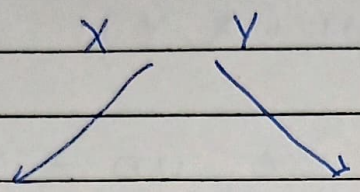
Unit-6 : Correlation and Regression

# Correlation :- $-1 \leq r \leq 1$




Positive / Direct

$X\uparrow \; Y\uparrow$

$X\downarrow \; Y\downarrow$

Negative / Diverse

$X\uparrow \; Y\downarrow$

$X\downarrow \; Y\uparrow$

If change in one variable effects change in another variable than two variables are said to be correlation.

- $r = -1 \longrightarrow$ perfect positive negative
- $r = 1 \longrightarrow$ perfect positive
- $r = 0 \longrightarrow$ no correlation

# Results on correlation coefficient -

- Correlation coefficient is independent of change of origin and change of scale.

$r(X, Y)$

$U = X - 10$

$V = Y + 2$ $\longrightarrow$ origin

$r(U, V) = r(X, Y)$

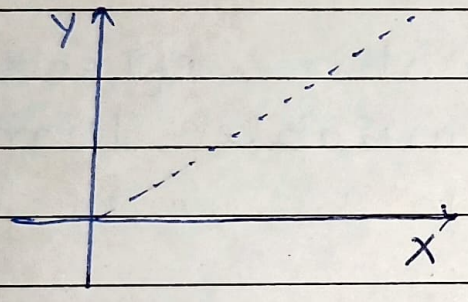
$U = 2X$ ⟶ Scale

$Y = \frac{V}{10}$

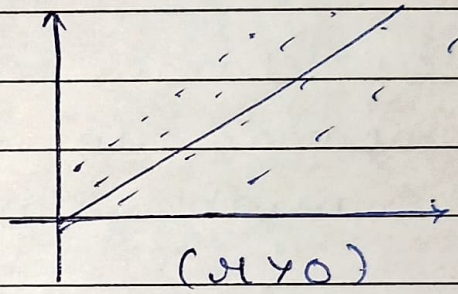
$r(U,V) = r(X,Y)$

- Two Independent variable are uncorrelated.

