MODULE - III

CORRELATION, REGRESSION, ESTIMATION

Definition: (Bivariate Data):- Let (X, Y) be a pair of two measured variables and takes a set of pair of values (x_i, y_i) ; $i = 1, 2, 3, \cdots n$ is known as a Bivariate distribution / Bivariate data

Definition: (Correlation):- Two measured variables x and y are said to be correlated, if the change in one variable affects the change in other variable.

Types of Correlation :-

(i) POSITIVE CORRELATION

(ii) NEGATIVE CORRELATION

(iii) UNCORRELATION

POSITIVE CORRELATION :- If two variables deviate in the same direction, then the correlation is said to be direct / positive.

Ex (i) Heights (x) and weights (y) of individual

(ii) Income (x) and Expenditure (y) of employees

NEGATIVE CORRELATION :- If 2 variables deviate in opposite direction, then correlation is said to be negative.

Ex (i) Praice (x) and Demand (y) of an item

(ii) Pressure (x) and volume (y) of ideal gas

UNCORRELATION :- If the change in one variable does not effects change in other variable, then the correlation is uncorrelation.

i.e

neither they are positively correlated nor negatively correlated.

SIMPLE, MULTIPLE CORRELATION :- If only two variables are studied, then it is a problem of simple correlation. When 3 / more variables are studied, it is a problem of multiple / partial correlation.

METHODS OF CORRELATION :-

(i) SCATTERED DIAGRAM

(ii) KARL PEARSON COEFFICIENT OF CORRELATION

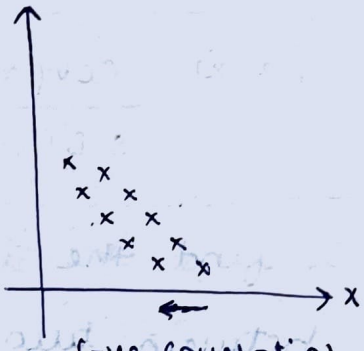
(iii) SPEARMAN RANK COEFFICIENT OF CORRELATION

SCATTERED DIAGRAM :-

8/8/17

(+ve correlation)

$r > 0$



(-ve correlation)

$r < 0$

uncorrelation

($r = 0$)

r = correlation b/w X and Y

→ Let (X, Y) be a bivariant data for set of n points (x_i, y_i), by taking the values of variable X along X-axis and values of Y along Y-axis.

→ Plot the points, then a dotted diagram is obtained, which is known as the scattered diagram.

KARL PEARSON COEFFICIENT OF CORRELATION :

$$r = (X, Y) = \frac{COV(X, Y)}{S.D(X) \cdot S.D(Y)}$$

→ To find the intensity of linear Relationship in between two measured variables X and Y, a mathematician discovered a formula known as Karl Pearson Coefficient of Correlation and it is denoted by $r(X, Y)$. i.e

$$r = (X, Y) = \frac{COV(X, Y)}{S.D(X) \cdot S.D(Y)}$$

$$r = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma (x - \bar{x})^2} \cdot \sqrt{\Sigma (y - \bar{y})^2}}$$ (for large values)

$$r = \frac{n \Sigma XY - \Sigma X \cdot \Sigma Y}{\sqrt{n \Sigma X^2 - (\Sigma X)^2} \cdot \sqrt{n \Sigma Y^2 - (\Sigma Y)^2}}$$ (for small values)

LIMITS :-

(i) Karl Pearson coefficient of correlation is always in between -1 and $+1$. i.e $-1 \leq r \leq 1$

If $-1 < r < 0$, then -ve correlation

If $0 < r < 1$, then +ve correlation

If $r = 0$, then uncorrelation

If $r = 1$, then perfect +ve correlation

If $r = -1$, then perfect -ve correlation

1. In a class test of two subjects, the marks obtained by ten students are given below:

Marks in Sub-I (x)	4	5	3	7	4	8	5	6	9	10
Marks in Sub-II (y)	6	7	2	8	9	6	3	5	8	10

Find K.P.C.C. Hence comment on the relation b/w marks in 2 subject

Sol:- $n = 10$

The following is the table

x	y	x^2	y^2	xy
4	6	16	36	24
5	7	25	49	35
3	2	9	4	6
7	8	49	64	56
4	9	16	81	36
8	6	64	36	48
5	3	25	9	15
6	5	36	25	30
9	8	81	64	72
10	10	100	100	100
$\Sigma x = 61$	$\Sigma y = 64$	$\Sigma x^2 = 421$	$\Sigma y^2 = 468$	$\Sigma xy = 422$

$$\therefore r = \frac{10(422) - (61)(64)}{\sqrt{10(421) - (61)^2}\sqrt{10(468) - (64)^2}}$$

$$= 0.5913$$

i.e $r > 0$

+ve correlation

Find K.P.C.C, for the following data

2.

Wage (x)	100	101	102	102	100	99	97	98	96	95
Cost of living (y)	98	99	99	97	95	92	95	94	90	91

Sol^r

$$r = \frac{\Sigma (x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma (x-\bar{x})^2} \cdot \sqrt{\Sigma (y-\bar{y})^2}}$$

$$\bar{x} = \frac{990}{10} = 99$$

$$\bar{y} = \frac{950}{10} = 95$$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
100	98	1	3	1	9
101	99	2	4	4	16
102	99	3	4	9	16
102	97	3	2	9	4
100	95	1	0	1	0
99	92	0	-3	0	9
97	95	-2	0	4	0
98	94	-1	-1	1	1
96	90	-3	-5	9	25
95	91	-4	-4	16	16
$\Sigma(x)=990$	$\Sigma(y)=950$	$\Sigma(x-\bar{x})=0$	0	= 54	= 96

$(x-\bar{x})(y-\bar{y})$ = 3 8 12 6 0 0 0 1 15 16 Σ = ~~99~~ 61

$$\therefore \quad r = \frac{61}{\sqrt{54}\sqrt{96}}$$

$$= \frac{61}{72}$$

$$= 0.847$$

NOTE :- Change of scale is independent.

