Correlation and Regression

Correlation

Introduction:- → If the Distribution involving only one variable it is called Univariate Variable distribution.

→ If the Distribution involving two variables it is called Bivariate distribution

→ If the Distribution involving Three variables it is called trivariate distribution

If the distribution involving more than three variables are called Multivariate distribution.

In our Correlation we have to study relationship between two variables.

Q → Define Correlation? Types of Correlation? Define Karl Pearson Coefficient of Correlation.

Def:- If change in one variable effects to change in another variable then the two variables are said to be correlated then that relationship is called correlation.

Correlation are two types.

① Positive Correlation (or) Direct Correlation

② Negative Correlation (or) Indirect Correlation

Positive correlation:- If the two variables are deviates in same direction is called (or) Direct (or) Positive Correlation ie If one variable is increases then another variable is also increases.

Negative (or) Indirect Correlation:-

If the two variables are deviates in opposite direction is called Negative (or) Indirect correlation. ie If one variable is increases then another variable is decreases.

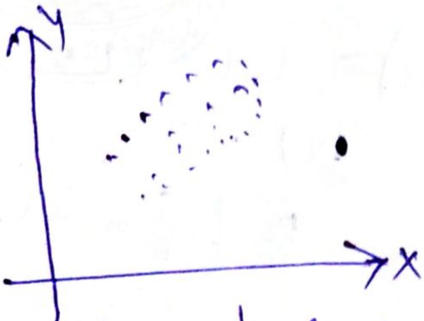
⇒ Scattered diagram:-

The diagrametic representation of the Bivariate data is called "Scattered diagram".

Using Scattered diagram we can get fairy good idea about the distribution.

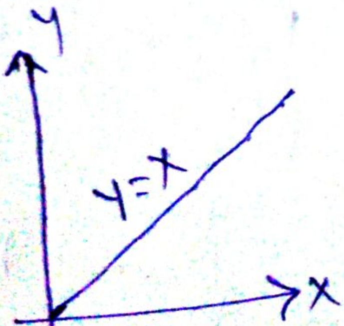
If the values are x and y plotted among XY plane. The diagram of dots obtained we can get fairy good idea. If the variables are correlated (or) not. If the points are scattered very close each other. Then there is high degree of correlation. and if the points are widely scattered then there is Poor correlation.

High Degree Positive Corre

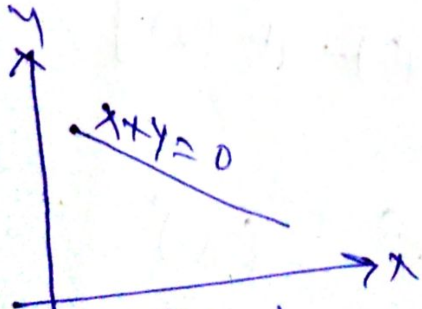


Low degree of -ve Corre

No Corre



$r=+1$



$r=-1$

Karl Pearson Coefficient of Correlation :- As a major intensity of association between two Variables Karl Pearson a bio mathematician and Statistician Developed a formulae is called Karl Pearson Coefficient of correlation. It is denoted with r_{xy} defined as

$$r_{xy} = \frac{Cov(xy)}{\sigma_x \cdot \sigma_y} \quad (or) \quad r_{xy} = \frac{Cov(xy)}{\sqrt{V(x)V(y)}}$$

Where $Cov(xy) = \frac{\sum x_i y_i}{n} - (\bar{x})(\bar{y})$

(or)

$$Cov(xy) = E(xy) - E(x) \cdot E(y)$$

(or)

$$Cov(xy) = E[\{x - E(x)\}\{y - E(y)\}]$$

$V(x) = E[x - E(x)]^2$

(or)

$V(x) = E(x^2) - [E(x)]^2$

(or)

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

$V(y) = E[y - E(y)]^2$

(or)

$V(y) = E(y^2) - [E(y)]^2$

(or)

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

⇒ Properties of Karl Pearson Coefficient of Correlation.

(1) Correlation Coefficient lies between -1 to $+1$ (or) $-1 \le r_{xy} \le +1$

(2) There is no origin and Scale effect on correlation.

Theorem:— Prove limits of correlation (or) Prove $-1 \le r_{xy} \le +1$

Proof:- Karl Pearson Coefficient of Correlation $r_{xy} = \dfrac{Cov(xy)}{\sqrt{V(x) \cdot V(y)}}$.

Where

$$Cov(xy) = E\left[\{x - E(x)\}\{y - E(y)\}\right]$$

$$Cov(xy) = \frac{1}{n}\sum\left[\{x_i - \bar{x}\}\{y_i - \bar{y}\}\right]$$

$V(x) = E[x - E(x)]^2$ | $V(y) = E[y - E(y)]^2$

$V(x) = \frac{1}{n}\sum(x_i - \bar{x})^2$ | $V(y) = \frac{1}{n}\sum(y_i - \bar{y})^2$

$$r_{xy} = \frac{Cov(xy)}{\sqrt{V(x)\cdot V(y)}}.$$

$$= \frac{\frac{1}{n}\sum\left[\{x_i-\bar{x}\}\{y_i-\bar{y}\}\right].}{\sqrt{\left\{\frac{1}{n}\sum(x_i-\bar{x})^2\right\}\left\{\frac{1}{n}\sum(y_i-\bar{y})^2\right\}}}$$

$$r_{xy} = \frac{\cancel{\frac{1}{n}}\sum(x_i-\bar{x})(y_i-\bar{y})}{\cancel{\frac{1}{n}}\sqrt{\sum(x_i-\bar{x})^2\cdot\sum(y_i-\bar{y})^2}}$$

~~Again~~ Let $x_i-\bar{x}=a_i$, $y_i-\bar{y}=b_i$

$$r_{xy} = \frac{\sum a_i b_i}{\sqrt{\sum a_i^2\cdot\sum b_i^2}}$$

Squaring on both sides.

$$r_{xy}^2 = \frac{\left[\sum a_i b_i\right]^2}{\sum a_i^2\cdot\sum b_i^2}.$$

According to Cauchy Schwartz Inequality

$$\left[E(xy)\right]^2 \le E(x^2)\cdot E(y^2)$$

$$\left[\sum a_i b_i\right]^2 \le \sum a_i^2\cdot\sum b_i^2$$

$$r_{xy}^2 \le 1$$

$r_{xy} \leq +1$

$\boxed{-1 \leq r_{xy} \leq +1}$ Hence the Proof

→ Prove That there is no origin and scale effect on correlation.

