqrt{(0.93)(0.93)}$$

$$r = 0.93$$

The sign of ' $r$ ' must be +ve since both the regression coefficients are +ve.

$$\text{Thus } r = 0.93.$$

③ Find the correlation coefficient & the eqn of the line of regression for the following values of  $x$  &  $y$ :

$x = 1, 2, 3, 4, 5$

$y = 2, 5, 3, 8, 7$

Soln:  $n = 5$

| $x$           | $y$           | $z = x - y$    | $x^2$           | $y^2$            | $z^2$           |
|---------------|---------------|----------------|-----------------|------------------|-----------------|
| 1             | 2             | -1             | 1               | 4                | 1               |
| 2             | 5             | -3             | 4               | 25               | 9               |
| 3             | 3             | 0              | 9               | 9                | 0               |
| 4             | 8             | -4             | 16              | 64               | 16              |
| 5             | 7             | -2             | 25              | 49               | 4               |
| $\sum x = 15$ | $\sum y = 25$ | $\sum z = -10$ | $\sum x^2 = 55$ | $\sum y^2 = 151$ | $\sum z^2 = 30$ |

$$\bar{x} = \frac{\sum x}{n}, \bar{y} = \frac{\sum y}{n}$$

$$\bar{z} = \frac{\sum z}{n}$$

$$\bar{x} = \frac{15}{5} = 3, \bar{y} = \frac{25}{5} = 5,$$

$$\bar{z} = \frac{-10}{5} = -2$$

$$\sigma_x^2 = \frac{\sum x^2}{n} - (\bar{x})^2 = \frac{55}{5} - (3)^2 = 2$$

$$\sigma_y^2 = \frac{\sum y^2}{n} - (\bar{y})^2 = \frac{151}{5} - (5)^2 = 5.2$$

$$\sigma_z^2 = \frac{\sum z^2}{n} - (\bar{z})^2 = \frac{30}{5} - (-2)^2 = 2$$

We have,  $\sigma_x^2 = 2, \sigma_y^2 = 5.2, \sigma_z^2 = 2$

$$\text{Now, } r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_z^2}{2\sigma_x \sigma_y}$$

$$r = \frac{2+5.2-2}{2\sqrt{2}\sqrt{5.2}} = 0.8062 \approx 0.81$$

Thus  $r = 0.81$

The eqn of the regression lines are as follows.

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

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

$$y - 5 = (0.81) \cdot \frac{\sqrt{5.2}}{\sqrt{2}} (x - 3),$$

$$x - 3 = 0.81 \cdot \frac{\sqrt{2}}{\sqrt{5.2}} (y - 5)$$

$y = 5 = 1.306(x-3)$ ,  $x-3 = 0.502(y-5)$   
 Thus  $y = 1.306x + 1.082$  &  $x = 0.502y + 0.49$   
 These are the lines of regression.

- ④ 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 |

Soln:- There  $n = 10$

| $x$ | $y$ | $z = x-y$ | $x^2$ | $y^2$ | $z^2$ |
|-----|-----|-----------|-------|-------|-------|
| 1   | 10  | -9        | 1     | 100   | 81    |
| 2   | 12  | -10       | 4     | 144   | 100   |
| 3   | 16  | -13       | 9     | 256   | 169   |
| 4   | 28  | -24       | 16    | 784   | 576   |
| 5   | 25  | -20       | 25    | 625   | 400   |
| 6   | 36  | -30       | 36    | 1296  | 900   |
| 7   | 41  | -34       | 49    | 1681  | 1156  |
| 8   | 49  | -41       | 64    | 2401  | 1681  |
| 9   | 40  | -31       | 81    | 1600  | 961   |
| 10  | 50  | -40       | 100   | 2500  | 1600  |

$$\sum x = 55 \quad \sum y = 307 \quad \sum z = -252 \quad \sum x^2 = 385 \quad \sum y^2 = 11387 \quad \sum z^2 = 7624$$