3. Calculate the coefficient of coulation b/w age of cars and annual maintainance cost

Age of Cars (Year) (x)	2	4	6	7	8	10	12
Maintain (Cost) (y)	1600	1500	1800	1900	1700	2100	2000

Sol :- $r = \frac{\Sigma (x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma (x-\bar{x})^2}\sqrt{\Sigma (y-\bar{y})^2}}$

$$\bar{x} = \frac{49}{7} = 7$$

$$\bar{y} = \frac{12600}{7} = \frac{1800}{100} = 18//$$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
					100	100
2	1600	-5	-200	25	4	1000
4	1500	-3	-300	9	9	900
6	1800	-1	0	1	0	0
7	1900	0	100	0	1	0
8	1700	1	-100	1	1	-100
10	2100	3	300	9	9	900
12	2000	5	200	25	4	1000
				70	28	3700

$$r = \frac{3700}{\sqrt{70}\cdot\sqrt{28}}$$

$$= \frac{3700}{(8.36)(5.291)}$$

$$= \frac{3700}{44.236}$$

$$= 8.36//$$

(or)

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y}/100)}{\sqrt{\Sigma(x-\bar{x})^2}\cdot\sqrt{\Sigma\left(\frac{y-\bar{y}}{100}\right)^2}}$$

$$r = 0.836//$$

9/8/17

3) Rank Corelation :- (SPEARMAN's) :-

Let (x_i, y_i) be the ranks of n individuals for characteristics A and B respectively. Then rank corelation coefficient

$$\rho = \rho(X,Y) = \frac{1 - 6\Sigma d_i^2}{n(n^2-1)} \quad n \neq 1$$

where $d_i = x_i - y_i$

LIMITS :-

$-1 \leq \rho \leq 1$

If $-1 \leq \rho < 0 \Rightarrow$ Negative correlation

If $0 < \rho \leq 1 \Rightarrow$ Positive correlation

If $\rho = 0 \Rightarrow$ uncorrelation.

Repeated Ranks :- Add the below formula

$\frac{1}{12} m(m^2-1)$ to Σdi^2 where m = ~~repeated times~~ number of times an item is repeated. This adjustment factor is added to each repeated item.

1. Find the Rank correlation coefficient to the following data :-

Marks in Sub-I (X)	14	16	17	8	15	19	20
Marks in Sub-II (Y)	7	10	14	20	18	16	5

Sol :- $n = 7$

Rank C.C

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

x	y	xi	yi	$d_i = x_i - y_i$	d_i^2
14	7	6	6	0	0
16	10	4	5	-1	1
17	14	3	4	-1	1
8	20	7	1	6	36
15	18	5	2	3	9
19	16	2	3	-1	1
20	5	1	7	-6	36
					$\Sigma d_i^2 = 84$

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

$$= \frac{1 - 6(84)}{7(48)} = \frac{1 - 504}{336} = \frac{-503}{336} = -0.5$$

∴ Marks in both subjects are negatively correlated.

2. Obtain the rank C.C for the following data :-

x_i	68	64	75	50	64	80	75	40	55	64
y_i	62	58	68	45	81	60	68	48	80	70

Sol:- $n = 10$

X	Y	x_i	y_i	$x_i - y_i$	d_i^2
68	62	4	5	-1	1
64	58	$\frac{5+6+7}{3} = 6$	7	-1	1
75	68	$\frac{2+3}{2} = 2.5$	$\frac{3+4}{2} = 3.5$	-1	1
50	45	9	10	-1	1
64	81	$\frac{5+6+7}{3} = 6$	1	5	25
80	60	1	6	-5	25
75	68	$\frac{2+3}{2} = 2.5$	$\frac{3+4}{2} = 3.5$	-1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	$\frac{5+6+7}{3} = 6$	2	4	16
					$\Sigma d_i^2 = 72$

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

$$= \frac{1 - 6(72)}{10(99)} = 1 - \frac{2}{}$$

In X - series 75 repeated 2 times $\Rightarrow m_1 = 2$

" " " 64 " 3 " $\Rightarrow m_2 = 3$

In Y - series 68 " 2 " $\Rightarrow m_3 = 2$

$$\rho = 1 - \frac{6\left[\sum d_i^2 + \frac{1}{12}m_1(m_1^2-1) + \frac{1}{12}m_2(m_2^2-1) + \frac{1}{12}m_3(m_3^2-1)\right]}{n(n^2-1)}$$

$$= 1 - \frac{6\left[72 + \frac{1}{12}2(3) + \frac{1}{12}3(8) + \frac{1}{12}2(3)\right]}{10 \times 99}$$

$$= 1 - \frac{6\left[72 + \frac{6}{12} + \frac{24}{12} + \frac{6}{12}\right]}{990}$$

$$= 1 - \frac{6\left[72 + \frac{36}{12}\right]}{990}$$

$$= 1 - \frac{6[75]}{990}$$

$$= -\frac{5 \times 75}{990}$$

$$= -\frac{5[864 + 36]}{12 \times 990}$$

$$= -\frac{5[900]}{12 \times 990} = \frac{25}{66}$$

$$= 0.378$$

$$1 - \left\{\frac{6[864 + 36]}{990 \times 12}\right\}$$

$$1 - 6\left\{\frac{10}{990 \times 12}\right\}2$$

$$1 - \frac{10}{22}$$

H.W.

4. Find the rank correlation coefficient for the following.

x :	48	33	40	9	16	16	65	24	16	57
y :	13	13	24	6	15	4	20	9	16	19

Sol :- $n = 10$

x	y	x_i	y_i	$x_i - y_i$	d_i^2
48	13	3	$\frac{6+7}{2} = 6.5$	-3.5	12.25
33	13	5	$\frac{6+7}{2} = 6.5$	-1.5	2.25
40	24	4	1	3	9
9	6	10	9	1	1
16	15	$\frac{7+8+9}{3} = 8$	5	3	9
16	4	$\frac{7+8+9}{3} = 8$	10	-2	4
65	20	1	2	-1	1
24	9	6	8	-2	4
16	16	$\frac{7+8+9}{3} = 8$	4	4	16
57	19	2	3	-1	1

$\Sigma d_i^2 = 59.5$

In x series 16 repeated 3 times $\Rightarrow m_1 = 3$

In y series 13 repeated 2 times $= m_2 = 2$

$$\rho = 1 - \frac{6\left[\Sigma di^2 + \frac{1}{12} m_1 [m_1^2 - 1] + \frac{1}{12} m_2 [m_2^2 - 1]\right]}{10(10^2 - 1)}$$

$$\rho = 1 - \frac{6\left[59.5 + \frac{1}{12} 3(8) + \frac{1}{12} 2(3)\right]}{10(99)}$$

$$\rho = 1 - \frac{6\left[59.5 + \frac{24}{12} + \frac{6}{12}\right]}{990}$$

$$\rho = 1 - \frac{6\left[59.5 + \frac{30}{12}\right]}{990}$$

$$\rho = 1 - \frac{6\left[\frac{714 + 30}{12}\right]}{990 \times 12}$$

$$\rho = 1 - \frac{6\left[744\right]}{990 \times 12}$$

$$\rho = 1 - \frac{4464}{11880}$$

$$\rho = \frac{7416}{11880}$$

$$\rho = 0.624 //$$

∴ Marks of x and y are positively correlated.

10/8/17

5.* Ten competitors in a musical test were ranked by the three judges A, B, C in the following order.

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

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

Sol: $n = 10$; given that here 10 competitors are there.

