# Module - 5

## Bivariate Data

Data based on the characteristics of two variables are called bivariate data.

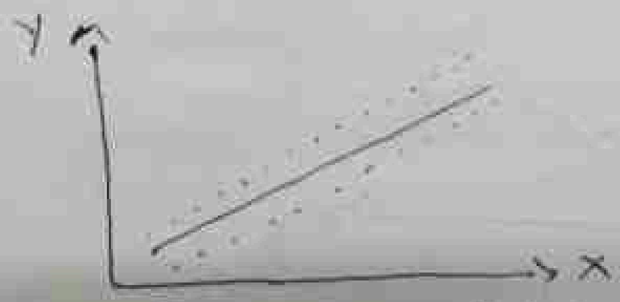
## Scatter Diagram

The scatter diagram is a technique used to examine the relationship b/w both the axis (x & y) with one variable. Each pair of values is plotted on the graph by means of a dot mark. If these points show same trend either upward or downward, these two variables are said to be correlated. If the plotted points do not show any trend, these two variables are not correlated.

Interpretation of the nature & degree of relation using scatter diagram

1. Positive linear Correlation

In the graph, the plotted points are distributed from lower left corner to upper right corner, the correlation is said to be positive.




Positively Correlated

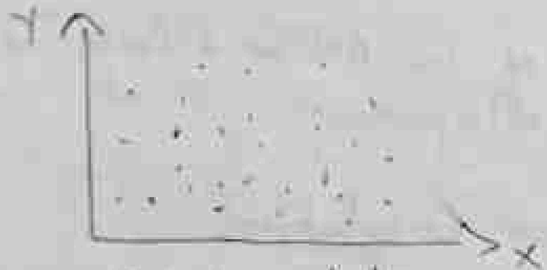
## 2. Negative Correlation

In the graph, the plotted points are distributed from upper left corner to lower right corner, the correlation is said to be negative.


Negatively Correlated

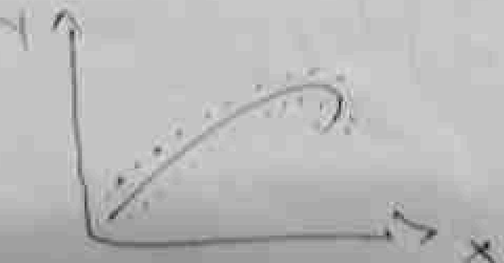
## 3. No Correlation

The points are scattered over the graph & do not show any specific pattern, then there is no correlation b/w two variables




No correlation

## 4. Non-linear Correlation



# Curve fitting - Principle of least Squares

Let $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ be $n$ pairs of observations of related data the general problem of finding a relation of the form $y = f(x)$ which fits best to the given data is called curve fitting

## Principle of least Squares

Let $y = f(x)$ be a best fit curve for the data $(x_1, y_1) (x_2, y_2) \ldots (x_n, y_n)$. $y_i$ is called the observed value of $y$ corresponding to $x_i$ & $f(x_i)$ is called the expected value of $y$ corresponding to $x_i$

Let $E_i = y_i - f(x_i)$ $i = 1, 2, \ldots, n$, where $E_i$ is called the error or residual for $y_i$ the principle of least squares states that for a best fit curve, the sum of squares of residuals is a minimum

i.e. $E = \sum_{i=1}^{n} [y_i - f(x_i)]^2$ is a minimum

fitting a straight line

$$y = A + Bx$$

The normal equations are

$$\Sigma y = nA + B\Sigma x$$
$$\Sigma xy = A\Sigma x + B\Sigma x^2$$

By solving these two eqns we can find A & B

1. Use the principle of least squares to fit a straight line to the following data.

i)

| x | 5 | 10 | 15 | 20 | 25 |
|---|---|---|---|---|---|
| y | 15 | 19 | 23 | 26 | 30 |

The normal equations are

$$\Sigma y = nA + B\Sigma x$$
$$\Sigma xy = A\Sigma x + B\Sigma x^2$$

| $x$ | $y$ | $x^2$ | $xy$ |
|---|---|---|---|
| 5 | 15 | 25 | 75 |
| 10 | 19 | 100 | 190 |
| 15 | 23 | 225 | 345 |
| 20 | 26 | 400 | 520 |
| 25 | 30 | 625 | 750 |
| 75 | 113 | 1375 | 1880 |

2

$\therefore$ Normal eqns become

$$112 = 5A + 75B \quad \text{--- (1)}$$
$$1880 = 75A + 1375B \quad \text{--- (2)}$$

By solving, multiply 15 on (1)

$$1695 = 75A + 1125B$$
$$1880 = 75A + 1375B$$

---

$$-185 = -250B$$
$$B = 1.35$$

Put $B = 1.35$ in (1)

$$A = 11.75$$

$\therefore$ The straight line is $y = 11.75 + 1.35x$

2.

| $x$ | 1 | 3 | 5 | 7 | 8 | 10 |
|---|---|---|---|---|---|---|
| $y$ | 8 | 12 | 15 | 17 | 18 | 20 |