Proof:- Karl Pearson Coefficient of correlation $r_{xy} = \frac{Cov(xy)}{\sqrt{V(x) \cdot V(y)}}$

$$r_{xy} = \frac{E[\{x - E(y)\}\{y - E(y)\}]}{\sqrt{E[x - E(x)]^2 \cdot E[y - E(y)]^2}}$$

To change of origin and scale in x and y variables.

$U = \frac{x-a}{h}$ $V = \frac{y-b}{K}$

$VK = y - b$

$Uh = x - a$

$x = a + Uh$

$E(x) = E[a + Uh]$

$= E(a) + E(Uh)$

$= a + h\,E(U)$

$y = VK + b$

Taking Expectation

$E(y) = E(VK + b)$

$E(y) = E(VK) + E(b)$

$E(y) = K\,E(V) + b$

$$\therefore r_{xy} = \frac{E[\{x-E(x)\}\{y-E(y)\}]}{\sqrt{[E\{x-E(x)\}^2][E\{y-E(y)\}^2]}}$$

$$= \frac{E[h\{u-E(u)\}\, k\{v-E(v)\}]}{\sqrt{E[h\{u-E(u)\}]^2 \cdot E[k\{v-E(v)\}]^2}}$$

$$= \frac{hk\, E[\{u-E(u)\}\{v-E(v)\}]}{\sqrt{h^2 E[u-E(u)]^2\, k^2 E[v-E(v)]^2}}$$

$$= \frac{\cancel{hk}\, E[\{u-E(u)\}\{v-E(v)\}]}{\cancel{hk}\sqrt{E[u-E(u)]^2 \cdot E[v-E(v)]^2}}$$

$$= r_{uv}.$$

$\therefore r_{xy} = r_{uv}.$

$\therefore$ There is no origin and scale effect on correlation.

Problem

① Compute Karl Pearson Coefficient between x and y for the given data.

x : 1 2 3 4 5

y : 10 20 30 40 50

Sol:- Karl Pearson Coefficient of Correlation.

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

where $Cov(x,y) = \frac{\sum x_i y_i}{n} - (\bar{x})(\bar{y})$

$$\sigma_x = \sqrt{\frac{\sum x_i^2}{n} - (\bar{x})^2}, \quad \sigma_y = \sqrt{\frac{\sum y_i^2}{n} - (\bar{y})^2}$$

$$\bar{x} = \frac{\sum x_i}{n}, \quad \bar{y} = \frac{\sum y_i}{n}.$$

x_i	y_i	$\sum x_i y_i$	$\sum x_i^2$	$\sum y_i^2$
1	10	10	1	100
2	20	40	4	400
3	30	90	9	900
4	40	160	16	1600
5	50	250	25	2500
15	150	550	55	5500

$\bar{x} = \frac{\sum x_i}{n} = \frac{15}{5} = 3$ $\boxed{\bar{x} = 3}$ | $\bar{y} = \frac{\sum y_i}{n} = \frac{150}{5}$, $\boxed{\bar{y} = 30}$

$$Cov(xy) = \frac{\sum x_i y_i}{n} - (\bar{x})(\bar{y})$$

$$= \frac{550}{5} - (3)(30)$$

$$= 110 - 90$$

$$\boxed{Cov(xy) = 20}$$

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

$$= \sqrt{\frac{55}{5} - (3)^2}$$

$$= \sqrt{2}$$

$$\boxed{\sigma x = 1.4142}$$

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

$$= \sqrt{\frac{5500}{5} - (30)^2}$$

$$= \sqrt{200}$$

$$\boxed{\sigma y = 14.1421}$$

$$r = \frac{Cov(xy)}{\sigma x \cdot \sigma y} = \frac{20}{(1.4142)(14.1421)}$$

$$= \frac{20}{19.9997} = 1$$

$$\boxed{r = 1}$$

Problem: Exam Papers.

(1) Find correlation coefficient between industrial production and export using the following data comment on result.

Production (x):	55	56	58	59	60	60	62
Export (Y):	35	38	38	39	44	43	45

(2) Calculate coefficient of correlation from the following data.

x:	50	60	70	90	100
y:	65	51	40	26	8

Rank correlation coefficient:

Def:- Arrangement of given data into order of merit is called 'Rank correlation coefficient.

Some times we come across statistical series in which the variables under correlated (or) not capable of qualitative measurement can be arranged in to the serial order. On that case of correlation does Karl Pearson coefficient

Q) When do we use Spearman's rank Correlation formulae.

Spearman's rank Correlation Coefficient

$$\rho = 1 - \left[\frac{6\sum d_i^2}{n(n^2-1)}\right]$$

where, $d_i = R_x - R_y$

n = number of pairs.

Theorem obtain rank Correlation coefficient formulae.