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

$$\bar{y} = \frac{\sum y}{n} = \frac{307}{10} = 30.7$$

$$\bar{z} = \frac{\sum z}{n} = \frac{-252}{10} = -25.2$$

$$\bar{S_x}^2 = \frac{\sum x^2}{n} - (\bar{x})^2 = \frac{385}{10} - (5.5)^2 = 8.25, \bar{S_x} = 2.87$$

$$\bar{S_y}^2 = \frac{\sum y^2}{n} - (\bar{y})^2 = 1138.7 - (30.7)^2 = 196.21, \bar{S_y} = 14.01$$

$$\bar{S_z}^2 = \bar{S_{x-y}}^2 = \frac{\sum z^2}{n} - (\bar{z})^2 = 762.4 - (-25.2)^2 = 127.36$$

we have,  $\gamma = \frac{\bar{S_x}^2 + \bar{S_y}^2 - \bar{S_{x-y}}^2}{2\bar{S_x}\bar{S_y}}$

$$= \frac{8.25 + 196.21 - 127.36}{2 \times 2.87 \times 14.01} = 0.96$$

Thus

Equation of the lines of regression are

$$y - \bar{y} = \gamma \frac{\bar{S_x}}{\bar{S_y}} (x - \bar{x}), \quad x - \bar{x} = \gamma \frac{\bar{S_x}}{\bar{S_y}} (y - \bar{y})$$

on substituting & simplifying we get,

$$y - 30.7 = 0.96 \times \frac{14.01}{2.87} (x - 5.5)$$

$$y = 4.686x + 4.927$$

$$x - 5.5 = 0.96 \times \frac{2.87}{14.01} (y - 30.7)$$

$$x = 0.197y - 0.548$$

These are the lines of regression.

- \* ⑤ find the coefficient of correlation for the following data.

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

Soln:- we have  $\bar{x} = \frac{\sum x}{n} = \frac{120}{6} = 20$

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

Let  $x = x - \bar{x}$ , &  $y = y - \bar{y}$

$$x = x - 20, \quad y = y - 21$$

| x  | y  | $x$ | y   | $x^2$            | $y^2$            | $xy$            |
|----|----|-----|-----|------------------|------------------|-----------------|
| 10 | 18 | -10 | -3  | 100              | 9                | 30              |
| 14 | 18 | -6  | -9  | 36               | 81               | 54              |
| 18 | 24 | -2  | 3   | 4                | 9                | -6              |
| 22 | 6  | 2   | -15 | 4                | 225              | -30             |
| 26 | 30 | 6   | 9   | 36               | 81               | 54              |
| 30 | 36 | 10  | 15  | 100              | 225              | 150             |
|    |    |     |     | $\sum x^2 = 280$ | $\sum y^2 = 630$ | $\sum xy = 252$ |

$$\gamma = \frac{\sum xy}{\sqrt{\sum x^2} \sqrt{\sum y^2}} = \frac{252}{\sqrt{280} \times \sqrt{630}} = 0.6$$

- \* ⑥ Find the correlation coefficient and the eqn of the line of regression for the following.

|   |   |   |   |   |   |
|---|---|---|---|---|---|
| x | 1 | 2 | 3 | 4 | 5 |
| y | 2 | 5 | 3 | 8 | 7 |