The normal eqns are

$$\Sigma y = nA + B\Sigma x$$
$$\Sigma xy = A\Sigma x + B\Sigma x^2$$

| $x$ | $y$ | $x^2$ | $xy$ |
|---|---|---|---|
| 1 | 8 | 1 | 8 |
| 3 | 12 | 9 | 36 |
| 5 | 15 | 25 | 75 |
| 7 | 17 | 49 | 119 |
| 8 | 18 | 64 | 144 |
| 10 | 20 | 100 | 200 |
| 34 | 90 | 248 | 582 |

Normal Eqns become

$$90 = 6A + 34B$$

$$582 = 34A + 248B$$

By Solving, $A = 7.63$ & $B = 1.3$

$$\therefore y = 7.63 + 1.3x$$

3. fit a st. line of the form $y = ax + b$ for

| | | | | |
|---|---|---|---|---|
| $x$: | 2 | 3 | 4 | 5 |
| $y$: | 7 | 9 | 10 | 11 |

Normal eqns are

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

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

| $x$ | $y$ | $x^2$ | $xy$ |
|---|---|---|---|
| 2 | 7 | 4 | 14 |
| 3 | 9 | 9 | 27 |
| 4 | 10 | 16 | 40 |
| 5 | 11 | 25 | 55 |
| 14 | 37 | 54 | 136 |

$\therefore$ Normal eqns are

$$37 = 14a + 4b$$

$$136 = 54a + 14b$$

By Solving, $a = 1.3$ & $b = 4.7$

$$y = 1.3x + 4.7$$

4. $x$ : 0 1 2 3 4

$y$ : 1 1.8 3.3 4.5 6.3

Normal eqns of $y = a + bx$ are

$$\Sigma y = na + b\Sigma x$$

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

| $x$ | $y$ | $x^2$ | $xy$ |
|---|---|---|---|
| 0 | 1 | 0 | 0 |
| 1 | 1.8 | 1 | 1.8 |
| 2 | 3.3 | 4 | 6.6 |
| 3 | 4.5 | 9 | 13.5 |
| 4 | 6.3 | 16 | 25.2 |
| 10 | 16.9 | 30 | 47.1 |

Solving $16.9 = 5a + 10b$

$47.1 = 10a + 30b$

Solving, $a = 0.72$ & $b = 1.33$

$$\therefore y = 0.72 + 1.33x$$

5. $x$ : 0 5 10 15 20

$y$ : 7 11 16 20 26

Normal eqns of $y = a + bx$ are

$$\Sigma y = nA + b\Sigma x$$

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

| $x$ | $y$ | $x^2$ | $xy$ |
|---|---|---|---|
| 0 | 7 | 0 | 0 |
| 5 | 11 | 25 | 55 |
| 10 | 16 | 100 | 160 |
| 15 | 20 | 225 | 300 |
| 20 | 26 | 400 | 520 |
| 50 | 80 | 750 | 1035 |

$$80 = 5a + 50b$$

$$1035 = 50a + 750b$$

Solving, $a = 6.6 \quad b = 0.94$

$$\therefore y = 6.6 + 0.94x$$

5. $x$ : 1941 1951 1961 1971 1981

$y$ : 8 10 12 10 16

Normal eqn of $y = ax + b$ are

$$\Sigma y = nA + b\Sigma x$$

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

| $x$ | $y$ | $x^2$ | $xy$ |
|---|---|---|---|
| 1941 | 8 | 3,767,481 | 15,552 |
| 1951 | 10 | 3,806,401 | 19,510 |
| 1961 | 12 | 3,845,521 | 23,532 |
| 1971 | 10 | 3,884,841 | 19,710 |
| 1981 | 16 | 3,924,361 | 31,696 |
| 9805 | 56 | 19,228,605 | 110,000 |

$$56 = 5a + 9805\,b$$

$$110{,}000 = 9805a + 19{,}228{,}605\,b$$

By solving $a = -302.56$ $b = 0.16$

$$\Rightarrow y = 0.16x - 302.56$$

fitting a Parabola

$$y = a + bx + cx^2$$

Normal equations are

$$\Sigma y = nA + b\Sigma x + c\Sigma x^2$$

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

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

fit a parabola by the method of least squares to the following data.

| | | | | | |
|---|---|---|---|---|---|
| $x$ : | 1 | 2 | 3 | 4 | 5 |
| $y$ : | 5 | 12 | 26 | 60 | 97 |

let $y = a + bx + cx^2$

Normal eqns are

$$\Sigma y = na + b\Sigma x + c\Sigma x^2$$

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

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

| $x$ | $y$ | $x^2$ | $x^3$ | $x^4$ | $xy$ | $x^2y$ |
|---|---|---|---|---|---|---|
| 1 | 5 | 1 | 1 | 1 | 5 | 5 |
| 2 | 12 | 4 | 8 | 16 | 24 | 48 |
| 3 | 26 | 9 | 27 | 81 | 78 | 234 |
| 4 | 60 | 16 | 64 | 256 | 240 | 960 |
| 5 | 97 | 25 | 125 | 625 | 485 | 2425 |
| 15 | 200 | 55 | 225 | 979 | 832 | 3672 |