Proof:- Assumming That no two individuals get a same rank.

let x_i is a rank of Attribute-A.

y_i is a rank of Attribute-B.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1+2+3+\cdots+n}{n} = \frac{\frac{n(n+1)}{2}}{n} = \frac{n+1}{2}$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{1+2+3+\cdots+n}{n} = \frac{\frac{n(n+1)}{2}}{n} = \frac{n+1}{2}$$

$$\boxed{\therefore \bar{x} = \bar{y}}$$

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

$$= \frac{1^2+2^2+3^2+\cdots+n^2}{n} - \left(\frac{n+1}{2}\right)^2$$

$$= \frac{\frac{\cancel{n}(n+1)(2n+1)}{6}}{\cancel{n}} - \left(\frac{n+1}{2}\right)^2$$

$$\sigma_x^2 = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}$$

$$\sigma_x^2 = \frac{n+1}{2}\left[\frac{2n+1}{3} - \frac{n+1}{2}\right]$$

$$= \frac{n+1}{2}\left[\frac{2(2n+1)-3(n+1)}{6}\right]$$

$$= \frac{n+1}{2}\left[\frac{4n+2-3n-3}{6}\right]$$

$$= \frac{n+1}{2}\left[\frac{n-1}{6}\right]$$

$$\sigma_x^2 = \frac{n^2-1}{6}$$

$||^{ly}$ $\sigma_y^2 = \frac{n^2-1}{6}$

$$\boxed{\therefore \sigma_x^2 = \sigma_y^2}$$

~~also $d_i = R_x - R_y$~~ also $d_i = x_i - y_i$

$d_i = x_i - \bar{x} + \bar{x} - y_i \Rightarrow d_i = x_i - \bar{x} + \bar{y} - y_i$

$d_i = (x_i - \bar{x}) - (y_i - \bar{y})$ $[\because \bar{x} = \bar{y}]$

Taking Σ divide by n and squaring on both sides.

$$\frac{\sum d_i^2}{n} = \frac{1}{n}\sum\left[(x_i-\bar{x})-(y_i-\bar{y})\right]^2$$

$$= \frac{1}{n}\left[\sum\left\{(x_i-\bar{x})^2+(y_i-\bar{y})-2(x_i-\bar{x})(y_i-\bar{y})\right\}\right]$$

$$= \frac{1}{n}\sum(x_i-\bar{x})^2 + \frac{\sum(y_i-\bar{y})^2}{n} - 2\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n}$$

$$= \sigma_x^2 + \sigma_y^2 - 2\,\mathrm{Cov}(xy)$$

$$\boxed{\begin{array}{l} r = \frac{\mathrm{Cov}(xy)}{\sigma_x\cdot\sigma_y} \\ \mathrm{Cov}(xy) = r\cdot\sigma_x\cdot\sigma_y \end{array}}$$

$$= \sigma_x^2 + \sigma_y^2 - r\cdot\sigma_x\cdot\sigma_y$$

$$\frac{\sum d_i^2}{n} = \sigma_x^2 + \sigma_x^2 - r\,\sigma_x\cdot\sigma_x \qquad [\because \sigma_x^2 = \sigma_y^2]$$

$$= 2\sigma_x^2 - r\,\sigma_x^2$$

$$= 2\sigma_x^2[1-r]$$

$$\boxed{\because \sigma_x^2 = \frac{n^2-1}{12}}$$

$$\frac{\sum d_i^2}{n} = \cancel{2}\left[\frac{n^2-1}{\cancel{12}_{6}}\right][1-r]$$

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

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

$$\boxed{1-\frac{6\sum d_i^2}{n(n^2-1)} = r}$$

$$\therefore \rho = r$$

$$\therefore \boxed{\rho = 1-\left[\frac{6\sum d_i^2}{n(n^2-1)}\right]}$$

→ Find Rank correlation Coefficient between marks of English and Computers

Marks in Eng: 48 50 47 55 43 60 68 45 52 64 58 56 72 40

Marks in Comp: 43 52 48 53 68 50 64 51 44 72 62 49 70 65

Sol:- Let us denote marks in English is variable x.
Marks in Computers is variable y.

x	Rank R_x	y	Rank R_y	$d_i = R_x - R_y$	Σd_i^2
48	10	43	14	10-14 = -4	16
50	9	52	8	9-8 = 1	1
47	11	48	12	11-12 = -1	1
55	7	53	7	0	0
43	13	68	3	10	100
60	4	50	10	-6	36
68	2	64	5	-3	9
45	12	51	9	3	9
52	8	44	13	-5	25
64	3	72	1	2	4
58	5	62	6	-1	1
56	6	49	11	-5	25
72	1	70	2	-1	1
40	14	65	4	10	100
					328

Spearman's correlation coefficient

$$\rho = 1 - \left[\frac{6 \sum d_i^2}{n(n^2-1)}\right]$$

$$\rho = 1 - \left[\frac{6(328)}{14(195)}\right]$$

$$\rho = 1 - \frac{1968}{2730}$$

$$= 1 - 0.7208$$

$$\boxed{\rho = 0.2792}$$

Problems. (1) Find Rank correlation coefficient between marks of Maths and Statistics.

Marks in Maths :	92	96	72	88	54	100	75	93	47	35
Marks in Stt :	100	47	74	94	75	99	72	89	50	40

Problem (2) The coefficient of rank correlation of marks obtained by 10 students in maths and statistics was found to be 0.5. It was later discovered that the difference in marks two subjects obtained by one of the student was wrongly taken as 3 insted of 7. find the correct rank correlation.

→ The following are the ranks assigned by two judges X and Y to 12 consistance in cooking competation. Find out what aggrement the judges had in judgement.

Entry No :	A	B	C	D	E	F	G	H	I	J	K	L
	1	9	2	10	3	11	8	4	12	7	5	6
Rank Judge-A :	1	9	2	10	3	11	8	4	12	7	5	6
Rank Judge-B :	2	9	1	7	4	10	8	3	12	6	5	11

Sol:- Let us denote the ranks given by judge-A is variable x.