Soln:- Let  $n=5$

| x             | y             | $z = x-y$      | $x^2$           | $y^2$            | $z^2$           |
|---------------|---------------|----------------|-----------------|------------------|-----------------|
| 1             | 2             | -1             | 1               | 4                | 1               |
| 2             | 5             | -3             | 4               | 25               | 9               |
| 3             | 3             | 0              | 9               | 9                | 0               |
| 4             | 8             | -4             | 16              | 64               | 16              |
| 5             | 7             | -2             | 25              | 49               | 4               |
| $\sum x = 15$ | $\sum y = 25$ | $\sum z = -10$ | $\sum x^2 = 55$ | $\sum y^2 = 151$ | $\sum z^2 = 30$ |

$$\bar{x} = \frac{\sum x}{n} = \frac{15}{5} = 3, \bar{y} = \frac{\sum y}{n} = \frac{25}{5} = 5, \bar{z} = \frac{\sum z}{n} = \frac{-10}{5} = -2$$

$$S_x^2 = \frac{\sum x^2}{n} - (\bar{x})^2 = \frac{55}{5} - (3)^2 = 2$$

$$S_y^2 = \frac{\sum y^2}{n} - (\bar{y})^2 = \frac{151}{5} - (5)^2 = 5.2$$

$$S_z^2 = S_{x-y}^2 = \frac{\sum z^2}{n} - (\bar{z})^2 = \frac{30}{5} - (-2)^2 = 2$$

$$\text{Now } = \frac{S_x^2 + S_y^2 - S_z^2}{2 S_x S_y}$$

$$\gamma = \frac{2+5.2-2}{2\sqrt{2}\sqrt{5.2}} = 0.8062 \approx 0.81$$

$$\text{Thus } \gamma = 0.81$$

The eqns of the regression lines are as follows.

$$y - \bar{y} = \gamma \frac{\sigma_y}{\sigma_x} (x - \bar{x}), \quad x - \bar{x} = \gamma \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

$$y - 5 = (0.81) \frac{\sqrt{5.2}}{\sqrt{2}} (x - 3), \quad x - 3 = (0.81) \frac{\sqrt{2}}{\sqrt{5.2}} (y - 5)$$

$$y - 5 = 1.306 (x - 3), \quad x - 3 = 0.502 (y - 5)$$

$$\text{Thus } y = 1.306x + 1.082 \quad & x = 0.502y + 0.49$$

These are the lines of regression.

- \* 7) 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 |

Soln:- Here  $n = 10$

| $x$ | $y$ | $z = x - y$ | $x^2$ | $y^2$ | $z^2$ |
|-----|-----|-------------|-------|-------|-------|
| 1   | 10  | -9          | 1     | 100   | 81    |
| 2   | 12  | -10         | 4     | 144   | 100   |
| 3   | 16  | -13         | 9     | 256   | 169   |
| 4   | 28  | -24         | 16    | 784   | 576   |
| 5   | 25  | -20         | 25    | 625   | 400   |
| 6   | 36  | -30         | 36    | 1296  | 1156  |
| 7   | 41  | -34         | 49    | 1681  | 1681  |
| 8   | 49  | -41         | 64    | 2401  | 961   |
| 9   | 40  | -31         | 81    | 1600  | 1600  |
| 10  | 50  | -40         | 100   | 2500  | 7624  |

$$\sum x = 55, \quad \sum y = 307, \quad \sum z = -252, \quad \sum x^2 = 385, \quad \sum y^2 = 11387$$

$$\bar{x} = \frac{\sum x}{n} = \frac{55}{10} = 5.5, \quad \bar{y} = \frac{\sum y}{n} = \frac{307}{10} = 30.7, \quad \bar{z} = \frac{\sum z}{n} = \frac{-252}{10} = -25.2$$

$$\bar{Sx}^2 = \frac{\sum x^2}{n} - (\bar{x})^2 = \frac{385}{10} - (5.5)^2 = 8.25, \bar{Sx} = 2.87$$

$$\bar{Sy}^2 = \frac{\sum y^2}{n} - (\bar{y})^2 = \frac{11387}{10} - (30.7)^2 = 196.21, \bar{Sy} = 14.01$$

$$\bar{Sz}^2 = \bar{Sx-y}^2 = \frac{\sum z^2}{n} - (\bar{z})^2 = \frac{7624}{10} - (-25.2)^2 = 127.36$$