$\therefore$ normal eqns are

$$200 = 5a + 15b + 55c$$

$$832 = 15a + 55b + 225c$$

$$3672 = 55a + 225b + 979c$$

By Solving $a = 10.3$, $b = -11$, $c = 5.7$

$$\therefore y = 10.3 - 11x + 5.7x^2$$

Fit a parabola of the type $y = a + bx + cx^2$ to the following data

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $x$: | 10 | 15 | 20 | 25 | 30 | 35 | 40 |
| $y$: | 11 | 13 | 16 | 20 | 27 | 34 | 41 |

normal eqns are

$$\Sigma y = na + b\Sigma x + c\Sigma x^2$$

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

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

| $x$ | $y$ | $x^2$ | $x^3$ | $x^4$ | $xy$ | $x^2y$ |
|---|---|---|---|---|---|---|
| 10 | 11 | 100 | 1000 | 10000 | 110 | 1100 |
| 15 | 13 | 225 | 3375 | 50,625 | 195 | 2925 |
| 20 | 16 | 400 | 8000 | 160000 | 320 | 6400 |
| 25 | 20 | 625 | 15625 | 3,90,625 | 500 | 12500 |
| 30 | 27 | 900 | 27000 | 8,10,000 | 810 | 24300 |
| 35 | 34 | 1225 | 42875 | 1500625 | 1190 | 41,650 |
| 40 | 41 | 1600 | 64000 | 2560000 | 1640 | 65,600 |
| 175 | 162 | 5075 | 161875 | 5481875 | 4765 | 154475 |

$\therefore$ normal eqns are

$$162 = 7a + 175b + 5075c$$

$$4765 = 175a + 5075b + 161875c$$

$$154475 = 5075a + 161875b + 5481875c$$

by solving, $a = 10$, $b = -2$, $c = 2.4$

$$\therefore y = 10 - 2x + 2.4x^2$$

fit the parabola $y = ax^2 + bx + c$ for the following data by the method of least squares. Estimate the value of $y$ when $x = 10$

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

normal eqns are

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

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

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

| $x$ | $y$ | $x^2$ | $x^3$ | $x^4$ | $xy$ | $x^2y$ |
|---|---|---|---|---|---|---|
| 1 | 2 | 1 | 1 | 1 | 2 | 2 |
| 2 | 6 | 4 | 8 | 16 | 12 | 24 |
| 3 | 7 | 9 | 27 | 81 | 21 | 63 |
| 4 | 8 | 16 | 64 | 256 | 32 | 128 |
| 5 | 10 | 25 | 125 | 625 | 50 | 250 |
| 6 | 11 | 36 | 216 | 1296 | 66 | 396 |
| 7 | 11 | 49 | 343 | 2401 | 77 | 539 |
| 8 | 10 | 64 | 512 | 4096 | 80 | 640 |
| 9 | 9 | 81 | 729 | 6561 | 81 | 729 |
| 45 | 74 | 285 | 2025 | 15,333 | 421 | 2771 |

$\therefore$ Eqns are

$$74 = 285a + 45b + 9c$$

$$421 = 2025a + 285b + 45c$$

$$2771 = 15333a + 2025b + 285c$$

by solving $a = -0.267$, $b = 3.51$, $c = -0.845$

$$\therefore y = -0.267x^2 + \underline{3.5}x - 0.845$$

fit a parabola of the form $y = a + bx + cx^2$ for the following data

$x$: 0 1 2 3 4

$y$: 1 5 10 22 38

normal eqns are

$$\Sigma y = na + b\Sigma x + c\Sigma x^2$$

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

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

| $x$ | $y$ | $x^2$ | $x^3$ | $x^4$ | $xy$ | $x^2y$ |
|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 1 | 5 | 1 | 1 | 1 | 5 | 5 |
| 2 | 10 | 4 | 8 | 16 | 20 | 40 |
| 3 | 22 | 9 | 27 | 81 | 66 | 198 |
| 4 | 38 | 16 | 64 | 256 | 152 | 608 |
| 10 | 76 | 30 | 100 | 354 | 243 | 851 |

$$76 = 5a + 10b + 30c$$

$$243 = 10a + 30b + 100c$$

$$851 = 30a + 100b + 354c$$

by solving $a = 1.42$, $b = 0.3$, $c = 2.2$

$$\therefore y = 1.42 + 0.3x + 2.2x^2$$

fit a parabola $y = ax^2 + bx + c$ for the following data

| | | | | | | |
|---|---|---|---|---|---|---|
| $x$: | 1 | 2 | 3 | 4 | 5 | 6 |
| $y$: | 5 | 14 | 30 | 44 | 77 | 96 |

normal eqns are

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

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

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

| $x$ | $y$ | $x^2$ | $x^3$ | $x^4$ | $xy$ | $x^2y$ |
|---|---|---|---|---|---|---|
| 1 | 5 | 1 | 1 | 1 | 5 | 5 |
| 2 | 14 | 4 | 8 | 16 | 28 | 56 |
| 3 | 30 | 9 | 27 | 81 | 90 | 270 |
| 4 | 44 | 16 | 64 | 256 | 176 | 704 |
| 5 | 77 | 25 | 125 | 625 | 385 | 1925 |
| 6 | 98 | 36 | 216 | 1296 | 588 | 3528 |
| 21 | 268 | 91 | 441 | 2275 | 1272 | 6488 |