Let us denote the ranks given by Judge-B is variable y

Spearman's rank correlation coefficient

$$\rho = 1 - \left[\frac{6 \Sigma d_i^2}{n(n^2-1)}\right] \quad \text{where } d_i = R_x - R_y$$

n = number of pairs

R_x	R_y	$d_i = R_x - R_y$	Σd_i^2
1	2	1-2 = -1	1
9	9	9-9 = 0	0
2	1	2-1 = 1	1
10	7	10-7 = 3	9
3	4	3-4 = -1	1
11	10	11-10 = 1	1
8	8	8-8 = 0	0
4	3	4-3 = 1	1
12	12	12-12 = 0	0
7	6	7-6 = 1	1
5	5	5-5 = 0	0
6	11	6-11 = -5	25
		Σd_i^2	40

$$\rho = 1 - \left[\frac{6 \Sigma d_i^2}{n(n^2-1)}\right]$$

$$\rho = 1 - \left[\frac{6(40)}{12(12^2-1)}\right]$$

$$= 1 - \left[\frac{240}{12(143)}\right]$$

$$= 1 - \frac{240}{1716}$$

$$= 1 - 0.1398$$

$$\boxed{\rho = 0.8609}$$

Problem
→ 10 competators in a musical fest were ranked by the three judges A, B and C in the following order.

Rank by Judge-A	1	6	5	10	3	2	4	9	7	8
Rank by Judge-B	3	5	8	4	7	10	2	1	6	9
Rank by Judge C	6	4	9	8	1	2	3	10	5	7

Using the rank correlation method discuss which pair of judges has the nearest approach to common liking music.

Sol:- Let us denote Ranks given by Judge-A is R_x.
" " " Judge-B is R_y
" " " Judge-C is R_z.

Rank corr b/w Judge-A Judge-B

$$\rho_{xy} = 1 - \left[\frac{6[\Sigma d^2_{xy}]}{n(n^2-1)}\right] \quad d_{xy} = R_x - R_y.$$

Rank corr b/w Judge-B and Judge-C

$$\rho_{yz} = 1 - \left[\frac{6[\Sigma d^2_{yz}]}{n(n^2-1)}\right] \quad d_{yz} = R_y - R_z$$

Rank corr b/w Judge-C and Judge-A

$$\rho_{zx} = 1 - \left[\frac{6[\Sigma d^2_{zx}]}{n(n^2-1)}\right] \quad d_{zx} = R_z - R_x$$

x	y	z	$d_{xy} = R_x - R_y$	d^2_{xy}	$d_{yz} = R_y - R_z$	d^2_{yz}	$d_{zx} = R_z - R_x$	d^2_{zx}
1	3	6	-2	4	-3	9	5	25
6	5	4	1	1	1	1	-2	4
5	8	9	-3	9	-1	1	4	16
10	4	8	6	36	-4	16	-2	4
3	7	1	-4	16	6	36	-2	4
2	10	2	-8	64	8	64	0	0
4	2	3	2	4	-1	1	-1	1
9	1	10	8	64	9	81	1	1
7	6	5	1	1	1	1	-2	4
8	9	7	-1	1	2	4	1	1
				200		214		60

$$\rho_{xy} = 1 - \left[\frac{6\sum d^2_{xy}}{n(n^2-1)}\right]$$

$$= 1 - \left[\frac{6(200)}{10(10^2-1)}\right]$$

$$= 1 - \left[\frac{1200}{990}\right]$$

$$= 1 - 1.2121$$

$$\boxed{\rho_{xy} = -0.2121}$$

$$\rho_{yz} = 1 - \left[\frac{6\sum d^2_{yz}}{n(n^2-1)}\right]$$

$$= 1 - \left[\frac{6(214)}{10(10^2-1)}\right]$$

$$= 1 - \left[\frac{1284}{990}\right]$$

$$= 1 - 1.2969$$

$$\boxed{\rho_{yz} = -0.2969}$$

$$\rho_{zx} = 1 - \left[\frac{6\sum d^2_{zx}}{n(n^2-1)}\right]$$

$$= 1 - \left[\frac{6(60)}{10(10^2-1)}\right]$$

$$= 1 - \frac{360}{990}$$

$$\rho_{zx} = 1 - 0.3636$$

$$\boxed{\rho_{zx} = 0.6364}$$

Conclusion :- $\rho_{21} = 0.6364$ is Maximum.
Judge C and Judge A are common approach to liking music.

Repeated Ranks (or) Tie Ranks:-

If any two (or) more individuals are equal in any classification with respect to characteristics A and B. If there is more than one item with the same value then the Spearman's rank correlation coefficient breaks down.

In this case common ranks are given to the repeated items. This common ranks is average of ranks with these items would have assumed. In this case Spearman's rank correlation coefficient formulae is

$$\rho = 1 - \left[\frac{6\left[\Sigma d_i^2 + CF\right]}{n(n^2-1)}\right]$$

where $CF = \frac{m(m^2-1)}{12}$

where m = number of times an item is repeated.