A (x)	B (y)	C (z)	$d_1 = x-y$	$d_2 = x-z$	$d_3 = y-z$	d_1^2	d_2^2	d_3^2
1	3	6	-2	-5	-3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	-3	-4	-1	9	16	1
10	4	8	6	2	-4	36	4	16
3	7	1	-4	2	6	16	4	36
2	10	2	-8	0	8	64	0	64
4	2	3	2	1	-1	4	1	1
9	1	10	8	-1	-9	64	1	81
7	6	5	1	2	1	1	4	1
8	9	7	-1	1	2	1	1	4
						200	60	214

∴ rank correlation = $\rho = 1 - \frac{6 \Sigma d_i^2}{n(n^2-1)}$

by (A, B) $\Rightarrow 1 - \frac{6(200)}{10(990)}$

$= 1 - \frac{1200}{990}$

$= \frac{990 - 1200}{990}$

$= \frac{-210}{990}$

$= \frac{-21}{99}$

$= \frac{-7}{33}$

$= -0.212$

$=$

by (A, C) $= 1 - \frac{6 \Sigma d_2^2}{n(n^2-1)}$

$= 1 - \frac{6(60)}{990}$

$= 1 - \frac{360}{990} = \frac{990 - 360}{990} = \frac{630}{990} = \frac{21}{33}$

$= \frac{7}{11}$

$= 0.636$

by (B, C) $= 1 - \frac{6 \Sigma d_3^2}{990}$

$= 1 - \frac{6(214)}{990}$

$= \frac{990 - 1284}{990} = \frac{-294}{990}$

$= -0.296$

$=$

Since, $\rho(x, z)$ is maximum, we conclude that the pair of judges A and C has nearest approach to common likings in music.

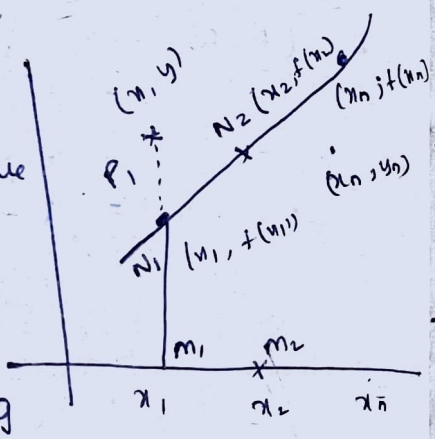
14/8/17

Curve fitting :-

→ Curve fitting is the method of finding the equation of a curve, that approximates a given set of data

→ Let the curve $y = a_0 + a_1 x + \cdots + a_m x^m$ be the fitted to the set of n data points (x_1, y_1), $(x_2, y_2) \cdots (x_n, y_n)$.

→ Here, we have to determine the constants $a_0, a_1, \ldots a_m$ such that it represents the curve of best fit



→ This can be possible using the following methods

(i) graphical method (X)

(ii) distribution method (X)

(iii) principle of least squares method ✓

PRINCIPLE OF LEAST SQUARE METHOD:-

This method is the most systematic procedure to fit a unique curve through the given data points.

→ Let $y = f(x)$, be the equation of the curve to be fitted to the given data points (x_1, y_1), (x_2, y_2), $(x_3, y_3) \cdots (x_n, y_n)$.

→ At $X = x_1$, the observed value of the ordinate P_1M_1 is y_1. and the corresponding value of the fitting curve is $N_1M_1 = f(x_1)$

$$\therefore \quad P_1N_1 = P_1M_1 - M_1N_1 = e_1$$

$$P_2N_2 = P_2N_2 - M_2N_2 = e_2$$

$$\vdots \qquad\qquad \vdots$$

$$P_nN_n = \cdots = e_n$$

→ Here some of the errors are positive and some of the errors are negative.

→ To make all errors positive, we square, each of these errors.

$$\therefore \quad E = e_1^2 + e_2^2 + \cdots + e_n^2$$

$$= \sum_{i=1}^{n} e_i^2$$

→ The curve of the best fit is that ~~to~~ which "THE SUM OF THE SQUARES OF THESE ERRORS IS MINIMUM", this is called principle of least squares.

Fitting of a straight line: Let $y = a + bx$ be a straight line.

To find unknown a, b we use method of least square.

$$E = \sum_{i=1}^{n} e_i^2$$

$$= \sum_{i=1}^{n} [y_i - f(x_i)]^2$$

$$\boxed{\sum_{i=1}^{n} (y_i - a - bx_i)^2} \longrightarrow (1)$$

The necessary conditions to minimum the error are $\frac{\partial E}{\partial a} = 0$; $\frac{\partial E}{\partial b} = 0$

$$\therefore \frac{\partial E}{\partial a} = 0 \Rightarrow \frac{\partial}{\partial a} \left(\sum_{i=1}^{n} (y_i - a - bx_i)^2 \right) = 0$$

$$= \sum_{i=1}^{n} 2(y_i - a - bx_i)(-1) = 0$$

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a - \sum_{i=1}^{n} bx_i = 0$$

$$\Rightarrow \boxed{\sum_{i=1}^{n} y_i = a \cdot n + b \cdot \sum_{i=1}^{n} x_i} \longrightarrow (2)$$

i.e $\frac{\partial E}{\partial b} = 0 = \frac{\partial}{\partial b} \left(\sum_{i=1}^{n} (y_i - a - bx_i)^2 \right) = 0$

$$= \frac{\partial}{\partial b} \left(\sum_{i=1}^{n} 2(y_i - a - bx_i)(-x_i) = 0 \right.$$

$$\Rightarrow \sum_{i=1}^{n} (y_i x_i - a x_i - b x_i^2) = 0$$

$$\Rightarrow \boxed{\sum_{i=1}^{n} y_i x_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2} \rightarrow (3)$$

Here (2) and (3) are known as normal equations of (1)

on solving (2) x (3) we get a, b which gives the best fit.

Fitting of a parabola :- Let $y = a + bx + cx^2$ be a parabola.

To find unknown a, b, c we used method of least squares.

$$E = \sum_{i=1}^{n} e_i^2$$

$$= \boxed{\sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)^2} \rightarrow (1)$$

The necessary conditions to minimise the errors are $\frac{\partial E}{\partial a} = 0$; $\frac{\partial E}{\partial b} = 0$; $\frac{\partial E}{\partial c} = 0$.

$$\frac{\partial E}{\partial a} = 0 \Rightarrow \frac{\partial}{\partial a} \left(\sum_{i=1}^{n} y_i - a - bx_i - cx_i^2 \right)^2 = 0$$

$$= \sum_{i=1}^{n} 2(y_i - a - bx_i - cx_i^2)(-1) = 0$$

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a - b \sum_{i=1}^{n} x_i - c \sum_{i=1}^{n} x_i^2 = 0$$

$$\Rightarrow \boxed{\sum_{i=1}^{n} y_i = a \cdot n + b \sum_{i=1}^{n} x_i + c \sum_{i=1}^{n} x_i^2} \rightarrow (2)$$

i.e $\frac{\partial \epsilon}{\partial b} = 0$

$\Rightarrow \frac{\partial}{\partial b} \left(\sum_{i=1}^{n} y - a - bxi - cxi^2 \right)^2 = 0$