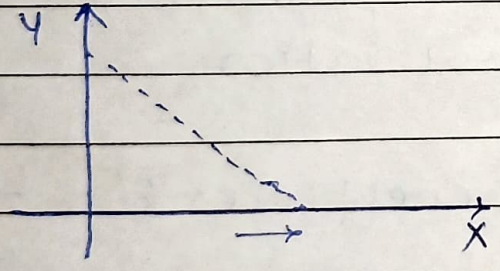
# Scatter plot / diagrams -



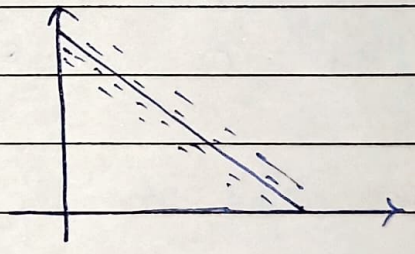

$(r = 1)$



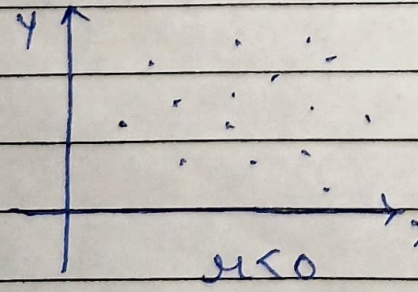
$(r > 0)$




$(r = -1)$



$(r < 0)$




$r < 0$

# Karl Pearson's correlation coefficient -

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

where, $\sigma_{xy} = \frac{1}{n} \Sigma x_i y_i - \bar{x}\bar{y}$

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

$$\sigma_x^2 = \frac{1}{n} \Sigma y_i^2 - \bar{y}^2$$

Q. Calculate the correlation coefficient for x and y.

→

| UV | X | Y | $U = X - 68$ | $V = Y - 69$ | $U^2$ | $V^2$ |
|---|---|---|---|---|---|---|
| 6 | 65 | 67 | -3 | -2 | 9 | 4 |
| 2 | 66 | 68 | -2 | -1 | 4 | 1 |
| 4 | 67 | 65 | -1 | -4 | 1 | 16 |
| 1 | 67 | 68 | -1 | -1 | 1 | 1 |
| 0 | 68 | 72 | 0 | 3 | 0 | 9 |
| 3 | 69 | 72 | 1 | 3 | 1 | 9 |
| 0 | 70 | 69 | 2 | 0 | 4 | 0 |
| 8 | 72 | 71 | 4 | 2 | 16 | 4 |
| 24 | | | 0 | 0 | 36 | 44 |

$\bar{U} = 0$

$\bar{V} = 0$

$$\sigma_U^2 = \frac{1}{8}(36 - 0) = 4.5$$

$$\sigma_V^2 = \frac{1}{8}(44 - 0) = 5.5$$

$$\sigma_{UV} = \frac{1}{8}(24) - 0 = 3$$

$$r(U,V) = \frac{3}{\sqrt{4.5 \times 5.5}} = 0.604$$

Since correlation coefficient is independent of change of origin therefore -

$$r(X,Y) = r(U,V) = 0.604$$

# Correlation coefficient for Bi-variate data and probability distribution -

Q. The following table gives according to age, the frequency of marks obtained by 100 students -

| | Age (x) → | 18 | 19 | 20 | 21 | Total |
|---|---|---|---|---|---|---|
| | Marks (Y) ↓ | u = -1 | 0 | 1 | 2 | h(v) |
| -2 | 10 - 20 | 4 (-8) | 2 (0) | 2 (2) | - | 8 |
| -1 | 20 - 30 | 5 (-5) | 4 (0) | 6 (6) | 4 (8) | 19 |
| 0 | 30 - 40 | 6 (0) | 8 (0) | 10 (0) | 11 (0) | 35 |
| 1 | 40 - 50 | 4 (-4) | 4 (0) | 6 (6) | 8 (16) | 22 |
| 2 | 50 - 60 | - | 2 (0) | 4 (8) | 4 (16) | 10 |
| 3 | 60 - 70 | - | 2 (0) | 3 (9) | 1 (6) | 6 |
| | Σuvf(u,v) | 9 | 0 | 13 | 30 | 52 |
| | Total G(u) | 19 | 22 | 31 | 28 | 100 |
| | u g(u) | -19 | 0 | 31 | 56 | 68 |
| | u² g(u) | 19 | 0 | 31 | 56 | 162 |

$u = x - 19$

$v = \frac{y - 35}{10}$

$$\rho(u,v) = \frac{\sigma_{uv}}{\sigma_u \sigma_v} = \frac{0.35}{\sqrt{1.1576 \times 1.6075}} = 0.26 \quad \text{Ans.}$$

$$\sigma_{uv} = \frac{1}{N} \sum_u \sum_v uv f(u,v) - \bar{u}\bar{v}$$

$$= \frac{52}{100} - (0.68)(0.25) = 0.35$$

$$\sigma_u^2 = \frac{1}{N}\sum_u u^2 g(u) - \bar{u}^2 = \frac{162}{100} - (0.68)^2 = 1.1576$$

$$\sigma_v^2 = \frac{1}{N}\sum_v v^2 h(v) - \bar{v}^2 = \frac{167}{100} - (0.25)^2 = 1.6075$$

$$\bar{u} = \frac{1}{N}\sum_u u g(u) = \frac{1}{100}(68) = 0.68$$

$$\bar{v} = \frac{1}{N}\sum_v v h(v) = \frac{25}{100} = 0.25$$

| $h(v)$ | $vh(v)$ | $v^2h(v)$ | $\sum uvf(u,v)$ |
|---|---|---|---|
| 8 | -16 | 32 | 4 |
| 19 | -19 | 19 | -9 |
| 35 | 0 | 0 | 0 |
| 22 | 22 | 22 | 18 |
| 10 | 20 | 40 | 24 |
| 6 | 18 | 54 | 15 |
| 100 | 25 | 167 | 52 |