Rank correlation Problems [Ties]

① Calculate rank correlation.

x: 68 64 75 50 64 80 75 40 55 64

y: 62 58 68 45 81 60 68 48 50 70

Sol: In the given Problem marks are tie

∴ Spearman's rank Cor coe $\rho = 1 - \left[\frac{6[\Sigma d_i^2 + CF]}{n(n^2-1)}\right]$

where $CF = \frac{m(m^2-1)}{12}$

x	Decending order of x-series	Ranks	Rearrange Ranks (Rx)	y	Decending order of y-series	Ranks	Rearrange Ranks (Ry)
68	~~75~~ 80	1	4	62	81 →	1	5
64	75 (2) $\frac{2+3}{2}$	2.5	6	58	70 →	2	7
75	75 (3)	2.5	2.5	68	68 (3) $\frac{3+4}{2}$ →	3.5	3.5
50	68 →	4	9	45	68 (4) →	3.5	10
64	64 (5) →	6	6	81	62 →	5	1
80	64 (6) $\frac{5+6+7}{3} = 6$ →	6	1	60	60 →	6	6
75	64 (7) →	6	2.5	68	58 →	7	3.5
40	55 →	8	10	48	50 →	8	9
55	50 →	9	8	50	48 →	9	8
64	40 →	10	6	70	45 →	10	2

$d_i = R_x - R_y$	Σd_i^2
4 − 5 = −1	1
6 − 7 = −1	1
2.5 − 3.5 = −1	1
9 − 10 = −1	1
6 − 1 = 5	25
6 − 6 = 0	0
2.5 − 3.5 = −1	1
10 − 9 = 1	1
8 − 8 = 0	0
6 − 2 = 4	16
	47

$$CF = \frac{m(m^2-1)}{12}$$

$$= \frac{2(2^2-1)}{12} + \frac{3(3^2-1)}{12} + \frac{2(2^2-1)}{12}$$

$$= \frac{6}{12} + \frac{24}{12} + \frac{6}{12}$$

$$= 0.5 + 2 + 0.5$$

$$CF = 3$$

$$\boxed{\begin{aligned} \Sigma d_i^2 + CF &= 47+3 \\ \Sigma d_i^2 + CF &= 50 \end{aligned}}$$

$\therefore$ Spearman's Rank Cor Coe $\rho = 1 - \left[\frac{6[\Sigma d_i^2 + CF]}{n(n^2-1)}\right]$.

$$= 1 - \left[\frac{6(50)}{10(10^2-1)}\right]$$

$$= 1 - \frac{300}{990}$$

$$= 1 - 0.3030$$

$$\boxed{\rho = 0.697}$$

Problem

① Calculate Rank Correlation coefficient from the following data

60 70 75 60 83

marks in statistics: 40 46 54 46 50 75 60 75 80.

marks in physics: 45 45 50 43

② Calculate Rank correlation coefficient from the following data.

69 78 64

x: 70 65 71 62 58 65 55 48.

y: 91 76 65 83 90

→ obtain Rank cor coe to the following data.

x: 115 109 112 87 98 98 120 100 98 118

y: 75 73 85 70 76 6[illegible] 82 73 68 80

REGRESSION

→ Define Regression

Regression means Stepping back towards the average. It is a mathematical measurement in which we have to take average relationship between two (or) more Variables. It was developed by "Sir Francis Galton". There are always two lines of regression.

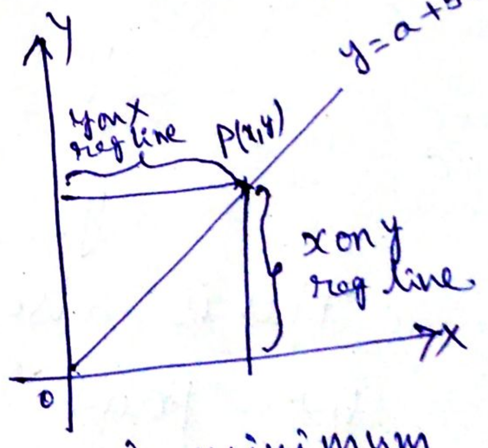
1. y on x regression

$$y-\bar{y} = r \cdot \frac{\sigma y}{\sigma x}(x-\bar{x})$$

2. x on y regression

$$x-\bar{x} = r \cdot \frac{\sigma x}{\sigma y}(y-\bar{y})$$

let us take two axis [x-axis, Y-axis]



In XY plane we plot straight line $y=a+bx$ and $P(x,y)$ is any point on the straight line.

The Parllel distance between Point $P(x,y)$ and X-axis is minimum we get x on y regression line.

The Parllel distance between Point $P(x,y)$ and Y-axis is minimum we get Y on x regression line.

Theorem

Derive the two regression lines

Proof:- Let (x_1, y_1) (x_2, y_2) (x_3, y_3) ..... (x_n, y_n) be the set of n pairs of observations. To fit y on x regression line we take straight line form $y = a + bx$. Here y is dependent variable and x is independent variable. a, b are unknown constants. To estimate these unknown constants we use principles of least squares.

Normal equations.

$$\sum y_i = na + b\sum x_i \quad \text{--- (1)}$$

$$\sum x_i y_i = a\sum x_i + b\sum x_i^2 \quad \text{--- (2)}$$

equation (1) is divide by n on both sides.

$$\frac{\sum y_i}{n} = \frac{\cancel{n}a}{\cancel{n}} + b\frac{\sum x_i}{n}$$

$$\bar{y} = a + b\bar{x} \quad \text{--- (3)}$$

This is also straight line passing through the point $(\bar{x}, \bar{y})$

also $\sigma_x^2 = \frac{\sum x_i^2}{n} - (\bar{x})^2$

$$\sigma^2 + (\bar{x})^2 = \frac{\sum x_i^2}{n} \quad \text{--- (4)}$$

equation (2) is divide by n and substitute (4)th equation.