we have  $\gamma = \frac{\bar{Sx}^2 + \bar{Sy}^2 - \bar{Sx-y}^2}{2\bar{Sx}\bar{Sy}}$

$$\gamma = \frac{8.25 + 196.21 - 127.36}{2 \times 2.87 \times 14.01} = 0.96$$

thus  $\gamma = 0.96$

eqn of the line of regression are

$$y - \bar{y} = \gamma \frac{\bar{Sx}}{\bar{Sx}} (x - \bar{x}), x - \bar{x} = \gamma \frac{\bar{Sx}}{\bar{Sx}} (y - \bar{y})$$

$$y - 30.7 = 0.96 \times \frac{14.01}{2.87} (x - 5.5), (x - 5.5) = 0.96 \times \frac{2.87}{14.01} (y - 30.7)$$

$$y = 4.686x + 4.927 \text{ and } x = 0.197y - 0.548$$

There are lines of regression.

- ⑧ find the regression line of 'y' on 'x' for the following data.

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

$$\text{Soln: } \bar{x} = \frac{\sum x}{n} = \frac{56}{8} = 7, \bar{y} = \frac{\sum y}{n} = \frac{40}{8} = 5$$

we denote  $x = x - \bar{x}$  and  $y = y - \bar{y}$   
 $x = x - 7, y = y - 5$

we have lines of regression in the form

$$y = \frac{\sum xy}{\sum x^2} x \quad x = \frac{\sum xy}{\sum y^2} \cdot y$$

(13)

| x  | y | $x = x - \bar{x}$ | $y = y - \bar{y}$ | $x^2$ | $y^2$ | $xy$ |
|----|---|-------------------|-------------------|-------|-------|------|
| 1  | 1 | -6                | -4                | 36    | 16    | 24   |
| 3  | 2 | -4                | -3                | 16    | 9     | 12   |
| 4  | 4 | -2                | -1                | 4     | 1     | 2    |
| 6  | 4 | -1                | -1                | 1     | 1     | 1    |
| 8  | 5 | 1                 | 0                 | 1     | 0     | 0    |
| 9  | 7 | 2                 | 2                 | 4     | 4     | 4    |
| 11 | 8 | 4                 | 3                 | 16    | 9     | 12   |
| 14 | 9 | 7                 | 4                 | 49    | 16    | 28   |

$$\sum x = 56 \quad \sum y = 40$$

$$\sum x^2 = 138 \quad \sum y^2 = 56 \quad \sum xy = 84$$

i.e  $y - \bar{y} = \frac{\sum xy}{\sum x^2} (x - \bar{x})$ ,  $x - \bar{x} = \frac{\sum xy}{\sum y^2} (y - \bar{y})$

$$y - 5 = \frac{84}{138} (x - 7) \quad x - 7 = \frac{84}{56} (y - 5)$$

$$y - 5 = 0.63 (x - 7) \quad x - 7 = 1.4285 (y - 5)$$

$$y - 5 = 0.63x - 4.47 \quad x - 7 = 1.4285y - 7.1425$$

$$y = 0.63x - 4.47 + 5 \quad x = 1.4285y - 7.1425 + 7$$

$$y = 0.63x + 0.53 \quad x = 1.4285y - 0.1425$$

These are the lines of regression.

\* ⑨ Calculate the Karl Pearson's coefficient

of correlation for 10 students who have obtained the following % of marks in mathematics & Electronics.

| Roll No              | 1  | 2  | 3  | 4-5 | 6  | 7  | 8  | 9  | 10 |
|----------------------|----|----|----|-----|----|----|----|----|----|
| Marks in mathematics | 78 | 36 | 98 | 25  | 75 | 82 | 90 | 62 | 65 |
| Marks in Electronics | 84 | 51 | 91 | 60  | 68 | 62 | 86 | 58 | 53 |