# Rank correlation -

⊙ Spearman's Rank correlation coefficient -

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

$$d_i = x_i - y_i$$

$$\sum d_i = 0$$

Q. The rank of same 16 students in maths and physics are given

| X | Y | $di$ | $di^2$ |
|---|---|---|---|
| 1 | 1 | 0 | 0 |
| 2 | 10 | -8 | 64 |
| 3 | 3 | 0 | 0 |
| 4 | 4 | 0 | 0 |
| 5 | 5 | 0 | 0 |
| 6 | 7 | -1 | 1 |
| 7 | 2 | 5 | 25 |
| 8 | 6 | 2 | 4 |
| 9 | 8 | 1 | 1 |
| 10 | 11 | -1 | 1 |
| 11 | 15 | -4 | 16 |
| 12 | 9 | 3 | 9 |
| 13 | 14 | -1 | 1 |
| 14 | 12 | 2 | 4 |
| 15 | 06 | -1 | 1 |
| 16 | 13 | 3 | 9 |
| | | | 136 |

Calculate the rank correlation coefficient for expertise of this group in mathematics and physics.

Sol.

$$\rho = 1 - \frac{6 \Sigma di^2}{16(16^2-1)}$$

$$= 1 - \frac{6(136)}{16(16^2-1)}$$

$$\rho = 0.8 \quad \text{—— Ans.}$$

Q. 10 competitors in a musical test were ranked three judges

| Ranked by A(X) | Rank by B(Y) | Rank by C(Z) |
|---|---|---|
| 1 | 3 | 6 |
| 6 | 5 | 5 |
| 5 | 8 | 9 |
| 10 | 4 | 8 |
| 3 | 7 | 1 |
| 2 | 10 | 2 |
| 4 | 2 | 3 |
| 9 | 1 | 10 |
| 7 | 6 | 5 |
| 8 | 9 | 7 |

Discuss which pair of judges has the nearest approach to common liking in music.

- Solution of previous question -

$d_1 = X - Y$
$d_2 = Y - Z$
$d_3 = X - Z$

| $d_1$ | $d_2$ | $d_3$ | $d_1^2$ | $d_2^2$ | $d_3^2$ |
|---|---|---|---|---|---|
| -2 | -3 | -5 | 4 | 9 | |
| 1 | 0 | 1 | 1 | 0 | |
| -3 | -1 | -4 | 9 | 1 | |
| 6 | -4 | 2 | 36 | 16 | |
| -4 | 6 | 2 | 16 | 36 | |
| -8 | 8 | 0 | 64 | 64 | |
| 2 | -1 | 1 | 4 | 1 | |
| 8 | -9 | -1 | 64 | 81 | |
| 1 | 1 | 2 | 1 | 1 | |
| -1 | 2 | 1 | 1 | 4 | |
| | | | 200 | 213 | |

$\Sigma d_1^2 = 200$
$\Sigma d_2^2 = 214$
$\Sigma d_3^2 =$

$$f(X,Y) = 1 - \frac{6\,\Sigma d_3^2}{n(n^2-1)} = 1 - \frac{6(200)}{10(99)} = -0.21$$

$$f(Y,Z) = 1 - \frac{6\,\Sigma d_2^2}{n(n^2-1)} = 1 - \frac{6(213)}{990} = -0.29$$

$$f(X,Z) = 1 - \frac{6\,\Sigma d_3^2}{n(n^2-1)} = 1 - \frac{6(57)}{990} = 0.65$$

Since rank $f(x,z)$ is maximum, the judges A and C has the nearest approach to common liking in music.

