

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 : $\rho = -1$

Methods to find Correlation :

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

(ii) Spearman's Rank correlation.

Karl Pearson's coefficient: (Direct Method)

x	y
x_1	y_1
x_2	y_2
\vdots	\vdots
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?

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

Soln

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

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

$$= \frac{4}{\sqrt{172}}$$

$$= 0.3049$$

$$\approx 0.305$$

Assumed Mean Method:

$$r = \frac{\sum dx dy}{\sqrt{\sum dx^2} \sqrt{\sum dy^2}}$$

X	Y	dx = (x - 60)	dy = (y - 80)	(dx - \bar{dx})	(dy - \bar{dy})	(dx - \bar{dx})(dy - \bar{dy})
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

Take assume mean of x as 60

" " " y " 80

$$\bar{dx} = 5.4$$

$$\bar{dy} = 7.6$$

(dx - \bar{dx})	(dy - \bar{dy})
2.6	-6.6
-3.4	-0.6
-2.4	4.4
-0.4	5.4
3.6	-2.6

$$r = \frac{\sum (dx - \bar{dx})(dy - \bar{dy})}{\sqrt{\sum (dx - \bar{dx})^2} \sqrt{\sum (dy - \bar{dy})^2}} = \frac{-37.2}{\sqrt{37.2} \sqrt{99.4}}$$

$$= -0.611 \quad \underline{\text{Ans}}$$

Spearman's Rank Correlation:
for qualitative data.

This is also applicable

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

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

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

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

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

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

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

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

{ 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.

Chi-Sq:

X	R ₁	Y	R ₂	d _i	d ²
97.8	5	73.2	7	-2	4
99.2	1	85.8	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

$$\chi^2 = 1 - \frac{6 \times 62}{7^3 - 7} = \underline{\underline{0.107}}$$

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

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

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

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

88.7

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

→ how many times a no. is repeated

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

$$\gamma_s = -0.0714$$

→ Negatively correlated and less correlation.

ft

Regression Analysis :

Regression Lines :

- ① y on x line \rightarrow Used for finding the value of y when x is given
- ② x on y line \rightarrow 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 $b_{xy} \rightarrow$ Regression coefficient x on y

$b_{yx} \rightarrow$ Regression coeff y on x .

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

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

Data points

x	y
x_1	y_1
x_2	y_2
\vdots	\vdots
x_n	y_n

Regression Coefficients:

$$b_{yx} = \frac{\sum \delta y}{\sum \delta x} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

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

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

and ρ is correlation coefficient.

and $b_{xy} = \frac{\sum \delta x}{\sum \delta y} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum y^2 - \frac{(\sum y)^2}{n}}$

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

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

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

Ex 1)

Regression lines

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

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

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

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

$$\begin{aligned} \tan \theta &= \left| \frac{m_1 - m_2}{1 + m_1 m_2} \right| = \left| \frac{r \frac{\sigma_y}{\sigma_x} - \frac{1}{r} \frac{\sigma_y}{\sigma_x}}{1 + r \frac{\sigma_y}{\sigma_x} \cdot \frac{1}{r} \frac{\sigma_y}{\sigma_x}} \right| \\ &= \left| \frac{r - \frac{1}{r}}{1 + \frac{\sigma_y^2}{\sigma_x^2}} \right| = \left| \frac{r - \frac{1}{r}}{\frac{\sigma_x^2 + \sigma_y^2}{\sigma_x^2}} \right| \end{aligned}$$

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

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

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

Both lines coincide.

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

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

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

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

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

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.

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

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

$$\therefore -1 \leq r \leq 1 \Rightarrow r^2 \leq 1$$

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

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

Soln
 $x = \text{Age of husband}$
 $y = \text{Age of wife}$

$$u = x - A = x - 26$$

$$v = y - B = y - 17$$

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

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

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

Regression coefficients!

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

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

Hence line of regression y on x

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

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

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

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

(i) When $y = 19$ ^(given) then $(x \text{ on } y)$

$$x = 2.23 \times 19 - 12.76$$

$$x = 29.6 \approx 30$$

Age of husband $x = 30$ years.