Soln:- we have  $\bar{x} = \frac{\sum x}{n} = \frac{650}{10} = 65$

$$\bar{y} = \frac{\sum y}{n} = \frac{660}{10} = 66$$

$$\text{Let } x = x - \bar{x} \quad \text{&} \quad y = y - \bar{y}$$

$$x = x - 65 \quad y = y - 66$$

| $x$ | $y$ | $x = x - \bar{x}$ | $y = y - \bar{y}$ | $x^2$ | $y^2$ | $xy$ |
|-----|-----|-------------------|-------------------|-------|-------|------|
| 78  | 84  | 13                | 18                | 169   | 324   | 234  |
| 36  | 51  | -29               | -15               | 841   | 225   | 435  |
| 98  | 91  | 33                | 25                | 1089  | 625   | 825  |
| 85  | 60  | -40               | -6                | 1600  | 36    | 240  |
| 75  | 68  | 10                | 2                 | 100   | 4     | 20   |
| 82  | 62  | 17                | -4                | 289   | 16    | 68   |
| 90  | 86  | 25                | 20                | 625   | 400   | 500  |
| 62  | 58  | -3                | -8                | 9     | 64    | 24   |
| 65  | 53  | 0                 | -13               | 0     | 169   | 0    |
| 39  | 47  | -26               | -19               | 676   | 361   | 494  |

$$\sum x^2 = 5398 \quad \sum y^2 = 2224 \quad \sum xy = 2840$$

$$\gamma = \frac{\sum xy}{\sqrt{\sum x^2} \sqrt{\sum y^2}} = \frac{2840}{\sqrt{5398} \sqrt{2224}} = \frac{2840}{\sqrt{73.47} \times \sqrt{47.15}} = \frac{2840}{3464.1105}$$

$$\gamma = 0.819 \approx 0.8$$

- \* ⑩ In a partially destroyed lab record, only the lines of regression of  $y$  on  $x$  &  $x$  on  $y$  are available as  $4x - 5y + 33 = 0$  &  $20x - 9y = 107$  respectively. calculate  $\bar{x}, \bar{y}$  and coefficient of correlation b/w  $x - \epsilon_i y$ .

Soln: we know that regression lines parallel through  $\bar{x}$  &  $\bar{y}$ :

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

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

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

we shall now rewrite the eqn of the regression lines to find the regression coefficients.

$$5y = 4x + 33 \quad \text{or} \quad y = 0.8x + 6.6 \rightarrow ①$$

$$20x = 9y + 107 \quad \text{or} \quad x = 0.45y + 5.35 \rightarrow ②$$

From ① & ②

$$\tau \cdot \frac{\bar{y}}{\bar{x}} = 0.8 , \quad \tau \cdot \frac{\bar{x}}{\bar{y}} = 0.45$$

correlation coefficient  $\tau = \sqrt{0.8 \times 0.45} = \pm 0.6$

Thus  $\tau = 0.6$

\* Show that if  $\theta$  is the angle b/w the lines of regression, then

$$\tan \theta = \frac{\bar{y} \cdot \bar{x}}{\bar{x}^2 + \bar{y}^2} \left( \frac{1 - \tau^2}{\tau} \right)$$

Soln:- W.K.T if  $\theta$  is a w/e, the angle b/w the lines  $y = m_1 x + c_1$  and  $y = m_2 x + c_2$  is given by.

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

we have the lines of regression,

$$y - \bar{y} = \tau \frac{\bar{y}}{\bar{x}} (x - \bar{x}) \quad \text{①} \quad x - \bar{x} = \tau \frac{\bar{x}}{\bar{y}} (y - \bar{y})$$

we write the second of the eqn as

$$y - \bar{y} = \frac{\bar{y}}{\tau \bar{x}} (x - \bar{x}) \rightarrow \text{②}$$

Slopes of ① & ② are respectively given by

$$m_1 = \tau \frac{\bar{y}}{\bar{x}} \quad \text{&} \quad m_2 = \frac{\bar{y}}{\tau \bar{x}}$$