$\sum_{i=1}^{n} 2(y - a - bxi - cxi^2)(-xi) = 0$

$\Rightarrow \sum_{i=1}^{n} yx_i - a \sum_{i=1}^{n} xi - b \sum_{i=1}^{n} xi^2 - c \sum_{i=1}^{n} xi^3$

$\Rightarrow$ $\boxed{\sum_{i=1}^{n} yx_i = a \sum_{i=1}^{n} xi + b \sum_{i=1}^{n} xi^2 + c \sum_{i=1}^{n} xi^3}$ → 3

i.e $\frac{\partial \epsilon}{\partial c} = 0$

$\Rightarrow \frac{\partial}{\partial c} \left(\sum_{i=1}^{n} y - a - bxi - cxi^2 \right)^2 = 0$

$\sum_{i=1}^{n} 2(y - a - bxi - cxi^2)(-xi^2) = 0$

$\Rightarrow \sum_{i=1}^{n} yxi^2 - a \sum_{i=1}^{n} xi^2 - b \sum_{i=1}^{n} xi^3 - c \sum_{i=1}^{n} xi^4 = 0$

$\therefore$ $\boxed{\sum_{i=1}^{n} yxi^2 = a \sum_{i=1}^{n} xi^2 + b \sum_{i=1}^{n} xi^3 + c \sum_{i=1}^{n} xi^4}$

→ (4)

Here (2) (3) and (4) are known as normal eqns of

(1) on solving (2)(3)(4) we get a, b, c which gives best fit for parabola

16/8/19

Fitting of a Power Curve (or) Geometric Function :-

Let $y = ax^b$ be a power curve $\longrightarrow$ (1)

Apply $\log_{10}$ on both sides

$$\log_{10} y = \log_{10} ax^b$$

$$\log_{10} y = \log_{10} a + \log_{10} x^b$$

$$\log_{10} y = \log_{10} a + b \log_{10} x$$

Now $Y = \log_{10} y$

$A = \log_{10} a$

$X = \log_{10} x$

$\therefore Y = A + bX \longrightarrow (2)$

$\rightarrow$ It is in the form of a st line

Normal Equation of (2) is

$$\sum_{i=1}^{n} Y_i = A \cdot n + b \sum_{i=1}^{n} x_i \longrightarrow (3)$$

$$\sum_{i=1}^{n} Y_i x_i = A \cdot \sum_{x_i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 \longrightarrow (4)$$

$\therefore$ By considering 3 and 4 we get A, b

But $a = 10^A$ (= antilog A)

Fitting of an exponential function :-

Let the exponential function is $y = ae^{bx}$

Applying $\log_{10} \Rightarrow \log_{10} y = \log_{10} a + \log_{10} e^{bx}$

$\Rightarrow \log_{10} y = \log_{10} a + bx \log_{10} e$

$$Y = A + B \cdot x \longrightarrow (2)$$

where $Y = \log_{10} y$

$A = \log_{10} a$

$B = b \log_{10} e$

(2) is a straight line

∴ Hence Normal Equations are

$$\sum_{i=1}^{n} y_i = A \cdot n + B \cdot \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} y_i x_i = A \cdot \sum_{i=1}^{n} x_i + B \cdot \sum_{i=1}^{n} x_i^2$$

On solving the both equations we get A, B

∴ $a = 10^A$ (= antilog A)

$b = \dfrac{B}{\log_{10} e}$

1. Find the best values of a & b so that $y = a + bx$ fits the data given in the table:

x	0	1	2	3	4
y	1.0	2.9	4.8	6.7	8.6

Sol: given $y = a + bx \longrightarrow (1)$

whose normal equations are

$$\sum_{i=1}^{n} y_i = a \cdot n + b \sum_{i=1}^{n} x_i \longrightarrow (2)$$

$$\sum_{i=1}^{n} y_i x_i = a \cdot \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 \longrightarrow (3)$$

x	y	x^2	xy
0	1.0	0	0
1	2.9	1	2.9
2	4.8	4	9.6
3	6.7	9	20.1
4	8.6	16	34.4
$\Sigma x_i = 10$	$\Sigma y_i = 24.0$	$\Sigma x^2 = 30$	$\Sigma xy = 67$

$$54.5$$
$$9.6$$
$$2.9$$
$$0$$

$\therefore$ (2) $\Rightarrow$ $\Sigma y_i = 24 = 5a + b \cdot 10$ (2)

(3) $\Rightarrow$ $67 = 10a + 30b$

$$48 = 10a + 20b$$
$$67 = 10a + 30b$$
$$-19 = -10b$$
$$b = -19/10$$

(2)
48
19

$\therefore a = 1$

$y = 1 + (1.9)x$ is the best fit straight line to given data.

2. Find a st line to fit the following data and find y at $x = 12$

x	1	2	3	4	5
y	5	7	9	10	11

Sol:- given st line $y = a + bx \longrightarrow (1)$

Normal eq'ns are

$$(2) = \sum_{i=1}^{n} y_i = a \cdot n + b \cdot \sum_{i=1}^{n} x_i$$

$$(3) = \sum_{i=1}^{n} y_i x_i = a \cdot \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2$$

x	y	x^2	xy
1	5	1	5
2	7	4	14
3	9	9	27
4	10	16	40
5	11	25	55
$\Sigma x_i = 15$	$\Sigma y_i = 42$	$\Sigma x_i^2 = 55$	$\Sigma xy = 141$

30

55

95
19
27
141

$\therefore$ (2) $\Rightarrow$ $42 = a\,5 + b \cdot 15$ (11)

(3) $\Rightarrow$ $141 = a \cdot 55 + b \cdot 55$

$462 = 55a + 165b$

(−) $141 = 55a + 55b$

$321 = +\ 110b$

$b = \frac{321}{110} \Rightarrow b = 2.9$

$\therefore\ 42 - \frac{321}{110}(15) = 5a$

$\Rightarrow$ ~~4620 − 321~~

$\frac{4620 - 4815}{110 \times 5} = a$

$\frac{195}{550} = a \Rightarrow a = 0.354$

(2) $\Rightarrow$ $42 = 5a + 15b$

(3) $\Rightarrow$ $141 = 15a + b \cdot 55$

$126 = 15a + 45b$

(−) $141 = 15a + 55b$

$-15 = -10b$

$b = \frac{3}{2} = 1.5$

$$42 = 5a + 15\left(\frac{3}{2}\right)$$

$$42 = 5a + \frac{45}{2}$$

$$42 - \frac{45}{2} = 5a$$

$$\frac{84 - 45}{10} = a$$

$$3.9 = a$$

$$y = a + bx$$

$$y = 3.9 + 1.5(12)$$

$$y = 3.9 + 18$$

$$y = 21.9$$

H.W

3. Find the best values of a, b, c ; so that $y = a + bx + cx^2$, fits the data

x	0	1	2	3	4
y	-4	-1	4	11	20

$y = x^2 + 2x - 4$

H.W

4. Find a, b, c so that $y = a + bx + cx^2$ fits the data

x	0	1	2	3	4
y	0	1.8	1.3	2.5	6.3

$y = 0.534 - 0.298x + 0.407x^2$

3.

