Correlation : Relationship between two variable X and Y is called correlation.

Degree of association between these two variables is called correlation coefficient(ρ). (Harvest, Rainfall)

Types of Correlation :

(i) Positive correlation : $\rho > 0$

(ii) Negative Correlation: $\rho < 0$

(iii) No correlation : $\rho = 0$

(iv) Perfect positive correlation: $\rho = 1$

(v) Perfect Negative correlation : $\rho = -1$

$-1 \leq \rho \leq 1$

Methods to find Correlation:

(i) Coefficient of correlation or Karl ~~Pearson~~ Pearson's correlation.

(ii) Spearman's Rank correlation.

Scanned with CamScanner

Karl Pearson's Coefficient: (Direct Method)

x	y
x_1	y_1
x_2	y_2
⋮	⋮
x_n	y_n

$$\bar{x} = \frac{\sum x_i}{n}$$

$$\bar{y} = \frac{\sum y_i}{n}$$

$$\rho = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2} \cdot \sqrt{\sum(y-\bar{y})^2}} = \frac{S_{xy}}{S_x \cdot S_y}$$

Ex:

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

find Karl pearson's coefficient?

Soln

$\bar{x} = 3$, $\bar{y} = 5.4$

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
1	7	-2	1.6	4	2.56	-3.2
2	3	-1	-2.4	1	5.76	2.4
3	4	0	-1.4	0	1.96	0
4	5	1	-0.4	1	0.16	-0.4
5	8	2	2.6	4	6.76	5.2

$$\rho = \frac{4}{\sqrt{10}\sqrt{17.2}}$$

$$= \frac{[illegible]}{}$$

$$= 0.3049$$

$$\simeq 0.305$$

Assumed Mean Method:

$$\boxed{\rho = \cancel{\sum d_x}}$$

x	y	$d_x = (x-60)$	$d_y = (y-80)$	$(d_x - \bar{d_x})^2$	$(d_y - \bar{d_y})^2$	$(d_x - \bar{d_x})(d_y - \bar{d_y})$
68	81	8	1	6.76	40.96	−17.16
62	87	2	7	11.56	0.16	2.04
63	92	3	12	5.76	21.16	−10.56
65	93	5	13	0.16	31.36	−2.16
69	85	9	5	12.96	5.76	−9.36

$(d_x - \bar{d_x})$	$(d_y - \bar{d_y})$
2.6	−6.6
−3.4	−0.6
−2.4	4.4
−0.4	5.4
3.6	−2.6

Take assume mean of x as 60

" " " y " 80

$\bar{d_x} = 5.4$

$\bar{d_y} = 7.6$

$$\rho = \frac{\sum (d_x - \bar{d_x})(d_y - \bar{d_y})}{\sqrt{\sum (d_x - \bar{d_x})^2}\sqrt{\sum (d_y - \bar{d_y})^2}} = \frac{-37.2}{\sqrt{37.2}\sqrt{99.4}}$$

$$= -0.611$$ Ans

Spearman's Rank Correlation: This is also applicable for qualitative data.

Suppose for a sample of size n, X_i, Y_i are converted to ranks $R(X_i)$, $R(Y_i)$ and then r_s is computed as

$$r_s = \rho_{R(X), R(Y)} = \frac{\text{cov}(R(X), R(Y))}{\sigma_{R(X)} \cdot \sigma_{R(Y)}}$$

$$= \frac{\sum (R_i - \bar{R}) \cdot (S_i - \bar{S})}{\sqrt{\sum (R_i - \bar{R})^2} \cdot \sqrt{(S_i - \bar{S})^2}}$$

where $\bar{R} = \frac{1}{n} \sum R_i$, $\bar{S} = \frac{1}{n} \sum S_i$

$$d_i^2 = (R_i - S_i)^2$$

$$\bar{R} = \bar{S} . = \sum i = \frac{(n+1)}{2}$$

$$\sum_{i=1}^{n} R_i^2 = \sum_{i=1}^{n} S_i^2 = \sum_{i=1}^{n} i^2 = \frac{n}{6}(n+1)(2n+1)$$

$$\boxed{r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2-1)}}$$

(There is no repeated rank)

(Ex):

(Quality of) Food	R_1	R_2	d	d^2
A	2	1	1	1
B	1	3	-2	4
C	4	2	2	4
D	3	4	-1	1
E	5	5	0	0
F	7	6	1	1
G	6	7	-1	1

$$r_s = 1 - \frac{6 \times 12}{7^3 - 7} = \underline{\underline{0.786}}$$

Correlation is high and positively related.