# Tied Ranks / Repeated Ranks -

Q. Obtained the rank correlation coefficient for the following data -

| X | Y | x | y | $d_i = x_i - y_i$ | $d_i^2$ |
|---|---|---|---|---|---|
| 68 | 62 | 4 | 5 | -1 | 1 |
| 64 | 58 | 6 | 7 | -1 | 1 |
| 75 | 68 | 2.5 | 3.5 | -1 | 1 |
| 50 | 45 | 9 | 10 | -1 | 1 |
| 64 | 81 | 6 | 1 | 5 | 25 |
| 80 | 60 | 1 | 6 | -5 | 25 |
| 75 | 68 | 2.5 | 3.5 | -1 | 1 |
| 40 | 48 | 10 | 9 | 1 | 1 |
| 55 | 50 | 8 | 8 | 0 | 0 |
| 64 | 70 | 6 | 2 | 4 | 16 |
| | | | | | 72 |

$$\Sigma d_i^2 = 72$$

$$f = 1 - \frac{6\left(\Sigma d_i^2 + \frac{m(m^2-1)}{12}\right)}{n(n^2-1)}$$

(correction factor)

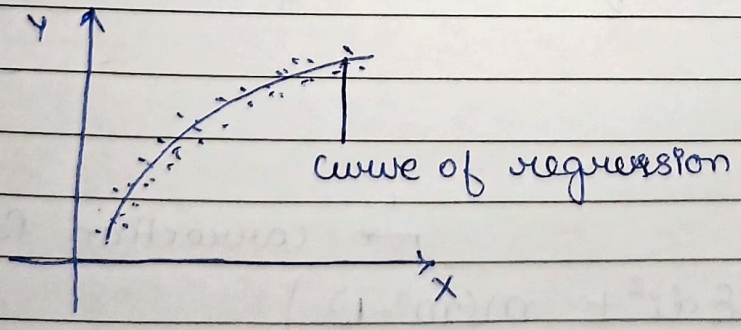
64 → 3
68 → 2
75 → 2

$$\rho = 1 - 6\left(\frac{72 + \frac{3(3^2-1)}{12} + \frac{2(2^2-1)}{12} + \frac{2(2^2-1)}{12}}{10(99)}\right)$$

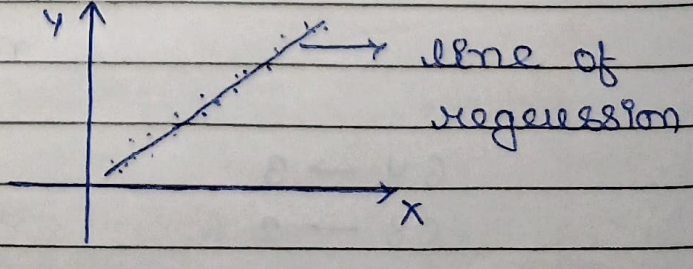
$$\rho = 0.55$$

# Regression -

Regression is the study of nature of relationship between variables so that one may be able to predict the unknown value of one variable for a known value of another variable. One variable is considered as independent and another is taken as dependent variable.
We predict the value of dependent variable using the value of independent variable.




# linear regression



# line of Regression of y on x

$$y = a + bx$$

The line of y on x is given by -

$y - \bar{y} = b_{yx}(x - \bar{x})$, where $b_{yx} = r\frac{\sigma_y}{\sigma_x}$ is regression coefficient of y on x.

# line of Regression of x on y

$$x = a + by$$

The line of x on y is -

$$x - \bar{x} = b_{xy}(y - \bar{y}),$$

$b_{xy} = r\frac{\sigma_x}{\sigma_y}$ is regression coefficient of x on y.

Q. Obtain the equations of two lines of regression for the following data

| x | y |
|---|---|
| 65 | 67 |
| 66 | 68 |
| 67 | 65 |
| 67 | 68 |
| 68 | 72 |
| 69 | 72 |
| 70 | 69 |
| 72 | 71 |

Also obtain the estimate of x for $y = 70$.