x	0	1	2	3	4
y	-4	-1	4	11	20

given $y = a + bx + cx^2 \longrightarrow (1)$

Normal equations are

$(2) \Rightarrow \Sigma y_i = a.n + b\,\Sigma x_i + c\,\Sigma x_i^2$

$(3) \Rightarrow \Sigma x_i y_i = a.\Sigma x_i + b\,\Sigma x_i^2 + c\,\Sigma x_i^3$

$(4) \Rightarrow \Sigma x_i^2 y_i = a.\Sigma x_i^2 + b\,\Sigma x_i^3 + c\,\Sigma x_i^4$

x	y	x^2	x^3	x^4	xy	x^2y
0	-4	0	0	0	0	0
1	-1	1	1	1	-1	-1
2	4	4	8	16	8	16
3	11	9	27	81	33	99
4	20	16	64	256	80	320
$\Sigma x = 10$	$\Sigma y = 30$	$\Sigma x^2 = 30$	$\Sigma x^3 = 100$	$\Sigma x^4 = 354$	$\Sigma xy = 120$	$\Sigma x^2y = 434$

∴ The equations we will be getting are as follows

$(2) \Rightarrow 30 = a.5 + b.10 + c.30$

$(3) \Rightarrow 120 = a.10 + b.30 + c.100$

$(4) \Rightarrow 434 = a.30 + b.100 + c.354$

$$\Rightarrow \begin{pmatrix} 5a + 10b + 30c = 30 \\ 10a + 30b + 100c = 120 \end{pmatrix} \Rightarrow \times 2$$

$$\Rightarrow \begin{array}{r} 10a + 20b + 60c = 60 \\ (-)\quad (-) \quad (-) \quad (-) \\ 10a + 30b + 100c = 120 \\ \hline -10b - 40c = -60 \end{array}$$

$$10b + 40c = 60 \longrightarrow (i)$$

$(5a + 10b + 30c = 30) \times 6$

$30a + 100b + 354c = 434$

$\Rightarrow 30a + 60b + 180c = 180$

$(-) \quad (-) \quad (-) \quad (-)$

$30a + 100b + 354c = 434$

$-40b - 174c = -254$

$\Rightarrow 40b + 174c = 254 \longrightarrow (ii)$

Solving (i) & (ii) we get

$(10b + 40c = 60) \times 4$

$40b + 174c = 254$

$\Rightarrow 40b + 160c = 240$

$(-) \quad (-) \quad (-)$

$40b + 174c = 254$

$-14c = -14$

$\boxed{c = 1}$

$\therefore 40b + 160 = 240$

$40b = 240 - 160 \Rightarrow 80 \Rightarrow \boxed{b = 2}$

$\therefore 5a + 20 + 30 = 30 \Rightarrow 5a = -20 \Rightarrow \boxed{a = -4}$

$\therefore$ The ~~eqn~~ best fit obtained is $y = -4 + 2x + x^2$

4. Sol:-

x :	0	1	2	3	4
y :	0	1.8	1.3	2.5	6.3

Here given $y = a + bx + cx^2$ (parabola) where

$n = 5$

Normal eqn's $\rightarrow$ (2) $= \Sigma y_i = an + b\Sigma x_i + c\Sigma x_i^2$

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

$\Sigma x_i^2 y_i = a\Sigma x_i^2 + b\Sigma x_i^3 + c\Sigma x_i^4$

x	y	x^2	x^3	x^4	xy	x^2y
0	0	0	0	0	0	0
1	1.8	1	1	1	1.8	1.8
2	1.3	4	8	16	2.6	5.2
3	2.5	9	27	81	7.5	22.5
4	6.3	16	64	256	25.2	100.8
$\Sigma x = 10$	$\Sigma y = 11.9$	$\Sigma x^2 = 30$	$\Sigma x^3 = 100$	$\Sigma x^4 = 354$	$\Sigma xy = 37.1$	$\Sigma x^2y = 130.3$

The equations obtained will be are as follows:

$(2) = 11.9 = 5.a + b.10 + c.30$

$(3) = 37.1 = 10a + b.30 + c.100$

$(4) = 130.3 = 30a + b.100 + c.354$

$$\Rightarrow \begin{aligned} 5a + 10b + 30c &= 11.9 \\ 10a + 30b + 100c &= 37.1 \end{aligned} \Rightarrow \begin{aligned} 10a + 20b + 60c &= 23.8 \\ (-)\ 10a + 30b + 100c &= 37.1 \\ \hline -10b - 40c &= -13.3 \end{aligned}$$

$$\Rightarrow \begin{aligned} 5a + 10b + 30c &= 11.9 \\ 30a + 100b + 354c &= 130.3 \end{aligned} \Rightarrow \begin{aligned} 30a + 60b + 180c &= 71.4 \\ (-)\ 30a + 100b + 354c &= 130.3 \\ \hline -40b - 174c &= -58.9 \end{aligned}$$

$$\begin{aligned} -40b - 160c &= -53.2 \\ (-)\ -40b - 174c &= -58.9 \\ \hline +14c &= 5.7 \end{aligned}$$

$$\boxed{c = 0.407}$$

$$-40b + 65.12 = 53.2 \Rightarrow -40b = -11.92 \Rightarrow \boxed{b = -0.298}$$

$$\Rightarrow 5a - 2.98 + 12.21 = 11.9$$

$$5a + 9.23 = 11.9 \Rightarrow \boxed{a = 0.534}$$

∴ The best fit obtained is

$$y = 0.534 - 0.298x + 0.407x^2$$

5. Fit a parabola $y = a + bx + cx^2$ with the data in table

x :	10	12	15	23	20
y :	14	17	23	25	21

Sol:- given its parabola i.e $y = a + bx + cx^2 \rightarrow (1)$

$$\bar{x} = \frac{10 + 12 + 15 + 23 + 20}{5} = 16$$

$$\bar{y} = \frac{14 + 17 + 23 + 25 + 21}{5} = 20$$

$$X = x - \bar{x} = x - 16$$

$$Y = y - \bar{y} = y - 20$$

∴ Parabola in X, Y is $Y = a + bX + cX^2$

Normal Eqns:- $\sum y_i = a \cdot n + b \sum x_i + c \sum x_i^2 \rightarrow (3)$

$$\sum y_i x_i = a \sum x_i + b \sum x_i^2 + c \sum x_i^3 \rightarrow (4)$$

$$\sum y_i x_i^2 = a \sum x_i^2 + b \sum x_i^3 + c \sum x_i^4 \rightarrow (5)$$

x	y	$X = x - 16$	$Y = y - 20$	x^2	x^3	x^4	xy	yx^2
10	14	-6	-6	36	-216	+1296	36	-216
12	17	-4	-3	16	-64	256	12	-48
15	23	-1	3	1	-1	1	-3	3
23	25	7	5	49	343	2401	35	245
20	21	4	1	16	64	256	4	16
80	100	0	0	118	126	4210	84	0

∴ ⑶ ⇒ 0 = 5a + 0·b + 118·c

⑷ ⇒ 84 = a·0 + 118b + 126c

⑸ ⇒ 0 = a·118 + b·126 + 4210c

$5a + 118c = 0$

$118b + 126c = 84$

$118a + 126b + 4210c = 0$

$a = \frac{-118c}{5}$

b.