Scanned with CamScanner

Chi:

X	R_1	Y	R_2	d_i	d^2
97.8	5	77.2	7	-2	4
99.2	1	85.0	2	-1	1
98.8	2	78.9	4	-2	4
98.3	4	75.8	6	-2	4
98.4	3	77.2	5	-2	4
96.7	7	87.2	1	6	36
97.1	6	83.8	3	3	9

$$r_s = 1 - \frac{6 \times 62}{7^3 - 7} = \underline{\underline{0.107}}$$

Repeated Rank

Ex.

X	R_1	Y	R_2	d	d^2
20	8	28	3	5	25
22	7	24	7.5	-0.5	0.25
20	3	24	7.5	-4.5	20.25
23	5.5	25	6	-0.5	0.25
30	1.5	26	5	-3.5	12.25
30	1.5	27	4	-2.5	6.25
23	5.5	32	1	4.5	20.25
24	4	30	2	2	4
					88.5

$\frac{1^{st} + 2^{nd}}{2} = (1.5)^{th}$

$\frac{5^{th} + 6^{th}}{2} \rightarrow 5.5$

$\frac{7^{th} + 8^{th}}{2} = 7.5$

Scanned with CamScanner

$$r_s = 1 - \frac{6\left\{\sum d_i^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \cdots\right\}}{n^3 - n}$$

→ how many times a no. is repeated

$$= 1 - \frac{6\left\{88.5 + \frac{1}{12}(2^3-2) + \frac{1}{12}(2^3-2) + \frac{1}{12}(2^3-2)\right\}}{8^3 - 8}$$

$$\boxed{r_s = -0.0714}$$

→ Negatively correlated and less correlation.

#

Regression Analysis :

Regression Lines :

① y on x line → Used for finding the value of y when x is given

② x on y line → Used for finding the value of x when y is given

The regression line y on x is defined as

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

and the regression line x on y is defined as

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

where b_{xy} → Regression coefficient x on y

b_{yx} → Regression coeff y on x.

and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

<u>Data points</u>

x	y
x_1	y_1
x_2	y_2
⋮	⋮
x_n	y_n

Regression Coefficient:

$$b_{yx} = \rho \frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$

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

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

and ρ is correlation coefficient.

and $$b_{xy} = \rho \frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma y^2 - \frac{(\Sigma y)^2}{n}}$$

Note: If the values in data points are very large then we consider

$u = x - A$

$v = y - B$

$A \rightarrow$ Assumed mean of x

$B \rightarrow$ Assumed mean of y.

then $b_{xy} = \frac{\Sigma uv - \frac{\Sigma u \Sigma v}{n}}{\Sigma v^2 - \frac{(\Sigma v)^2}{n}}$, $b_{yx} = \frac{\Sigma uv - \frac{\Sigma u \Sigma v}{n}}{\Sigma u^2 - \frac{(\Sigma u)^2}{n}}$

Properties of Regression lines :-

(i) The regression lines always intersect at the pt $(\bar{x}, \bar{y})$.

(ii) If two regression lines are given and θ be the acute angle between them is

$$\tan\theta = \left|\frac{1-\rho^2}{\rho}\right| \left|\frac{\sigma_x \sigma_y}{\sigma_x^2+\sigma_y^2}\right|$$ where ρ is correlation coeff.

Proof:

Regression lines

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

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

$$m_1 = \rho \frac{\sigma_y}{\sigma_x}$$

$$m_2 = \frac{1}{\rho} \frac{\sigma_y}{\sigma_x}$$

$$\tan\theta = \left|\frac{m_1 - m_2}{1+m_1 \cdot m_2}\right| = \frac{\left|\rho \cdot \frac{\sigma_y}{\sigma_x} - \frac{1}{\rho}\frac{\sigma_y}{\sigma_x}\right|}{\left|1+\rho \cdot \frac{\sigma_y}{\sigma_x} \cdot \frac{1}{\rho}\frac{\sigma_y}{\sigma_x}\right|}$$

$$= \frac{\left|\rho - \frac{1}{\rho}\right| \left|\frac{\sigma_y}{\sigma_x}\right|}{\left|\frac{\sigma_x^2+\sigma_y^2}{\sigma_x^2}\right|}$$

$$= \left|\frac{\rho^2-1}{\rho}\right| \cdot \left|\frac{\sigma_x \cdot \sigma_y}{\sigma_x^2+\sigma_y^2}\right|$$

Remark! (i) If $\rho = 1$ (perfectly correlated)

$$\tan\theta = 0 \Rightarrow \theta = 0$$

Both lines coincides.