Sol. From previous solved question in correlation.

$$u = 0.604$$
$$\sigma_x^2 = \sigma_u^2 = 4.5$$
$$\sigma_y^2 = \sigma_v^2 = 5.5$$

$$\bar{x} = 68 \qquad \bar{y} = 69$$

$$\sigma_{ax+b}^2 = \sigma_x^2$$

$$b_{yx} = u\frac{\sigma_y}{\sigma_x} = 0.604\,\frac{\sqrt{5.5}}{\sqrt{4.5}} = 0.67$$

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 69 = 0.67(x - 68)$$

$$y = 69 + 0.67x - 0.67 \times 68$$
$$y = 0.67x + 23.66$$

x on y

$$x = 0.55y + 30.05$$

↓

when $y = 70$

$$x = 0.55 \times 70 + 30.05$$
$$x = 68.55$$

# Properties of regression coefficient -

1. The correlation coefficient is the geometric mean between the regression coefficients. Denoted by $r$

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

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

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

$$r^2 = b_{xy} + b_{yx}$$

$$r = \pm \sqrt{b_{xy} + b_{yx}}$$

$r$, $b_{xy}$, $b_{yx}$

2. $r$, $b_{xy}$, $b_{yx}$ all Have the same signs. ( signs of regression coefficient should be same )

3. If one regression coefficient is greater than 1 than another must be less than 1.

4. Regression coefficients are independent of change of origin but not of scale.

$$u = \frac{x-a}{h}, \quad v = \frac{y-b}{k}$$

$$b_{xy} = \frac{h}{k} b_{uv}$$

$$b_{yx} = \frac{k}{h} b_{vu}$$

5) The point of intersection of the regression line is ordered pair $(\bar{x}, \bar{y})$.

6) The modulus of the arithmatic mean between two regression coefficients is greater than the modulus of the correlation coefficients.

$$\left| \frac{b_{xy} + b_{yx}}{2} \right| > |r|$$

# Angle between two regression line

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

If $r = 0$, $\theta = \frac{\pi}{2}$

If two variables are uncorrelated, the lines of regression are perpendicular to each other.

If $r = \pm 1$, $\theta = 0$ or $\pi$

In the case of perfect correlation, +ve or -ve the two lines of regression coincide.

Q. In a partially destroyed laboratory only the following result are readable. The variance of $x$ is 9 and the regression lines are $8X - 10Y + 66 = 0$

$40X - 18Y = 214$

(i) What are the means of X and Y — $\bar{x}$ and $\bar{y}$ ?

(ii) What is the correlation coefficient between Y and X

(iii) What are standard deviation of Y ($\sigma_y = ?$)?

Sol. $8\bar{X} - 10\bar{Y} + 66 = 0$

$40\bar{X} - 18\bar{Y} = 214$

$\bar{Y} = 17 \qquad \bar{X} = 13$

X on Y

$$X = \frac{10}{8}Y - \frac{66}{8}$$

$$b_{xy} = \frac{10}{8}$$

Y on X

$$Y = \frac{40}{18}X - \frac{214}{18}$$

$$b_{yx} = \frac{40}{18}$$

$$\frac{10}{8} \times \frac{40}{18} = \frac{50}{18}$$

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

$$b_{yx} = \frac{8}{10} \qquad b_{xy} = \frac{18}{40}$$

$$r^2 = \frac{8}{10} \times \frac{18}{40} = \frac{9}{25}$$

$$\boxed{r = \pm \frac{3}{5}}$$

- $r = \frac{3}{5} \longrightarrow$ positive

(iii) $$b_{yx} = r \frac{\sigma_y}{\sigma_x}$$

$$\frac{8}{10} = \frac{3}{5} \times \frac{\sigma_y}{9}$$

$$\boxed{\sigma_y = 4}$$

Q. Find the most likely price in mumbai corresponding to the price of rupees 70 at kolkata from the following

| | Kolkata (X) | Mumbai (Y) |
|---|---|---|
| Average Price | 65 | 67 |
| Standard deviation | 2.5 | 3.5 |

The correlation coefficient between the prices in two cities is 0.8.