$$268 = 91a + 21b + 6c$$

$$1272 = 441a + 91b + 21c$$

$$6488 = 2275a + 441b + 91c$$

by Solving $a = 2.371$, $b = 2.48$, $c = 24.96$

$$\Rightarrow y = 2.371x^2 + 2.48x + 24.96$$

Linear Correlation & Regression - karl Pearson's Coefficient of Correlation

The Degree of relationship b/w the variables is called Coefficient of Correlation. It is usually denoted by $r_{xy}$ or $r$ or $r_{xy}$

$$r_{xy} = \frac{\frac{\Sigma xy}{n} - \bar{x}\bar{y}}{\sqrt{\frac{\Sigma x^2}{n} - \bar{x}^2} \cdot \sqrt{\frac{\Sigma y^2}{n} - \bar{y}^2}}$$

where $\bar{x} = \frac{\Sigma x}{n}$ & $\bar{y} = \frac{\Sigma y}{n}$

Properties

* Coefficient of correlation lies b/w -1 & 1
* If $r=0$, then the two variables are uncorrelated
* The Coefficient of correlation is independent of the change of origin & scale of measurements.

Problems

find the correlation & Coefficient for the following data:

$x$ : 3 5 6 7 10 11
$y$ : 8 12 11 14 16 17

| $x$ | $y$ | $x^2$ | $y^2$ | $xy$ |
|---|---|---|---|---|
| 3 | 8 | 9 | 64 | 24 |
| 5 | 12 | 25 | 144 | 60 |
| 6 | 11 | 36 | 121 | 66 |
| 7 | 14 | 49 | 196 | 98 |
| 10 | 16 | 100 | 256 | 160 |
| 11 | 17 | 121 | 289 | 187 |
| 42 | 78 | 340 | 1070 | 595 |

$\bar{x} = \frac{\Sigma x}{n} = \frac{42}{6} = \underline{7}$

$\bar{y} = \frac{\Sigma y}{n} = \frac{78}{6} = \underline{13}$

$r = \frac{\Sigma xy/n - \bar{x}\bar{y}}{\sqrt{\frac{\Sigma x^2}{n} - \bar{x}^2} \cdot \sqrt{\frac{\Sigma y^2}{n} - \bar{y}^2}} = \frac{545/6 - 7 \times 13}{\sqrt{\frac{3440}{6} - 7^2} \cdot \sqrt{\frac{1070}{6} - 13^2}}$

$= \underline{0.97}$

x : 51 54 55 59 65 60 70

y : 38 44 33 36 33 23 10

let $u = x - 60$ $v = y - 33$

| x | y | u | v | $u^2$ | $v^2$ | uv |
|---|---|---|---|---|---|---|
| 51 | 38 | -9 | 5 | 81 | 25 | -45 |
| 54 | 44 | -6 | 11 | 36 | 121 | -66 |
| 55 | 33 | -5 | 0 | 25 | 0 | 0 |
| 59 | 36 | -1 | 3 | 1 | 19 | -3 |
| 65 | 33 | 5 | 0 | 25 | 0 | 0 |
| 60 | 23 | 0 | -10 | 0 | 100 | 0 |
| 70 | 10 | 10 | -23 | 100 | 529 | -230 |
| | | -6 | -14 | 268 | 784 | -344 |

$\bar{u} = \frac{\Sigma u}{n} = \frac{-6}{7} = -0.86$

$\bar{v} = \frac{\Sigma v}{n} = \frac{-14}{7} = -2$

$$r = \frac{\frac{\sum uv}{n} - \bar{u}\bar{v}}{\sqrt{\frac{\sum u^2}{n} - \bar{u}^2}\sqrt{\frac{\sum v^2}{n} - \bar{v}^2}}$$

$$= \frac{\frac{-344}{7} - 0.86 \times 2}{\sqrt{\frac{268}{7} - (-0.86)^2} \cdot \sqrt{\frac{784}{7} - (-2)^2}}$$

$$= \underline{-0.798}$$

3. $x$ : 9 8 7 6 5 4 3 2

$y$ : 15 16 14 13 11 12 10 8

| $x$ | $y$ | $x^2$ | $y^2$ | $xy$ |
|---|---|---|---|---|
| 9 | 15 | 81 | 225 | 135 |
| 8 | 16 | 64 | 256 | 128 |
| 7 | 14 | 49 | 196 | 98 |
| 6 | 13 | 36 | 169 | 78 |
| 5 | 11 | 25 | 121 | 55 |
| 4 | 12 | 16 | 144 | 48 |
| 3 | 10 | 9 | 100 | 30 |
| 2 | 8 | 4 | 64 | 16 |
| 1 | 9 | 1 | 81 | 9 |
| 45 | 108 | 285 | 1356 | 597 |

$$r = \frac{\frac{\sum xy}{n} - \bar{x}\bar{y}}{\sqrt{\frac{\sum x^2}{n} - \bar{x}^2} \cdot \sqrt{\frac{\sum y^2}{n} - \bar{y}^2}}$$

$\bar{x} = \frac{\sum x}{n} = \frac{45}{9}$ , $\bar{y} = \frac{\sum y}{n} = \frac{108}{9}$

$$r = \frac{547/9 - 45/9 \times \frac{108}{9}}{\sqrt{\frac{285}{9} - \left(\frac{45}{9}\right)^2} \cdot \sqrt{\frac{1356}{9} - \left(\frac{108}{9}\right)^2}}$$