(ii) $\rho = 0$ (No correlation)

$$\tan\theta = \infty$$

$$\Rightarrow \theta = \frac{\pi}{2}$$ (Both lines are perpendicular)

(iii) The coefficient of correlation is G.M of Regression coefficients i.e $\rho = \sqrt{b_{yx} \cdot b_{xy}}$

Pf: $\because b_{yx} = \rho \cdot \frac{\sigma_y}{\sigma_x}$

$$b_{xy} = \rho \cdot \frac{\sigma_x}{\sigma_y}$$

$$\sqrt{b_{yx} \cdot b_{xy}} = \sqrt{\rho^2 \cdot \frac{\sigma_y}{\sigma_x} \cdot \frac{\sigma_x}{\sigma_y}}$$

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

Note! b_{yx} and b_{xy} either both are +ve or both are −ve. If both are +ve the ρ is +ve and if both are −ve then ρ is −ve

(iv) If one regression coefficient is greater than 1 then other should be less than 1.

Pf

$$\because \quad \rho = \sqrt{b_{yx} \cdot b_{xy}}$$

$$\rho^2 = b_{yx} \cdot b_{xy}$$

$$\because \; -1 \leq \rho \leq 1 \quad \Rightarrow \quad \rho^2 \leq 1$$

$$\Rightarrow \quad b_{yx} \cdot b_{xy} \leq 1$$

#4

Q: Find the lines of regression from the following data

Age of Husband: 25, 22, 28, 26, 35, 20, 22, 40, 20, 18.

Age of wife: 18, 15, 20, 17, 22, 14, 16, 21, 15, 14.

Hence estimate (i) the age of husband when age of wife is 19

(ii) the age of wife, when the age of husband is 30

(iii) correlation coefficient.

Sol^n

$x =$ Age of husband

$y =$ Age of wife

$u = x - A = x - 26$

$v = y - B = y - 17$

Scanned with CamScanner

x	y	u	v	u^2	v^2	uv
25	18	-1	1	1	1	-1
22	15	-4	-2	16	4	8
28	20	2	3	4	9	6
26	17	0	0	0	0	0
35	22	9	5	81	25	45
20	14	-6	-3	36	9	18
22	16	-4	-1	16	1	4
40	21	14	4	196	16	56
20	15	-6	-2	36	4	12
18	14	-8	-3	64	9	24
256	172	-4	2	450	78	172

$n = 10$, $\bar{x} = \frac{256}{10} = 25.6$

$\bar{y} = \frac{172}{10} = 17.2$

Regression Coefficients:

$$b_{yx} = \frac{\Sigma uv - \frac{\Sigma u \Sigma v}{n}}{\Sigma u^2 - \frac{(\Sigma u)^2}{n}} = \frac{172 - \frac{(-4).(2)}{10}}{450 - \frac{(-4)^2}{10}} = 0.385$$

$$b_{xy} = \frac{172 - \frac{(-4)(2)}{10}}{78 - \frac{(2)^2}{10}} = 2.23$$

Hence line of regression y on x

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

$$\Rightarrow \boxed{y = 0.385x + 7.34}$$

Scanned with CamScanner

x on y

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

$\Rightarrow \boxed{x = 2.23y - 12.76}$

(i) When $y = 19$ (given) then (x on y)

$x = 2.23 \times 19 - 12.76$

$x = 29.6 \simeq 30$

Age of husband $x = 30$ years.

(ii) When $x = 30$ (given)

$y = 0.385 \times 30 + 7.34$ (y on x)

$= 18.89 \simeq 19$

Hence age of wife is 19 when $x = 30$.

(iii) $\rho = \pm\sqrt{b_{yx} \cdot b_{xy}} = \sqrt{0.385 \times 2.23}$

$= 0.927$

Q: If regression lines are

$5x - y = 22$ and

$64x - 45y = 24$.

find (i) mean values of x and y.

(ii) regression coefficients

(iii) Coefficient of correlation

(iv) Standard deviation of y i.e σ_y if the variance of x is 25.

Soln (i) $x = 6$, $y = 8$

i.e $\bar{x} = 6$, $\bar{y} = 8$ } Soln of regression lines are mean.

(ii) Regression Coefficients:

The regression line x on y

$5x - y = 22$

$\Rightarrow x = \frac{y}{5} + \frac{22}{5}$

$b_{xy} = \frac{1}{5}$

regression line y on x

$64x - 45y = 24$

$y = \frac{64x}{45} - \frac{24}{45}$; $b_{yx} = \frac{64}{45}$

$$\rho = \pm\sqrt{\frac{64}{45} \times \frac{1}{5}}$$

$$\rho = 0.533$$

(iv) $\sigma_x^2 = 25 \Rightarrow \sigma_x = 5$

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

$$\frac{\cancel{64}^{8}}{\cancel{45}_{3}} = \frac{\cancel{8}}{\cancel{15}} \times \frac{\sigma_y}{5}$$

$$\Rightarrow \boxed{\sigma_y = \frac{40}{3} = 13.33} \quad \#$$

Scanned with CamScanner