Sol: $r = 0.8$

$\bar{X} = 65, \quad \bar{Y} = 67$

$\sigma_X = 2.5, \quad \sigma_Y = 3.5$

Y on X

$$Y - 67 = 0.8\left(\frac{3.5}{2.5}\right)(X - 65)$$

$$Y = 1.12X - 5.8$$

$$Y = 1.12(70) - 5.8$$

$$\hat{Y} = 72.6$$

# Fitting of a curve -

~~The del~~

- Fitting of a straight line :

$$Y = a + bX$$

Normal Equation :

$$\Sigma Y = \Sigma a + \Sigma bX$$

$$\Sigma Y = na + b\Sigma X \quad - \text{multiple by } 1$$

$$\Sigma XY = \Sigma aX + \Sigma bX^2$$

Normal Equations

$$\Sigma XY = a\Sigma X + b\Sigma X^2 \quad - \text{multiple by } X$$

- Fitting a parabola -

$$Y = a + b_1X + b_2X^2$$

$$\Sigma y = na + b_1 \Sigma x + b_2 \Sigma x^2 \quad \text{— (i)}$$

$$\Sigma xy = a\Sigma x + b_1 \Sigma x^2 + b_2 \Sigma x^3 \quad \text{— (ii)}$$

$$\Sigma x^2 y = a\Sigma x^2 + b_1 \Sigma x^3 + b_2 \Sigma x^4 \quad \text{— (iii)}$$

Normal Equation

Q. For 10 randomly selected observation the following data were recorded were values of x and y -

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

Find the regression coefficient and the regression equation using the non linear form $y = a + b_1x + b_2x^2$

Sol.

$$\Sigma y = na + \Sigma b_1x + b_2 \Sigma x^2$$
$$\Sigma xy = a\Sigma x + b_1 \Sigma x + b_2 \Sigma x^2$$
$$\Sigma x^2y = a\Sigma x^2 + b_1 \Sigma x^3 + b_2 \Sigma x^4$$

$\Sigma x = 34$
$\Sigma x^2 = 154$
$\Sigma y = 95$
$\Sigma x^3 = 820$
$\Sigma x^4 = 4774$
$\Sigma xy = 377$
$\Sigma x^2y = 1849$

$$10a + 34b_1 + 154b_2 = 95$$
$$34a + 154b_1 + 820b_2 = 377$$
$$154a + 820b_1 + 4774b_2 = 1849$$

$a = 1.8$

$b_1 = 3.49$

$b_2 = -0.27$

$$Y = 1.8 + 3.49X - 0.27X^2$$

Q. Fit an exponential curve $Y = ab^X$ to the following data

| X | Y | $X^2$ | $U = \log_{10} Y$ |
|---|---|---|---|
| 1 | 1 | | 0 |
| 2 | 1.2 | | 0.0792 |
| 3 | 1.8 | | 0.2553 |
| 4 | 2.5 | | 0.3979 |
| 5 | 3.6 | | 0.5563 |
| 6 | 4.7 | | 0.6721 |
| 7 | 6.6 | | 0.8195 |
| 8 | 9.1 | | 0.9590 |

Sol $\log_{10} Y = \log_{10} a + X \log_{10} b$

$U = A + BX$

$U = \log_{10} Y$ $\quad B = \log_{10} b$

$A = \log_{10} a$,

$a = 10^A$, $\quad b = 10^B$

$$U = A + BX$$
$$\Sigma U = nA + B\Sigma X$$
$$\Sigma UX = A\Sigma X + B\Sigma X^2$$

$$\Sigma X = 36$$
$$\Sigma X^2 = 204$$

$$U = \log_{10} Y$$

$$8A + 36B = 3.7393$$
$$36A + 204B = 22.7385$$

$$A = -0.16632 \Rightarrow a = 10^{-0.1662} = 0.682$$
$$B = 0.1408 \Rightarrow b = 10^{0.1408} = 1.383$$

$$Y = (0.682)(1.383)^X$$