$$\frac{\sum x_i y_i}{n} = a\frac{\sum x_i}{n} + b\frac{\sum x_i^2}{n},$$

$$\frac{\sum x_i y_i}{n} = a\bar{x} + b\frac{\sum x_i^2}{n}.$$

$$\frac{\sum x_i y_i}{n} = a\bar{x} + b\left[\sigma_x^2 + (\bar{x})^2\right] \quad \text{— (5)}$$

and also $\text{Cov}(xy) = \frac{\sum x_i y_i}{n} - (\bar{x})(\bar{y})$

let $\text{Cov}(xy) = \mu_{11}$

$$\mu_{11} = \frac{\sum x_i y_i}{n} - (\bar{x})(\bar{y})$$

$$\mu_{11} + (\bar{x})(\bar{y}) = \frac{\sum x_i y_i}{n}.$$

Substitute this value in to equation 5

$$\mu_{11} + (\bar{x})(\bar{y}) = a\bar{x} + b\sigma_x^2 + b(\bar{x})^2 \quad \text{— (6)}$$

equation (3) is multiply by $\bar{x}$ and substract from equation (6)

equation (3) $\times \bar{x}$.

$$\bar{x}\bar{y} = a\bar{x} + b(\bar{x})^2 \quad \text{— (7)}$$

(6) − (7)

$$\mu_{11} + \cancel{(\bar{x})(\bar{y})} = a\bar{x} + b\sigma_x^2 + b\cancel{(\bar{x})^2}$$

$$\underline{-\cancel{(\bar{x})(\bar{y})} = -a\bar{x} \qquad -b\cancel{(\bar{x})^2}}$$

$$\underline{\mu_{11} = b\sigma_x^2.}$$

$$b = \frac{\mu_{11}}{\sigma_x^2}.$$

$$r = \frac{\text{Cov}(xy)}{\sigma_x \cdot \sigma_y}$$

$$b = \frac{\text{Cov}(xy)}{\sigma_x^2}.$$

$$\boxed{r \cdot \sigma_x \cdot \sigma_y = \text{Cov}(xy)}$$

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

$(\bar{x}, \bar{y})$ is the point $b_1 = r \cdot \frac{\sigma y}{\sigma x}$ is the slope then y on x regression line is

$$y - \bar{y} = b_1 (x - \bar{x})$$

$$\boxed{y - \bar{y} = r \cdot \frac{\sigma y}{\sigma x} (x - \bar{x})}$$

Problems

Q → Why two lines of Regression

A:- There are always two lines of Regression.

① y on x regression line

$$y - \bar{y} = r \cdot \frac{\sigma y}{\sigma x} (x - \bar{x})$$

② x on y regression line

$$x - \bar{x} = r \cdot \frac{\sigma x}{\sigma y} (y - \bar{y})$$

To Estimate (or) predict the value of y to the given value of x we use y on x regression line. Since in y on x regression line we take straight line form $y = a + bx$. Here x is independent variable and y is dependent variable.

To Estimate (or) predict the value of x to the given value of y we use x on y regression line. Since in x on y regression line we take straight line form $x = a + by$. Here y is independent variable.

x is dependent variable.

Two regression lines are not Interchangable or reversable. Because of simple reason that the parllel distance between the point (x,y) and x-axis is minimum we get x on y regression line. The parllel distance between the point P(x,y) and Y-axis is minimum we get Y on x regression line.

In case of perfect correlation $r = \pm 1$ two regression lines are coinside

Y on x regression line

$$y - \bar{y} = r \cdot \frac{\sigma y}{\sigma x}(x - \bar{x})$$

$\underline{r = \pm 1}$ $$y - \bar{y} = \pm 1 \cdot \frac{\sigma y}{\sigma x}(x - \bar{x}) \quad \text{--- (1)}$$

x on y regression line

$$x - \bar{x} = r \cdot \frac{\sigma x}{\sigma y}(y - \bar{y})$$

$\underline{r = \pm 1}$ $$x - \bar{x} = \pm 1 \cdot \frac{\sigma x}{\sigma y}(y - \bar{y})$$

$$(x - \bar{x})\left(\pm 1 \cdot \frac{\sigma y}{\sigma x}\right) = y - \bar{y}.$$

$$\therefore y - \bar{y} = \pm \cdot \frac{\sigma y}{\sigma x}(x - \bar{x}) \quad \text{--- (2)}$$

equation (1) and (2) are equal.

except perfect correlation $(r = \pm 1)$ two regression lines are coinside. In any other case two regression lines are entairly different.

Regression Coefficients :-

In y on x regression line $y-\bar{y} = r.\frac{\sigma y}{\sigma x}(x-\bar{x})$.

In this $r.\frac{\sigma y}{\sigma x}$ is called regression coefficient of y on x regression line. It is denoted by b_{yx}.

$$\boxed{b_{yx} = r.\frac{\sigma y}{\sigma x}}$$

In x on y regression line $x-\bar{x} = r.\frac{\sigma x}{\sigma y}(y-\bar{y})$

In this $r.\frac{\sigma x}{\sigma y}$ is called regression coefficient of x on y regression line. It is denoted with b_{xy}

$$b_{xy} = r.\frac{\sigma x}{\sigma y}$$

Note:- (1) In Correlation $r_{xy} = r_{yx}$.

(2) In regression $b_{xy} \neq b_{yx}$.

Q → Properties of two regression lines:-

① Two regression lines are passing through the point $(\bar{x}, \bar{y})$

· y on x regression line $y-\bar{y} = r \cdot \frac{\sigma y}{\sigma x}(x-\bar{x})$

x on y regression line $x-\bar{x} = r \cdot \frac{\sigma x}{\sigma y}(y-\bar{y})$

② If $(\bar{x}, \bar{y})$ be the origin i.e $(0,0)$ the two regression lines are.