$= \underline{\underline{0.95}}$

4. x: 77 54 27 52 14 35 90 25 96 60

y: 35 58 60 40 50 40 35 56 34 42

| x | y |
|---|---|
| 77 | 35 |
| 54 | 58 |
| 27 | 60 |
| 52 | 40 |
| 14 | 50 |
| 35 | 40 |
| 90 | 35 |
| 25 | 56 |
| 96 | 34 |
| 60 | 42 |

0
2

5. $x$ : 6 2 10 4 8
$y$ : 9 11 5 8 7

| $x$ | $y$ | $x^2$ | $y^2$ | $xy$ |
|---|---|---|---|---|
| 6 | 9 | 36 | 81 | 54 |
| 2 | 11 | 4 | 121 | 22 |
| 10 | 5 | 100 | 25 | 50 |
| 4 | 8 | 16 | 64 | 32 |
| 8 | 7 | 64 | 49 | 56 |
| 30 | 40 | 220 | 340 | 214 |

$$r_{xy} = \frac{\frac{\Sigma xy}{n} - \bar{x}\bar{y}}{\sqrt{\frac{\Sigma x^2}{n} - \bar{x}^2} \cdot \sqrt{\frac{\Sigma y^2}{n} - \bar{y}^2}}$$

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

$\bar{x} = \frac{30}{5} = 6$ , $\bar{y} = \frac{40}{5} = 8$

$$= \frac{\frac{214}{5} - 6 \times 8}{\sqrt{\frac{220}{5} - 6^2} \cdot \sqrt{\frac{340}{5} - 8^2}}$$

$= -0.92$

Specsmen's rank correlation Coefficient

Instead of giving the values of the variables, the ranks of the observations are given, the correlation coefficient so obtained is called rank correlation coefficient. It is denoted by $\rho$ or $\rho_{xy}$

i.e $\rho = 1 - \frac{6 \Sigma D^2}{n(n^2-1)}$ where $D = x-y$
(difference b/w ranks)

If there are ties in ranks, we substitute each of the tied observations the mean of the ranks that they jointly occupy if there are two ranks equal to 4, they are given the mean rank $= \frac{4+5}{2} = 4.5$

∴ The formula becomes,

$$\rho = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12} m_1(m_1^2-1) + \frac{1}{12} m_2(m_2^2-1) + \cdots \frac{1}{12} m_k(m_k^2-1)\right]}{n(n^2-1)}$$

where $m_1, m_2 \cdots$ stands for the no of times the rank repeats

1. Calculate the rank corelation for

ranks in $x$ : 1 2 3 4

ranks in $y$ : 3 4 2 1

| $x$ | $y$ | $D = x - y$ | $D^2$ |
|---|---|---|---|
| 1 | 3 | -2 | 4 |
| 2 | 4 | -2 | 4 |
| 3 | 2 | 1 | 1 |
| 4 | 1 | 3 | 9 |
| | | | 18 |

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

$$= 1 - \frac{6 \times 18}{4(4^2-1)} = -0.8$$

2. In a music contest two judges awarded the following ranks to 10 competitors. On the basis of the data do you think that the judges appear to agree in their standard

judge 1 : 5 4 2 6 7 10 9 1 8 3

judge 2 : 4 1 5 7 8 9 7 6 3 2

| $x$ | $y$ | $D = x - y$ | $D^2$ |
|---|---|---|---|
| 5 | 4 | 1 | 1 |
| 4 | 1 | 3 | 9 |
| 2 | 5 | -3 | 9 |
| 6 | 7 | -1 | 1 |
| 7 | 8 | -1 | 1 |
| 10 | 9 | 1 | 1 |
| 9 | 7 | 2 | 4 |
| 1 | 6 | -5 | 25 |
| 8 | 3 | 5 | 25 |
| 3 | 2 | 1 | 1 |
| | | | 77 |

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

$$= 1 - \frac{6 \times 77}{10(10^2-1)} = \underline{\underline{0.53}}$$

3. Calculate the rank correlation coefficient for the following data on weight of fathers & sons

wg. of father : ~~57~~ 57 54 55 59 65 60 70

wg. of son : 38 44 33 36 34 23 10

| ag. of father | ag. of son | rank (x) | rank (y) | D | $D^2$ |
|---|---|---|---|---|---|
| 57 | 38 | 1 | 6 | -5 | 25 |
| 54 | 44 | 2 | 7 | -5 | 25 |
| 55 | 33 | 3 | 3 | 0 | 0 |
| 59 | 36 | 4 | 5 | -1 | 1 |
| 65 | 34 | 6 | 4 | -2 | 4 |
| 60 | 23 | 5 | 2 | 3 | 9 |
| 70 | 10 | 7 | 1 | 6 | 36 |
| | | | | | 100 |