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

$$y = 0.385 \times 30 + 7.34 \quad (y \text{ on } x)$$

$$= 18.89 \approx 19$$

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

$$5x - y = 22 \text{ and}$$

$$64x - 45y = 24.$$

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

$$5x - y = 22$$

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

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

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)

$$\sigma_x^2 = 25 \Rightarrow \sigma_x = 5$$

$$\text{by } \rho = \rho \cdot \frac{\sigma_y}{\sigma_x}$$

$$\frac{64 \cancel{8}}{4 \cancel{8} 3} = \frac{\cancel{8}}{\cancel{18}} \times \frac{\sigma_y}{5}$$

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

①

gf $X \sim N(\mu, \sigma^2)$

then

Then $\left(\frac{X-\mu}{\sigma}\right) \sim N(0,1)$

and

$\left(\frac{X-\mu}{\sigma}\right)^2 \sim \chi^2_{(1)}$

gf $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$

then $\sum \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2_n$

~~$\chi^2 = \sum \frac{(O-E)^2}{E}$ → O, observed frequency, E → expected frequency~~

②

gf $X \sim N(0,1)$ and $Y \sim \chi^2_{(n)}$ and are independent

⊗

then $T = \frac{X}{\sqrt{\frac{Y}{n}}} \sim t_{(n)}$

③

gf $X \sim \chi^2_{(m)}$ and $Y \sim \chi^2_{(n)}$

and X and Y are independent then

$F = \frac{X/m}{Y/n} \sim F_{(m,n)}$

Chi-Square test!

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$f_o \rightarrow$ observed frequency

$f_e \rightarrow$ expected frequency

degree of freedom = 2

Conditions: ① Total frequency should be large <<
 $\sum f_o = N > 30$

② $\sum f_o = \sum f_e$

③ No expected frequency should be less than 5.

H_0 : No association between both attributes.

χ^2 - test of Independence of Attributes.

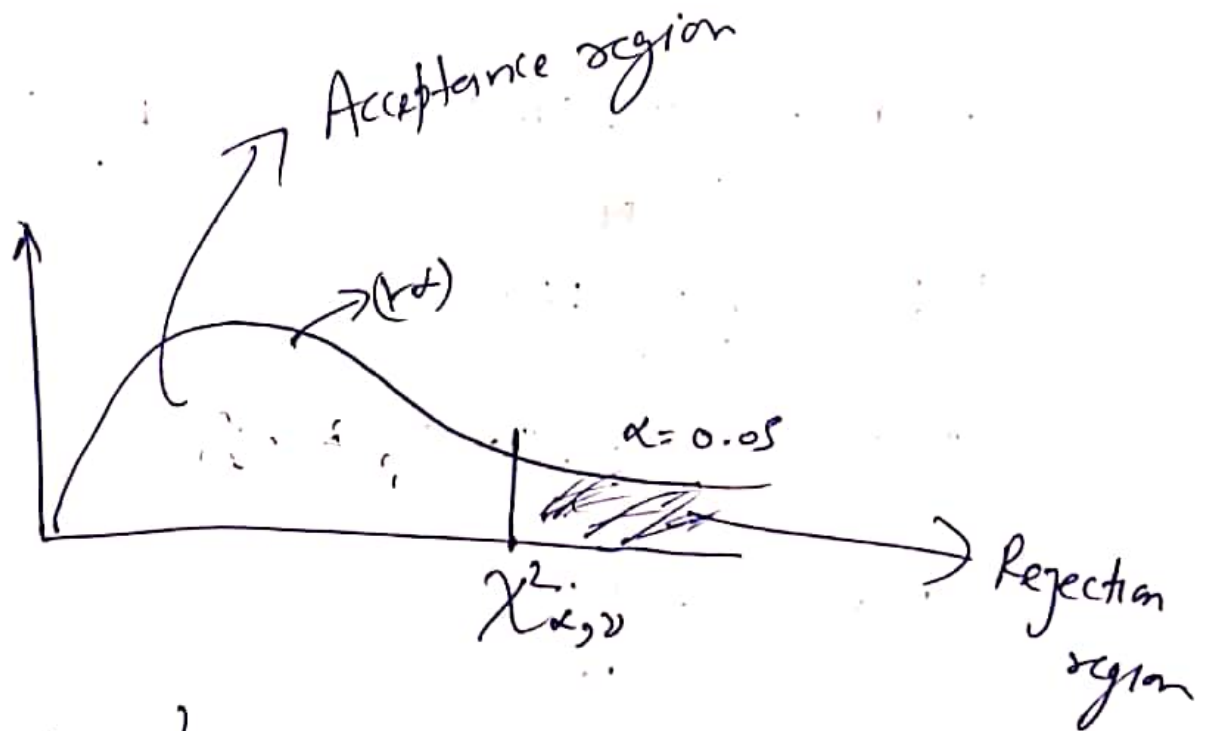
Whether two attributes are independent or not.