ⓐ y on x regression line $y = r \cdot \frac{\sigma y}{\sigma x}(x)$

ⓑ x on y regression line $x = r \cdot \frac{\sigma x}{\sigma y}(y)$

③ The Geomentric mean of two regression coefficient gives correlation coefficient.

i.e $\sqrt{b_{yx} \cdot b_{xy}}$

$$= \sqrt{r \cdot \frac{\cancel{\sigma y}}{\cancel{\sigma x}} \cdot r \cdot \frac{\cancel{\sigma x}}{\cancel{\sigma y}}}$$

$$= \sqrt{r^2} = r.$$

④ If one regression coefficient is increases then other regression coefficient decrease.

⑤ If $r=0$ then two regression lines are perpenticular.

⑥ There is no origin effect on two regression lines. But there is scale effect.

(7) Two regression lines are not coinside except incase of perfect correlation i.e $r = \pm 1$

(8) In correlation $r_{xy} = r_{yx}$. But in regression $b_{yx} \neq b_{xy}$.

Theorem Angle between two regression lines.

Proof:- y on x regression line $y - \bar{y} = r \cdot \frac{\sigma_y}{\sigma_x}(x - \bar{x})$ — (1)

x on y regression line $x - \bar{x} = r \cdot \frac{\sigma_x}{\sigma_y}(y - \bar{y})$

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

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

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

in equation (1) $m_1 = r \cdot \frac{\sigma_y}{\sigma_x}$.

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

Angle between two regression lines

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

$$\tan\theta = \frac{r \cdot \frac{\sigma_y}{\sigma_x} - \frac{\sigma_y}{r \cdot \sigma_x}}{1 + \cancel{r} \cdot \frac{\sigma_y}{\sigma_x} \cdot \frac{1}{\cancel{r}} \cdot \frac{\sigma_y}{\sigma_x}}$$

$$= \frac{\frac{r^2 \sigma_y - \sigma_y}{r\sigma_x}}{\frac{r \cdot \sigma_x^2 + r \cdot \sigma_y^2}{r \cdot \sigma_x^2}} = \frac{\frac{\sigma_y [r^2 - 1]}{r \cdot \sigma_x}}{\frac{r[\sigma_x^2 + \sigma_y^2]}{r \cdot \sigma_x^2}}$$

$$= \frac{\sigma_y [r^2 - 1]}{r[\sigma_x^2 + \sigma_y^2]} \times \frac{\cancel{r} \cdot \sigma_x^2}{\cancel{r} \cdot \cancel{\sigma_x}}$$

$$\tan\theta = \frac{\sigma_x \cdot \sigma_y [r^2 - 1]}{r[\sigma_x^2 + \sigma_y^2]}$$

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

Q → Difference between Correlation and Regression.

Correlation	Regression
① If change in one variable effects to change in another variable. Then the two variables are said to be correlated. Then that relationship is called correlation.	① Regression means stepping back towards the average in which we have to take average relationship between two (or) more variables.

(2) There is no cause and effect relationship between the variables.	(2) There is cause and effect relationship between the variables.
(3) There is no origin and scale effect on correlation.	(3) There is no origin effect but there is scale effect on regression.
(4) In correlation $r_{xy} = r_{yx}$.	(4) In regression $b_{yx} \neq b_{xy}$.
(5) It is limited only to linear relationship between the variables.	(6) It has wide scope to study linear as well as non linear relationship between the variables.

Problems

① Fit two regression equations from the following data.

$x: 1 \quad 5 \quad 3 \quad 2 \quad 1 \quad 1 \quad 7 \quad 3$

$y: 6 \quad 1 \quad 0 \quad 0 \quad 1 \quad 2 \quad 1 \quad 5$.

Sol:- Two two regression lines are.

① y on x regression line $y-\bar{y} = r.\frac{\sigma y}{\sigma x}(x-\bar{x})$

② x on y regression line $x-\bar{x} = r.\frac{\sigma x}{\sigma y}(y-\bar{y})$

where $r = \frac{Cov(xy)}{\sigma x . \sigma y}$

$Cov(xy) = \frac{\Sigma x_i y_i}{n} - (\bar{x})(\bar{y})$, $\bar{x} = \frac{\Sigma x_i}{n}$

$\bar{y} = \frac{\Sigma y_i}{n}$.

$\sigma x = \sqrt{\frac{\Sigma x_i^2}{n} - (\bar{x})^2}$

$\sigma y = \sqrt{\frac{\Sigma y_i^2}{n} - (\bar{y})^2}$

x	y	$\Sigma x_i y_i$	Σx_i^2	Σy_i^2
1	6	6	1	36
5	1	5	25	1
3	0	0	9	0
2	0	0	4	0
1	1	1	1	1
1	2	2	1	4
7	1	7	49	1
3	5	15	9	25
23	16	36	99	68

$$\bar{x} = \frac{\sum x_i}{n} = \frac{23}{8}$$

$$\boxed{\bar{x} = 2.875}$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{16}{8}$$

$$\boxed{\bar{y} = 2}$$

$$Cov(xy) = \frac{\sum x_i y_i}{n} - (\bar{x})(\bar{y})$$

$$= \frac{36}{8} - (2.875)(2)$$

$$= 4.5 - 5.75$$

$$\boxed{Cov(xy) = -1.25}$$

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

$$= \sqrt{\frac{99}{8} - (2.875)^2}$$

$$= \sqrt{12.375 - 8.2656}$$

$$= \sqrt{4.1094}$$

$$\boxed{\sigma x = 2.0271}$$

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

$$= \sqrt{\frac{68}{8} - (2)^2}$$

$$= \sqrt{8.5 - 4}$$

$$= \sqrt{4.5}$$

$$\boxed{\sigma y = 2.1213}$$

$$r = \frac{Cov(xy)}{\sigma x \cdot \sigma y} = \frac{-1.25}{(2.0271)(2.1213)} = \frac{-1.25}{4.300}$$

$$\boxed{r = -0.2906}$$

y on x regression line $y - \bar{y} = r \cdot \frac{\sigma y}{\sigma x}(x - \bar{x})$