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

$$= 1 - \frac{6\times 100}{7(7^2-1)} = -0.79$$

Calculate the rank correlation coefficient for the following data on heights of fathers & sons

ht. of father : 65 63 67 64 68 62 70 66 68 67

ht. of son : 68 66 68 65 67 66 68 65 71 67

| ht. of father | ht. of son | rank (x) | rank (y) | D | $D^2$ |
|---|---|---|---|---|---|
| 65 | 68 | 7 | 4 | 3 | 9 |
| 63 | 66 | 9 | 7.5 | 1.5 | 2.25 |
| 67 | 68 | 4.5 | 4 | 0.5 | 0.25 |
| 64 | 65 | 8 | 9.5 | -1.5 | 2.25 |
| 68 | 67 | 2.5 | 2 | 0.5 | 0.25 |
| 62 | 66 | 10 | 7.5 | 2.5 | 6.25 |
| 70 | 68 | 1 | 4 | -3 | 9 |
| 66 | 65 | 6 | 9.5 | -3.5 | 12.25 |
| 68 | 71 | 2.5 | 1 | 1.5 | 2.25 |
| 67 | 67 | 4.5 | 6 | -1.5 | 2.25 |
| | | | | | 46 |

$$\rho = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12} m_1(m_1^2-1) + \cdots\right]}{n(n^2-1)}$$

$$= 1 - \frac{6\left[46 + \frac{1}{12}\times 2(2^2-1) + \frac{1}{12}\times 2(2^2-1) + \frac{1}{12}\times 3(3^2-1) + \frac{1}{12}\times 2(2^2-1) + \frac{1}{12}\times 2(2^2-1)\right]}{10(10^2-1)}$$

$$= \underline{0.70}$$

5 Calculate the rank correlation coefficient for the following data

rank of x : 1 2 2 2 3 4 5 5 6

rank of y : 1 2 3 3 3 3 4 5 6

| rank (x) | rank (y) | $D = x-y$ | $D^2$ |
|---|---|---|---|
| 1 | 1 | 0 | 0 |
| 2 | 2 | 0 | 0 |
| 2 | 3 | -1 | 1 |
| 2 | 3 | -1 | 1 |
| 3 | 3 | 0 | 0 |
| 4 | 3 | 1 | 1 |
| 5 | 4 | 1 | 1 |
| 5 | 5 | 0 | 0 |
| 6 | 6 | 0 | 0 |
| 7 | 6 | 1 | 1 |
| | | | 5 |

$$\rho = 1 - \frac{6\left(\Sigma D^2 + \frac{1}{12}\times 3(3^2-1) + \frac{1}{12}\times 2(2^2-1) + \frac{1}{12}\times 4(4^2-1) + \frac{1}{12}\times 2(2^2-1)\right]}{n(n^2-1)}$$

$$= 1 - \frac{6[5+2+2+5+2]}{10(10^2-1)} = 0.897$$

rank of $x$ : 3 5 8 4 7 10 2 1 6 9

rank of $y$ : 6 4 9 8 1 2 3 10 5 7

| $x$ | $y$ | $D = x - y$ | $D^2$ |
|---|---|---|---|
| 3 | 6 | -3 | 9 |
| 5 | 4 | 1 | 1 |
| 8 | 9 | -1 | 1 |
| 4 | 8 | -4 | 16 |
| 7 | 1 | 6 | 36 |
| 10 | 2 | 8 | 64 |
| 2 | 3 | -1 | 1 |
| 1 | 10 | -9 | 81 |
| 6 | 5 | 1 | 1 |
| 9 | 7 | 2 | 4 |
| | | | 214 |

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

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

| x | y | rank(x) | rank(y) | D = x − y | $D^2$ |
|---|---|---|---|---|---|
| 50 | 11 | 2 | 1 | 1 | 1 |
| 50 | 13 | 2 | 3 | −1 | 1 |
| 50 | 14 | 4 | 5.5 | −1.5 | 2.25 |
| 58 | 14 | 6 | 9.5 | −3.5 | −12.25 |
| 60 | 16 | 9 | 9.5 | −0.5 | 0.25 |
| 65 | 16 | 9 | 7.5 | 1.5 | 2.25 |
| 65 | 16 | 9 | 7.5 | 1.5 | 2.25 |
| 65 | 15 | 6 | 5.5 | 0.5 | 0.25 |
| 60 | 14 | 6 | 3 | 3 | 9 |
| 60 | 13 | 2 | 3 | −1 | 1 |
| 50 | 13 | | | | 34.5 |

$$\rho = 1 - \frac{6\left[\sum D^2 + \frac{1}{12}m_1(m_1^2-1) + \dots\right]}{n(n^2-1)}$$

$$= 1 - \frac{6\left[34.5 + \frac{1}{12}3(3^2-1) + \frac{1}{12}3(3^2-1) + \frac{1}{12}3(3^2-1) + \frac{1}{12}3(3^2-1) + \frac{1}{12}2(2^2-1) + \frac{1}{12}2(2^2-1) + \frac{1}{12}2(2^2-1)\right]}{10(10^2-1)}$$

$$= \frac{1 - 6\left(\frac{79+3}{2}\right)}{10(10^2-1)} = \underline{\underline{0.752}}$$

## Regression

The eqn which gives the relationship b/w the variables is called regression eqn

### Regression lines

Regression line of $x$ on $y$ is

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

where $\sigma_{xy} = \frac{\Sigma xy}{n} - \bar{x}\bar{y}$

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

$\frac{\sigma_{xy}}{\sigma_y^2}$ is called the regression Coefficient of $x$ on $y$ & is denoted as $b_{xy}$ i.e $b_{xy} = \frac{\sigma_{xy}}{\sigma_y^2}$