Substituting these in the formula for  $\tan \theta$   
we have,

$$\tan \theta = \frac{\frac{\bar{y}}{\tau \bar{x}} - \frac{\bar{y}}{\bar{x}}}{1 + \frac{\bar{y}}{\tau \bar{x}} \cdot \frac{\bar{y}}{\bar{x}}} = \frac{\frac{\bar{y}}{\bar{x}} \left( \frac{1}{\tau} - 1 \right)}{1 + \frac{\bar{y}^2}{\bar{x}^2}}$$

$$\text{Thus } \tan \theta = \frac{\frac{\bar{y}}{\bar{x}} \left( \frac{1 - \tau^2}{\tau} \right)}{\frac{\bar{x}^2 + \bar{y}^2}{\bar{x}^2}} = \frac{\bar{x} \bar{y}}{\bar{x}^2 + \bar{y}^2} \left( \frac{1 - \tau^2}{\tau} \right)$$

# RANK CORRELATION

## 18MAT41

### MODULE-IV

#### Spearman's rank Correlation :

The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the rank variables and it will be denoted by the symbol  $\rho$  ( Row).

Suppose  $x$  and  $y$  are the marks scored in the two subjects , let  $R_x$  and  $R_y$  are the ranks,  $d_i = R_x - R_y$  be the differences of the ranks then the Spearman's rank correlation will be calculated by the following formula.

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n^3 - n}$$

1. The scores for 9 students in Physics and Mathematics are as follows

|             |    |    |    |    |    |    |    |   |    |
|-------------|----|----|----|----|----|----|----|---|----|
| Physics     | 35 | 23 | 47 | 17 | 10 | 43 | 9  | 6 | 28 |
| Mathematics | 30 | 33 | 45 | 23 | 8  | 49 | 12 | 4 | 31 |

Compute the ranks of the students in the two subjects and also compute the Spearman's rank correlation.

**Sol.**

| Phy( $x$ ) | Mat( $y$ ) | $R_x$ | $R_y$ | $d_i = R_x - R_y$ | $d_i^2$           |
|------------|------------|-------|-------|-------------------|-------------------|
| 35         | 30         | 3     | 5     | -2                | 4                 |
| 23         | 33         | 5     | 3     | 2                 | 4                 |
| 47         | 45         | 1     | 2     | -1                | 1                 |
| 17         | 23         | 6     | 6     | 0                 | 0                 |
| 10         | 8          | 7     | 8     | -1                | 1                 |
| 43         | 49         | 2     | 1     | 1                 | 1                 |
| 9          | 12         | 8     | 7     | 1                 | 1                 |
| 6          | 4          | 9     | 9     | 0                 | 0                 |
| 28         | 31         | 4     | 4     | 0                 | 0                 |
|            |            |       |       |                   | $\sum d_i^2 = 12$ |

and  $n = 9$

We know that The Spearman correlation coefficient is

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n^3 - n}$$

$$\Rightarrow \rho = 1 - \frac{6(12)}{9^3 - 9}$$

$$\Rightarrow \rho = 1 - \frac{72}{720}$$

$$\Rightarrow \rho = 1 - 0.1$$

$$\Rightarrow \rho = 0.9$$

2. The scores for 9 students in English and Mathematics are as follows

|             |    |    |    |    |    |    |    |    |    |    |
|-------------|----|----|----|----|----|----|----|----|----|----|
| English     | 56 | 75 | 45 | 71 | 62 | 64 | 58 | 80 | 76 | 61 |
| Mathematics | 66 | 70 | 40 | 60 | 65 | 56 | 59 | 77 | 67 | 63 |

Compute the ranks of the students in the two subjects and also compute the Spearman's rank correlation.