$a = -\frac{118c}{5}$

$-\frac{(118)(118)}{c} + 126b + 4210c = 0$

⇒

∴ On solving we get;

$a = 1.67$

$b = 0.79$

$c = -0.07$

⑵ ⇒ $y = 1.67 + 0.79x - (0.07)x^2$

⇒ $y - 20 = 1.67 + 0.79(x-16) - 0.07(x-16)^2$

6. H.W Fit a straight line to the following data

Years (x) :	1961	1971	1981	1991	2001
Production (y) :	8	10	12	10	16

Sol:- The straight line equation obtained is

$$y = a + bx \longrightarrow (1)$$

$X = x - \bar{x}$ $\Rightarrow \bar{x} = 1981$

x	y	$(X)/10$	X^2	Xy
1961	8	-20	400	-16
1971	10	-10	100	-10
1981	12	0	0	0
1991	10	10	100	10
2001	16	20	400	32
	56	0	10	16

(2) $\Rightarrow \Sigma y_i = a \cdot n + b \Sigma X_i$

$\Rightarrow 56 = 5a + b \cdot 0$

$\Rightarrow 56 = 5a$

$\boxed{a = 11.2}$

(3) $\Rightarrow \Sigma X_i y_i = a \cdot \Sigma X_i + b \Sigma X_i^2$

$16 = a \cdot 0 + b \cdot 10$

$\boxed{b = 1.6}$

By solving with the normal equations, the values obtained of a and b are 11.2 & 1.6 respectively.

The best fit obtained is as follows

$\therefore y = a + bx \Rightarrow y = 11.2 + 1.6x$

$\Rightarrow y = 11.2 + 1.6\left(\frac{x - 1981}{10}\right)$

NOTE :- Fitting of Exponential Curve ($y = ab^x$) :-

Consider of $y = ab^x \longrightarrow (1)$

Apply $\log_{10}$ on b.s

$\log_{10} y = \log_{10} a + x \log_{10} b$

$Y = A + x \cdot B \longrightarrow (2)$

where $Y = \log_{10} y$; $A = \log_{10} a$; $B = \log_{10} b$

(2) is a straight line, whose normal.

$\Sigma Y_i = A \cdot n + B \Sigma x_i$

$\Sigma Y_i x_i = A \cdot \Sigma x_i + B \Sigma x_i^2$

on solving we get A, B

finally $a = 10^A$ (or) $a =$ Antilog A

$b = 10^B$ (or) $b =$ Antilog B

1. An experiment gave the following values.

u :	350	400	500	600
t :	61	26	7	2·6

Its known that u & t are connected by relation $u = at^b$. Find the best possible values of ~~u and t~~ a & b.

Sol:- given $u = at^b \longrightarrow (1)$

Apply $\log_{10}$ on b.s

$$\log_{10} u = \log_{10} a + b \log_{10} t$$

$$U = A + b \cdot T \longrightarrow (2)$$

where $U = \log_{10} u$; $A = \log_{10} a$; $T = \log_{10} t$

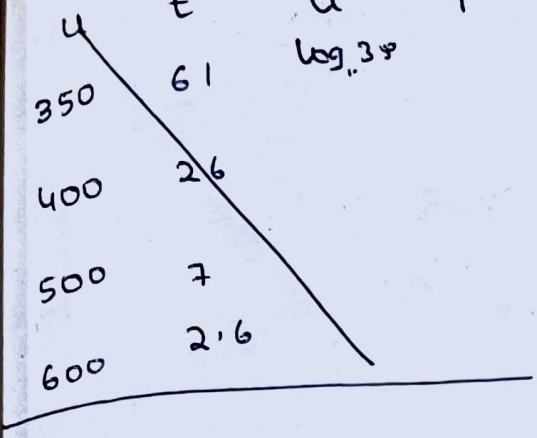
(2) is st line ; so the normal equn's are

$$\Sigma U_i = A \cdot n + b \Sigma T_i \longrightarrow (3)$$

$$\Sigma U_i T_i = A \Sigma T_i + b \Sigma T_i^2 \longrightarrow (4)$$

Here $n = 4$

$$\bar{u} = u - \bar{u} = \frac{1850}{4} = 1400$$



u	t	$U = \log_{10} u$	$T = \log_{10} t$	T^2	UT
350	61	2.5441	1.7853	3.187	
400	26	2.6021	1.4150		
500	7	2.69890	0.8451		
600	2.6	2.7782	0.4150		
		$\Sigma U = 10.6234$	$\Sigma T = 4.4604$	$\Sigma T^2 = 6.075$	$\Sigma UT = 11.658$

(3) $\Rightarrow$ $10.6234 = A \cdot 4 + b \cdot 4.4604$

(4) $\Rightarrow$ $11.658 = A(4.4604) + b \cdot (6.075)$

$\Rightarrow$ $4A + 4.4604b = 10.6234$

$4.4604A + 6.075b = 11.658$

Solving we get $A = 2.845$

$b = -0.1697$

$a = 10^A$ / a = antilog A = antilog (2.845) = 699.8 //

$b = -0.1697$

$\therefore$ (1) $\Rightarrow$ $u = (6.99 \cdot 8)\, t^{-0.1697}$

2. Fit an exponential curve of form $y = ab^x$ to the following data

x	1	2	3	4	5	6	7	8
y	1.0	1.2	1.8	2.5	3.6	4.7	6.6	9.1

Sol: Given $y = ab^x \longrightarrow (1)$

Apply $\log_{10}$ on b.s

$$\log_{10} y = \log_{10} a + x \log_{10} b$$

$$Y = A + Bx \longrightarrow (2)$$

where $Y = \log_{10} y$; $A = \log_{10} a$; $B = \log_{10} b$

∴ Normal eqn's are

$$\Sigma Y_i = A \cdot n + B \cdot \Sigma x_i \longrightarrow (3)$$

$$\Sigma Y_i x_i = A \Sigma x_i + B \Sigma x_i^2 \longrightarrow (4)$$

x	y	Y	x^2	$Y_i x_i$
1	1.0	0	1	0
2	1.2	0.0791	4	
3	1.8	0.2552	9	
4	2.5	0.3980	16	
5	3.6	0.5564	25	
6	4.7	0.6720	36	
7	6.6	0.8195	49	
8	9.1	0.9590	64	
36		3.7892	204	22.7385

$\therefore$ (3) $\Rightarrow$ $3.7393 = A \cdot 8 + B \cdot 36$

(4) $\Rightarrow$ $22.7385 = A \cdot 36 + B \cdot 204$

$\Rightarrow$ $8A + 36B = 3.7393$

$36A + 204B = 22.7385$

$A = -0.1662$

$= \bar{1}.8338$

$B = 0.1408$

$b =$ antilog $B = 1.383$

$a =$ " $A = 0.6821$

$y = (0.6821) \cdot (1.38)^x$

19/8/17

Regression :- It is the estimation or prediction of one variable (unknown values) from known values of another variable.