or

	Gender	
	M	F
Bigger		
Small		
Can		

→ Contingency table

Crucifers



Note: Cal $\chi^2_{\alpha, v} < \text{tab } \chi^2_{\alpha, v}$ then

H_0 accepted

If Cal $\chi^2_{\alpha, v} > \text{tab } \chi^2_{\alpha, v}$ then

H_0 rejected

How to compute expected frequency:

	R_T		
	a	b	$a+b$
	c	d	$c+d$
C_T	$a+c$	$b+d$	$N = a+b+c+d$

Expected frequency f_e

$$E(a) = \frac{(a+b) \cdot (a+c)}{N} \quad \text{ie} \quad \frac{R_T \times C_T}{N}$$

$$E(b) = \frac{(a+b) \cdot (b+d)}{N}$$

$$E(c) = \frac{(c+d) \cdot (a+c)}{N}$$

$$E(d) = \frac{(c+d) \cdot (b+d)}{N}$$

f_o	f_e
a	$E(a)$
b	$E(b)$
c	$E(c)$
d	$E(d)$

$$\begin{aligned}
 d.f. &= (r-1) \cdot (c-1) \\
 &\quad \downarrow \quad \quad \downarrow \\
 &\quad \text{no of row} \quad \text{no of column} \\
 &= (2-1) \times (2-1) \\
 &= 1
 \end{aligned}$$

then $\chi^2_{\alpha} < \text{tab } \chi^2_{\alpha}$ then H_0 is accepted.

(Ex) In an anti malarial campaign in a certain area, quinine was administered to 812 persons out of total population of 3248.

The no of fever cases is shown below:

Treatment	Fever	No fever	Total
Quinine	20	792	812
No quinine	220	2216	2436
Total	240	3008	3248

Discuss the usefulness of quinine in checking Malaria at 5% level of significance

H_0 : There is no association between quinine and Malaria

Treatment	Fever	No fever	Total
Quinine	20	792	812
No quinine	220	2216	2436
	240	3008	$N = 3248$

$$E(20) = \frac{812 \times 240}{3248} = 60$$

$$E(792) = \frac{812 \times 3008}{3248} = 752$$

$$E(220) = \frac{2436 \times 240}{3248} = 180$$

$$E(2216) = \frac{2436 \times 3008}{3248} = 2256$$

f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
20	60	-40	1600	$1600/60 = 26.6$
792	752	40	1600	2.12
220	180	40	1600	8.88
2216	2256	40	1600	0.70
				<u>38.37</u>

$$\chi^2_{\text{cal}} = 38.37$$

$$df = (2-1)(2-1) = 1, \quad \text{level of significance } \alpha = 0.05$$

$$\chi^2_{\text{tab}} = 3.84$$

$$\therefore \chi^2_{\text{cal}} > \chi^2_{\text{tab}}$$

then we reject H_0 .

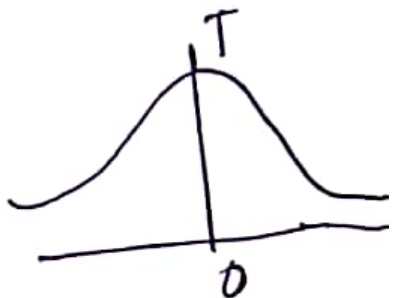
Definitely there is association between quinine and malaria

T-test

$$X \sim N(0,1), \quad Y \sim \chi^2_{(n)}$$

and X and Y are ind.

then $T = \frac{\cancel{X}}{\sqrt{\frac{Y}{n}}} \sim t_{(n)}$



$$H_0: \bar{X} = \mu$$

$$t = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \sqrt{n} \left(\frac{\bar{X} - \mu}{s} \right)$$

$\bar{X} \rightarrow$ sample mean, $n \rightarrow$ sample size

$\mu \rightarrow$ population mean; $s \rightarrow$ s.d. of sample

where

$$S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

$$d.f = n-1$$

$$9) \quad \text{Cal } |t_{\alpha, n}| < |t_{\text{tab } t_{\alpha, n}}|$$

then H_0 accept.