Regression line of $y$ on $x$ is

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

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

$\frac{\sigma_{xy}}{\sigma_x^2}$ is called the regression Coeff$^t$ of $y$ on $x$ denoted by $b_{yx}$ i.e, $b_{yx} = \frac{\sigma_{xy}}{\sigma_x^2}$

**Note** - Correlation Coeff$^t$ is the geometric mean of the regression Coeff$^{ts}$

i.e, $r = \pm\sqrt{b_{xy} \times b_{yx}}$

Both $b_{xy}$ & $b_{yx}$ are $+$, $r$ also $+$
$-$, $r$ also $-$

1. from the following data obtain two regression eqns.

| $x$ | 6 | 2 | 10 | 4 | 8 |
|---|---|---|---|---|---|
| $y$ | 9 | 11 | 5 | 8 | 7 |

| $x$ | $y$ | $x^2$ | $y^2$ | $xy$ |
|---|---|---|---|---|
| 6 | 9 | 36 | 81 | 54 |
| 2 | 11 | 4 | 121 | 22 |
| 10 | 5 | 100 | 25 | 50 |
| 4 | 8 | 16 | 64 | 32 |
| 8 | 7 | 64 | 49 | 56 |
| 30 | 40 | 220 | 340 | 214 |

$$\bar{x} = \frac{\Sigma x}{n} = \frac{30}{5} = 6$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{40}{5} = 8$$

$$\sigma_{xy} = \frac{\Sigma xy}{n} - \bar{x}\bar{y} = \frac{214}{5} - 6 \times 8 = -5.2$$

$$\sigma x^2 = \frac{\Sigma x^2}{n} - \bar{x}^2 = \frac{220}{5} - 36 = 8$$

$$\sigma y^2 = \frac{\Sigma y^2}{n} - \bar{y}^2 = \frac{340}{5} - 64 = 4$$

Regression line of $x$ on $y$ is

2

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

$$x - 6 = \frac{-5.2}{4}(y - 8)$$

$$\text{or } x = \underline{16.4 - 1.3y}$$

Regression line of y on x is

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

$$y - 8 = \frac{-5.2}{8}(x - 6)$$

$$y = \underline{11.9 - 0.65x}$$

2. The lengths & weights of a Sample of 6 articles manufactured by a factory are given below find the correlation Coefficient & the two regression lines

| | | | | | | |
|---|---|---|---|---|---|---|
| length ($x$) : | 3 | 5 | 6 | 7 | 10 | 11 |
| weighth ($y$) : | 8 | 12 | 11 | 14 | 16 | 17 |

| $x$ | $y$ | $x^2$ | $y^2$ | $xy$ |
|---|---|---|---|---|
| 3 | 8 | 9 | 64 | 24 |
| 5 | 12 | 25 | 144 | 60 |
| 6 | 11 | 36 | 121 | 66 |
| 7 | 14 | 49 | 196 | 98 |
| 10 | 16 | 100 | 256 | 160 |
| 11 | 17 | 121 | 289 | 187 |
| 42 | 78 | 340 | 1070 | 595 |

$$\bar{x} = \frac{\Sigma x}{n} = \frac{42}{6} = \underline{7}$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{78}{6} = \underline{13}$$

$$\sigma_{xy} = \frac{\Sigma xy}{n} - \bar{x}\bar{y}$$

$$= \frac{595}{6} - 7 \times 13 = \underline{8.17}$$

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

$$= \frac{340}{6} - 7^2 = \underline{7.67}$$

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

$$= \frac{1070}{6} - 13^2 = \underline{9.33}$$

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{8.17}{\sqrt{7.67 \times 9.33}} = \underline{0.97}$$

Regression eqn of y on x is

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

$$y - 13 = \frac{8.17}{7.67}(x - 7)$$

regression eqn of x on y is

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

$$= x - 7 = \frac{8.17}{9.33}(y - 13)$$

from the following data find the most likely value of $y$ when $x = 24$ (use regression eqn)

| | $y$ | $x$ |
|---|---|---|
| mean | 985.8 | 18.1 |
| S.D | 36.4 | 2.0 |

$r = 0.58$

regression eqn of $y$ on $x$ is

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

$$= \frac{r \cdot \sigma_x \cdot \sigma_y}{\sigma_x^2}(x - \bar{x})$$

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

$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

$\sigma_{xy} = r\sigma_x \sigma_y$

$$y - 985.8 = \frac{0.58 \times 36.4}{2}(x - 18.1)$$

$$y = 10.556x + 794.74$$

when $x = 24$, $\Rightarrow y = 1048.084$

two lines of regression are $x + 2y - 5 = 0$ & $2x + 3y - 8 = 0$ & variance of $x = 2$ find $\bar{x}, \bar{y}, r$ & $\sigma_y^2$

the lines are $x + 2y - 5 = 0$ & $2x + 3y - 8 = 0$

Solving these two, $\bar{x} = 1$ & $\bar{y} = -2$, neglecting the signs

take $\bar{x} = 1$ & $\bar{y} = 2$

assume $x + 2y - 5 = 0$ is the regression line of $y$ on $x$ & $2x + 3y - 8 = 0$ is the regression line of $x$ on $y$

i.e., $y = \frac{5-x}{2}$ & $x = \frac{8-3y}{2}$

$b_{yx} = -1/2$ & $b_{xy} = -3/2$

$r = \sqrt{b_{yx} \cdot b_{xy}}$

$= \sqrt{-1/2 \times -3/2} = \pm 0.866$

Since $b_{yx}$ & $b_{xy}$ are $-ve$, $r = -0.866$

$b_{yx} = \frac{r \sigma_y}{\sigma_x}$

$\sigma_y = \frac{b_{yx} \times \sigma_x}{r} = \frac{-1/2 \times \sqrt{12}}{-0.866} = 2.00009$