→ Regression means " Stepping back towards the average ".

→ Regression analysis, is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data.

Linear Regression :- If the two variables X and Y are correlated, then the scattered diagram will be more or less concentrated along a curve. This curve is called the "CURVE OF REGRESSION".

→ If the curve is a straight line, it is called Line of Regression. and regression is said to be linear.

Linear Regression Equations :-

Let $y = a + bx$ be a straight line $\longrightarrow (1)$

whose normal eqn's are $\Sigma y_i = a \cdot n + b \Sigma x \longrightarrow (2)$

$\Sigma y_i x_i = a \cdot \Sigma x + b \Sigma x^2 \longrightarrow (3)$

From (2) $\Sigma y = a \cdot n + b \Sigma x$.

Dividing by $n \Rightarrow \quad \Sigma y/n = a + b \Sigma x/n$

$\bar{y} = a + b \cdot \bar{x} \longrightarrow (4)$

from $1 - 4 \Rightarrow \quad y - \bar{y} = b(x - \bar{x}) \longrightarrow (5)$

From (3) $\Sigma yx = a \Sigma x + b \Sigma x^2$

On changing origin to $(\bar{x}, \bar{y})$ we get

$\Sigma (y - \bar{y})(x - \bar{x}) = a \Sigma (x - \bar{x}) + b \Sigma (x - \bar{x})^2$

$\Rightarrow \Sigma (y - \bar{y})(x - \bar{x}) = b \Sigma (x - \bar{x})^2$

$\therefore \left(\Sigma (x - \bar{x}) = 0 \right)$

Since deviations are taken from actual means, in that case $\Sigma (x - \bar{x}) = 0$; $\Sigma (y - \bar{y}) = 0$

$\Rightarrow b = \dfrac{\Sigma (x-\bar{x})(y-\bar{y})}{\Sigma (x-\bar{x})^2}$ $\qquad$ S.D $\sigma_x = \sqrt{\dfrac{\Sigma(x-\bar{x})^2}{n}}$

because of S.D $\Rightarrow \sigma_x = \sqrt{\dfrac{\Sigma(x-\bar{x})^2}{n}}$

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

$\therefore$ ~~$\Sigma (x-\bar{x})(y-\bar{y})$~~

$b = \dfrac{\Sigma (x-\bar{x})(y-\bar{y})}{n \cdot \sigma_x^2}$

$= \dfrac{\Sigma (x-\bar{x})(y-\bar{y})}{n \cdot \sigma_x^2} \times \dfrac{\sigma_y}{\sigma_y}$ (Multiply & divide by σ_y)

$= \dfrac{\Sigma (x-\bar{x})(y-\bar{y}) \cdot \times \sigma_y}{n \cdot \sigma_x \cdot \sigma_y \times \sigma_x}$

$b = r \cdot \dfrac{\sigma_y}{\sigma_x}$ $\qquad \left(\because r = \dfrac{\Sigma (x-\bar{x})(y-\bar{y})}{n \cdot \sigma_x \cdot \sigma_y} \right)$

$\therefore$ b in (5)

$\Rightarrow \boxed{y - \bar{y} = r \cdot \dfrac{\sigma_y}{\sigma_x} (x - \bar{x})}$ is line of linear regression of y on x.

$\Rightarrow \boxed{(x - \bar{x}) = r \cdot \dfrac{\sigma_x}{\sigma_y} (y - \bar{y})}$ of x on y.

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

(or) $r \cdot \frac{\sigma_x}{\sigma_y} = b_{xy}$

Here b_{yx} is called regression coefficient of y on x

and b_{xy} is called regression coefficient of x on y

$\therefore$ (6) becomes :- $y - \bar{y} = b_{yx}(x - \bar{x}) \longrightarrow 6'$

$x - \bar{x} = b_{xy}(y - \bar{y}) \longrightarrow 6'$

NOTE: 1 If $r = 0$, then the two lines of regression become $y = \bar{y}$ and $x = \bar{x}$ respectively which are two straight lines parallel to x-axis and y-axis respectively and passing through their means

NOTE: 2 If $r = \pm 1$, the two lines of regression will coincide.

Properties of Regression Coefficients

(i) Correlation coefficient is the geometric mean between the two regression coefficients.

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

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

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

$r = \sqrt{b_{yx} \cdot b_{xy}}$ // Geometric Mean of Regression coefficients

(ii) If one of the regression coefficient is greater than unity, then other must be less than one. ~~because~~ :-

Proof :- Let $b_{yx} > 1$,

$\Rightarrow \frac{1}{b_{yx}} < 1$

we know that $-1 \leq r \leq 1$

$\Rightarrow r^2 \leq 1$

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

but $r^2 \leq 1$

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

$\Rightarrow b_{xy} \leq \frac{1}{b_{yx}}$

but $\frac{1}{b_{yx}} < 1$

$\therefore b_{xy} \leq \frac{1}{b_{yx}} < 1$

$b_{xy} < 1$

$\therefore b_{yx} > 1 \Rightarrow b_{xy} < 1$

(iii) The correlation coefficient and the two regression coefficients have the same sign.

$\Rightarrow$ Since σ_x, σ_y are always +ve, then r, b_{yx} b_{xy} have same signs

(iv) A.M of regression coefficients is greater than the correlation coefficient.

Proof :

Since

$$(\sigma_x - \sigma_y)^2 > 0$$

$$\Rightarrow \sigma_x^2 + \sigma_y^2 - 2\sigma_x \cdot \sigma_y > 0$$

$$\Rightarrow \frac{\sigma_x^2 + \sigma_y^2 > 2\sigma_x\sigma_y}{\sigma_x \cdot \sigma_y}$$

$$\Rightarrow \frac{\sigma_x}{\sigma_y} + \frac{\sigma_y}{\sigma_x} > 2$$

Multiply whole with r

$$r \cdot \frac{\sigma_x}{\sigma_y} + r \cdot \frac{\sigma_y}{\sigma_x} > 2r$$

$$b_{xy} + b_{yx} > 2r$$

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

To prove

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

$$b_{xy} + b_{yx} > 2r$$

$$r\frac{\sigma_x}{\sigma_y} + r\cdot\frac{\sigma_y}{\sigma_x} > 2r$$

$$r\frac{\sigma_x^2 + r\sigma_y^2}{\sigma_x\sigma_y} > 2r$$

$$r\sigma_x^2 + r^2\sigma_y^2 > 2r\sigma_x\sigma_y$$

$$r^2\sigma_y^2 - 2r\sigma_x\sigma_y + \sigma_x^2 > 0$$

Since $(\sigma_x - r\sigma_y)^2 > 0$ is true

Hence $\frac{b_{xy} + b_{yx}}{2} > r$

(i) Angle between two regression lines

(ii) Correlation v/s Regression.