Sol.

| Eng(x) | Mat(y) | $R_x$ | $R_y$ | $d_i = R_x - R_y$ | $d_i^2$           |
|--------|--------|-------|-------|-------------------|-------------------|
| 56     | 66     | 9     | 4     | 5                 | 25                |
| 75     | 70     | 3     | 2     | 1                 | 1                 |
| 45     | 40     | 10    | 10    | 0                 | 0                 |
| 71     | 60     | 4     | 7     | -3                | 9                 |
| 62     | 65     | 6     | 5     | 1                 | 1                 |
| 64     | 56     | 5     | 9     | -4                | 16                |
| 58     | 59     | 8     | 8     | 0                 | 0                 |
| 80     | 77     | 1     | 1     | 0                 | 0                 |
| 76     | 67     | 2     | 3     | 1                 | 1                 |
| 61     | 63     | 7     | 6     | 1                 | 1                 |
|        |        |       |       |                   | $\sum d_i^2 = 54$ |

and  $n=10$

We know that The Spearman correlation coefficient is

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n^3 - n}$$

$$\begin{aligned}\Rightarrow \rho &= 1 - \frac{6(54)}{10^3 - 10} \\ \Rightarrow \rho &= 1 - \frac{324}{990} \\ \Rightarrow \rho &= 1 - 0.33 \\ \Rightarrow \rho &= 0.67\end{aligned}$$

3. A random sample of recent repair jobs was selected and Estimated cost and Actual cost were recorded.

|                   |     |     |     |     |     |     |     |     |
|-------------------|-----|-----|-----|-----|-----|-----|-----|-----|
| Estimated Cost(x) | 300 | 450 | 800 | 250 | 500 | 975 | 475 | 400 |
| Actual Cost(y)    | 273 | 486 | 734 | 297 | 631 | 872 | 396 | 457 |

Compute the ranks and the Spearman's rank correlation.

Sol.

| Estimated Cost(x) | Actual Cost(y) | $R_x$ | $R_y$ | $d_i = R_x - R_y$ | $d_i^2$          |
|-------------------|----------------|-------|-------|-------------------|------------------|
| 300               | 273            | 7     | 8     | -1                | 1                |
| 450               | 486            | 5     | 4     | 1                 | 1                |
| 800               | 734            | 2     | 2     | 0                 | 0                |
| 250               | 297            | 8     | 7     | 1                 | 1                |
| 500               | 631            | 3     | 3     | 0                 | 0                |
| 975               | 872            | 1     | 1     | 0                 | 0                |
| 475               | 396            | 4     | 6     | -2                | 4                |
| 400               | 457            | 6     | 5     | 1                 | 1                |
|                   |                |       |       |                   | $\sum d_i^2 = 8$ |

and  $n = 8$

We know that The Spearman correlation coefficient is

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n^3 - n}$$

$$\begin{aligned}\Rightarrow \rho &= 1 - \frac{6(8)}{8^3 - 8} \\ \Rightarrow \rho &= 1 - \frac{48}{504} \\ \Rightarrow \rho &= 1 - 0.095 \\ \Rightarrow \rho &= 0.9047\end{aligned}$$

4. The rank of 10 students of same batch in two subjects A and B are given below. Calculate the rank correlation coefficient.

| Rank of A | 1 | 2 | 3 | 4  | 5 | 6 | 7 | 8 | 9 | 10 |
|-----------|---|---|---|----|---|---|---|---|---|----|
| Rank of B | 6 | 7 | 5 | 10 | 3 | 9 | 4 | 1 | 8 | 2  |

Sol.