$\cancel{\frac{-0.866 \times 12}{0.75}}$

find the regression eqn from the following data

| | | | |
|---|---|---|---|
| x | 0 | 1 | 2 |
| y | 4 | 7 | 12 |

| x | y | $x^2$ | $y^2$ | xy |
|---|---|---|---|---|
| 0 | 4 | 0 | 16 | 0 |
| 1 | 7 | 1 | 49 | 7 |
| 2 | 12 | 4 | 144 | 24 |
| 3 | 23 | 5 | 209 | 31 |

$\bar{x} = \frac{\Sigma x}{n} = \frac{3}{3} = 1$

$\bar{y} = \frac{\Sigma y}{n} = \frac{23}{3} = 7.6$

$\sigma xy = \frac{31}{3} - 1 \times 7.6 = 2.73$

$\sigma x^2 = \frac{\Sigma x^2}{n} - \bar{x}^2 = \frac{5}{3} - 1 = 0.66$

$\sigma y^2 = \frac{\Sigma y^2}{n} - \bar{y}^2 = \frac{209}{3} - 57.76 = 11.90$

regression eq of y on x

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

$y - 7.6 = \frac{2.73}{0.66}(x - 1)$

eq of x on y

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

$x - 1 = \frac{2.73}{11.90}(y - 7.6)$

- The eqns of two regression lines obtained in a correlation analysis are as follows. $4y = 9x + 15$, $25x = 6y + 7$. Identify the regression eqns & obtain the correlation coeff & mean values of x & y

$4y = 9x + 15$

$y = \frac{9}{4}x + \frac{15}{4}$

$\therefore b_{yx} = 9/4$

$25x = 6y + 7$ — ②

$x = \frac{6}{25} y + \frac{7}{25}$

$\therefore b_{xy} = \frac{6}{25}$

$r^2 = b_{yx} \times b_{xy}$

$= \frac{9}{4} \times \frac{6}{25} = \frac{54}{100}$

$r = \pm \sqrt{54/100} = \pm 0.735$

Since both $b_{xy}$ & $b_{yx}$ are positive

$r = 0.735$

Solving ① & ②, $x = 2.6$ & $y = 9.61$

$\therefore \bar{x} = 2.6$ & $\bar{y} = 9.61$

③ $8x - 10y + 66 = 0$, $40x - 18y = 214$ $\sigma_x^2 = 9$.

find $\bar{x}, \bar{y}, r, \sigma_y$ $\bar{x} = 13$, $\bar{y} = 17$, $r = 0.6$, $\sigma_y^2 = 16$

$8x - 10y + 66 = 0$ — ①

~~$8x = 66 + 10y$~~

$8x = 10y - 66$ — ①

$x = \frac{10y - 66}{8}$ $10y = \cancel{-8x + 66}$

$b_{xy} = \frac{10}{8}$ $y = \frac{8x + 66}{10}$

$= \frac{5}{8}$ $b_{yx} = \frac{8}{10} = \frac{4}{5}$

$40x - 18y = 214 \quad - (2)$

$18y = 40x - 214$ $\qquad$ $214 + 18y = 40x$

$y = \frac{40x - 214}{18}$ $\qquad$ $x = \frac{214 + 18y}{40}$

$b_{yx} = \frac{40}{18}$ $\qquad$ $b_{xy} = \frac{18}{40} = \frac{9}{20}$

from (1) & (2)

$8x - 10y = -66 \quad \times 5$

$40x - 18y = 214$

$40x - 50y = -330$

$40x - 18y = 214$

$-32y = -544$

$y = 17$

$8x - 10 \times 17 = -66$

$8x - 170 = -66$

$8x = -66 + 170$

$x = 13$

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

$r^2 = b_{yx} \times b_{xy}$

$= \cancel{\frac{40}{18}} \cancel{\times \frac{10}{8}} \; \frac{4}{5} \times \frac{9}{20}$

$r = \pm\sqrt{\frac{36}{100}} = +0.6$

$$b_{yx} = \frac{\cancel{\sigma_{xy}}\ \sigma_{xy}}{\sigma_x^2}$$

$$= \frac{r \cdot \cancel{\sigma_x} \cdot \sigma_y}{\sigma_x^{\cancel{2}}}$$

$$b_{yx} = \frac{r \cdot \sigma_y}{\sigma_x}$$

$$\sigma_y = \frac{b_{yx} \cdot \sigma_x}{r}$$

$$= \frac{\frac{4}{5} \times \sqrt{13}}{0.6}$$

$$= \underline{\underline{4}}$$

$$\underline{\underline{\sigma_y^2 = 16}}$$