$$y - 2 = (-0.2906)\frac{2.1213}{2.0271}(x - 2.875)$$

$$y - 2 = \frac{-0.6164}{2.0271}(x - 2.875)$$

$$= (-0.3041)(x - 2.875)$$

$$y - 2 = -0.3041x + 0.8742 \Rightarrow y = -0.3041x + 0.8742 + 2$$

$$\boxed{y = -0.3041x + 2.8742}$$

x on y regression line

$$x-\bar{x} = r \cdot \frac{\sigma x}{\sigma y}(y-\bar{y})$$

$$x-2.875 = (-0.2906)\frac{2.0271}{2.1213}(y-2)$$

$$x-2.875 = \frac{-0.6164}{2.1213}(y-2)$$

$$x-2.875 = (-0.2906)(y-2)$$

$$x-2.875 = -0.2906y+0.5812$$

$$x = -0.2906y+0.5812+2.875$$

$$\boxed{x = -0.2906y+3.4562}$$

Problem ① Fit two regression equations from the data given below.

x :	30	32	35	40	45
y :	20	28	30	32	35

② Fit two regression equations from the following data

x :	2	4	5	6	8	11
y :	18	12	10	8	7	5

Problem ① The regression equations are $8x-10y+66=0$ and $40x-18y-214=0$ Variance of the variable x is 9. Then find Means of x and y and also find σ_y and r_{xy}.

Sol:- Given two regression equations are

$8x - 10y + 66 = 0$ —— ①

$40x - 18y - 214 = 0$ —— ②

① × 5 − ②

$$\begin{array}{r} \cancel{40x} - 50y + 330 = 0 \\ -\cancel{40x} \mp 18y \mp 214 = 0 \\ \hline -32y + 544 = 0 \end{array}$$

$y = \frac{544}{32} = 17$

$\boxed{\bar{y} = 17}$

Substitute $y = 17$ in equation ①

$8x - 10y + 66 = 0$

$8x - 10(17) + 66 = 0$

$8x - 170 + 66 = 0$

$8x = 104 \implies x = \frac{104}{8} = 13$

∴ means of x and y is $\boxed{\bar{x} = 13, \bar{y} = 17}$

① ⟹ $8x - 10y + 66 = 0$

$8x + 66 = 10y$

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

$\boxed{\therefore b_{yx} = \frac{8}{10}}$

② ⟹ $40x - 18y - 214 = 0$

$40x = 18y + 214$

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

$\boxed{b_{xy} = \frac{18}{40}}$

We know that $r = \sqrt{b_{yx} \cdot b_{xy}}$

$r = \sqrt{\left(\frac{8}{10}\right)\left(\frac{18}{40}\right)} = \sqrt{\frac{18}{50}}$

$\boxed{r = 0.6}$

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

We are given $\sigma_x^2 = 9$

$\sigma_x = 3$.

$\frac{8}{10} = (\cancel{0.6}^{0.2}) \cdot \frac{\sigma_y}{\cancel{3}}$

$\frac{8}{(10)(0.2)} = \sigma_y$.

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

$\therefore$ Variance of y is $\sigma_y^2 = 16$.

Problem (2) You are given below following information about advertising and sales.

	adv expenditure in (Lakhs Rs)	Sales (in Rs crores)
mean	20	100
S.D	5	12

correlation coefficient is 0.8.

(1) Calculate two regression lines.

(2) Find the likely sales when advertising expenditure is Rs 25 lakhs.

(3) What should be advertisement if the company wants to attain sales target of Rs 130 crores.

Sol:- We are given

$\bar{x} = 20, \quad \bar{y} = 100 \quad r = 0.8$

$\sigma x = 5, \quad \sigma y = 12$

y on x regression line $y - \bar{y} = r \cdot \frac{\sigma y}{\sigma x}(x - \bar{x})$

$$y - 100 = (0.8)\left(\frac{12}{5}\right)(x - 20)$$

$$y = 1.92x - 38.4 + 100$$

$$\boxed{y = 1.92x + 61.6}$$

(ii) To find likely sales (y) when adv expenditure $x = 25$.

To find y value we use y on x reg line

$$y = 1.92x + 61.6$$

$$y = 1.92(25) + 61.6$$

$$y = 48 + 61.6$$

$$y = 109.6 \text{ crores}$$

x on y regression line $x - \bar{x} = r \cdot \frac{\sigma x}{\sigma y}(y - \bar{y})$

$$x - 20 = (0.8)\left(\frac{5}{12}\right)(y - 100)$$

$$x = 0.3333y - 33 + 20$$

$$\boxed{x = 0.3333y - 13.33}$$

(iii) What should be adv (x) when company wants to attain sales target $y = 130$ crores

To estimate x value when $y = 130$ we use x on y regression line

$$x = 0.33y - 13.$$

If $y = 130 \Rightarrow x = (0.33)(130) - 13$

$$x = 42.9 - 13$$

$$\boxed{x = 29.9 \text{ crores}}.$$

Problem ② The following details of marks in Maths and Statistics of 2nd B.Tech Students in our College.

	Maths	Statistics
Average marks	39.5	49.5
S.D	10.8	16.8

The Correlation Coefficient between Maths and Statistics is 0.8 by using this find regression lines and Estimate marks in Maths when marks in Statistics is 52.

Sol:- let us denote marks in maths is x.
marks in st is y.

We are given $\bar{x} = 39.5$, $\bar{y} = 49.5$, $r = 0.8$
$\sigma_x = 10.8$, $\sigma_y = 16.8$

Two regression lines — y on x reg line

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

x on y regression line $x - \bar{x} = r \cdot \frac{\sigma_x}{\sigma_y}(y - \bar{y})$