Given

The ranks of no. of students  $n = 10$

| $R_A$ | $R_B$ | $d_i = R_A - R_B$ | $d_i^2$            |
|-------|-------|-------------------|--------------------|
| 1     | 6     | -5                | 25                 |
| 2     | 7     | -5                | 25                 |
| 3     | 5     | -2                | 4                  |
| 4     | 10    | -6                | 36                 |
| 5     | 3     | 2                 | 4                  |
| 6     | 9     | -3                | 9                  |
| 7     | 4     | 3                 | 9                  |
| 8     | 1     | 7                 | 49                 |
| 9     | 8     | 1                 | 1                  |
| 10    | 2     | 8                 | 64                 |
|       |       |                   | $\sum d_i^2 = 226$ |

We know that The Spearman correlation coefficient is

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n^3 - n}$$

$$\begin{aligned}\Rightarrow \rho &= 1 - \frac{6(226)}{10^3 - 10} \\ \Rightarrow \rho &= 1 - \frac{1356}{990} \\ \Rightarrow \rho &= 1 - 1.37 \\ \Rightarrow \rho &= -0.37\end{aligned}$$

5. The participants in a contest are ranked by two judges as follows.

|      |   |   |   |    |   |   |   |    |   |   |
|------|---|---|---|----|---|---|---|----|---|---|
| $x:$ | 1 | 6 | 5 | 10 | 3 | 2 | 4 | 9  | 7 | 8 |
| $y:$ | 6 | 4 | 9 | 8  | 1 | 2 | 3 | 10 | 5 | 7 |

Compute the Spearman's rank correlation.

Sol.

Given

The ranks of no. of students  $n = 10$

| $R_x$ | $R_y$ | $d_i = R_x - R_y$ | $d_i^2$           |
|-------|-------|-------------------|-------------------|
| 1     | 6     | -5                | 25                |
| 6     | 4     | 2                 | 4                 |
| 5     | 9     | -4                | 16                |
| 10    | 8     | 2                 | 4                 |
| 3     | 1     | 2                 | 4                 |
| 2     | 2     | 0                 | 0                 |
| 4     | 3     | 1                 | 1                 |
| 9     | 10    | -1                | 1                 |
| 7     | 5     | 2                 | 4                 |
| 8     | 7     | 1                 | 1                 |
|       |       |                   | $\sum d_i^2 = 60$ |

We know that The Spearman correlation coefficient is

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n^3 - n}$$

$$\begin{aligned}\Rightarrow \rho &= 1 - \frac{6(60)}{10^3 - 10} \\ \Rightarrow \rho &= 1 - \frac{360}{990} \\ \Rightarrow \rho &= 1 - 0.3636 \\ \Rightarrow \rho &= 0.6363\end{aligned}$$

6. The participants in a contest are ranked by two judges as follows.

|      |   |   |   |   |   |   |    |   |    |   |
|------|---|---|---|---|---|---|----|---|----|---|
| $x:$ | 6 | 4 | 3 | 1 | 2 | 7 | 9  | 8 | 10 | 5 |
| $y:$ | 4 | 1 | 6 | 7 | 5 | 8 | 10 | 9 | 3  | 2 |

Compute the Spearman's rank correlation.

Sol.

Given

The ranks of no. of students  $n=10$

| $R_x$ | $R_y$ | $d_i = R_x - R_y$ | $d_i^2$            |
|-------|-------|-------------------|--------------------|
| 6     | 4     | 2                 | 4                  |
| 4     | 1     | 3                 | 9                  |
| 3     | 6     | -3                | 9                  |
| 1     | 7     | -6                | 36                 |
| 2     | 5     | -3                | 9                  |
| 7     | 8     | -1                | 1                  |
| 9     | 10    | -1                | 1                  |
| 8     | 9     | -1                | 1                  |
| 10    | 3     | 7                 | 49                 |
| 5     | 2     | 3                 | 9                  |
|       |       |                   | $\sum d_i^2 = 128$ |

We know that The Spearman correlation coefficient is

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n^3 - n}$$

$$\begin{aligned}\Rightarrow \rho &= 1 - \frac{6(128)}{10^3 - 10} \\ \Rightarrow \rho &= 1 - \frac{768}{990} \\ \Rightarrow \rho &= 1 - 0.7758 \\ \Rightarrow \rho &= 0.2242\end{aligned}$$

\*\*\*