$$\left(\text{AM} > \frac{a+b}{2}\right)$$

PROBLEMS

21/8/16

1. Find the two lines of regressions y on x and x on y for the following data

x :	3	5	6	8	9	11
y :	2	3	4	6	5	8

Further estimate value of y when $x = 2$ and also estimate value of x when $y = 10$

$$\begin{array}{r|l} 6 & 28 \\ & 24 \\ \hline & 40 \end{array}$$

Sol: We know that

Line of regression of y on $x \Rightarrow y - \bar{y} = b_{yx}(x - \bar{x})$

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

$\bar{x} = 7$; $\bar{y} = 4.66$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
3	2	-4	-2.66	16	7.0756	10.64
5	3	-2	-1.66	4	2.7556	3.32
6	4	-1	-0.66	1	0.4356	0.66
8	6	1	1.34	1	1.7956	1.34
9	5	2	0.34	4	0.1156	0.65
11	8	4	3.34	16	11.1556	13.36
$\Sigma x = 42$	$\Sigma y = 28$	0	0	42	23.3336	29.97

$$\sigma_x = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}} = \sqrt{\frac{42}{6}} = \sqrt{7} = 2.645$$

$$\sigma_y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n}} = \sqrt{\frac{23.3336}{6}} = \sqrt{3.8893} = 1.987.$$

~~$r = \frac{\Sigma xy}{n \cdot \sigma_x \cdot \sigma_y}$~~

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n \cdot \sigma_x \cdot \sigma_y} = \frac{29.97}{6 \times (2.645)(1.987)}$$

$$= 0.9504$$

∴ Regression coefficient of y on $x \Rightarrow$

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

$$b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x} = 0.9504 \left(\frac{1.987}{2.645}\right) = 0.7139$$

Regression of x on y $\quad b_{xy} = r \cdot \frac{\sigma_x}{\sigma_y}$

$$= 0.9504 \times \frac{2.64}{1.987}$$

$$= 1.265$$

∴ Line of regression of y on x is

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

i.e

$$y - 4.66 = (0.7139)(x - 7)$$

$$y = 0.7139x - 0.3786 \longrightarrow (1)$$

Line of regression of x on y is

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

i.e

$$(x - 7) = 0.9504(1.265)(y - 4.66)$$

i.e $\quad x = 1.265y + 1.0589 \longrightarrow (2)$

for

(i) when $x = 2$ from (1)

$$y = 1.4278 - 0.3786$$

$$y = 1.0492$$

(ii) when $y = 10$ from (2)

$$x = 12.65 + 1.0589$$

$$x = 13.7089$$

$\Rightarrow$

2. Calculate correlation coefficient and least square regression line of y on x and x on y, also obtain an estimate of y which should correspond on the average to $x = 6.2$ for following data

x :	1	2	3	4	5	6	7	8	9
y :	9	8	10	12	11	13	14	16	15

Sol: We have y on x as $(y-\bar{y}) = b_{yx}(x-\bar{x})$ &

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

$$\bar{x} = \frac{45}{9} = 5 \; ; \quad \bar{y} = \frac{108}{9} = 12$$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
1	9	-4	-3	16	9	12
2	8	-3	-4	9	16	12
3	10	-2	-2	4	4	4
4	12	-1	0	1	0	0
5	11	0	-1	0	1	0
6	13	1	1	1	1	1
7	14	2	2	4	4	4
8	16	3	4	9	16	12
9	15	4	3	16	9	12
$\Sigma x = 45$	$\Sigma y = 108$	$\Sigma \bar{x} - \bar{x} = 0$	$\Sigma y - \bar{y} = 0$	60	60	57

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

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

$$= \sqrt{\frac{60}{9}} = \sqrt{6.666}$$

$$\boxed{\sigma_x = 2.5818}$$

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

$$= \sqrt{\frac{60}{9}}$$

$$\boxed{\sigma_y = 2.5818}$$

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

$$= \frac{57}{9 \times (2.5818)(2.5818)}$$

$$\boxed{r = 0.9501}$$

x on $y \Rightarrow b_{xy} = r \cdot \frac{\sigma_x}{\sigma_y}$

$$= 0.9501 \times 1$$

$$\boxed{b_{xy} = 0.9501}$$

y on $x \Rightarrow b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x}$

$$= 0.9501 \times 1$$

$$\boxed{b_{yx} = 0.9501}$$

Line of regression ~~of~~ y on x is

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

$$(y-12) = 0.9501(x-5)$$

$$y - 12 = 0.9501x - 4.7505$$

$$y = 0.9501x - 4.7505 + 12$$

$y = 0.9501x + 7.2495$

Line of regression of x on y is

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

$x - 5 = 0.9501(y - 12)$

$x - 5 = 0.9501y - 11.4012$

$x = 0.9501y - 11.4012 + 5$

$x = 0.9501y - 6.4012$

$y = 0.95(6.12) + 7.2495$

$= 5.89 + 7.2495$

$y = 13.14$

22/8/17

3. The two lines of regression of the variables x & y are $8x - 10y + 66 = 0$ and $40x - 18y - 214 = 0$. Then find

(i) Correlation coefficient b/w x and y

(ii) Mean of x and Mean of y.

Sol: Line of regression of y on x given by

$$y = \frac{8}{10}x + \frac{66}{10} \longrightarrow (1)$$

Line of regression of x on y given by

$$x = \frac{18}{40}y + \frac{214}{40} \longrightarrow (2)$$

from (1) $b_{yx} = 8/10 = 0.8$

(2) $b_{xy} = \frac{18}{40} = 0.45$

Correlation coefficient $r = \sqrt{byx \cdot bxy}$

$$r = \sqrt{0.8 \times 0.45}$$

$$= \sqrt{0.36}$$

$$r = \pm 0.6 //$$

* Since byx and bxy are positive, then r is 0.6

(ii) Mean of x & Mean of Y

* Regression lines always pass through their mean.

Solving $\Rightarrow$ $8\bar{x} - 10\bar{y} + 66 = 0 \longrightarrow (1)$

$40\bar{x} - 18\bar{y} - 214 = 0$

Multiply 1 by 5

$$\begin{array}{r} 40\bar{x} - 50\bar{y} + 330 = 0 \\ (-)\quad (+) \quad (+) \\ 40\bar{x} - 18\bar{y} - 214 = 0 \\ \hline -32\bar{y} + 544 \end{array}$$

$\bar{y} = 17$

On substituting $\bar{y}$; we get $\bar{x} = 13$

∴ The Mean of x is $\bar{x} = 13$.

The Mean of y is $\bar{y} = 17$.

Estimation :-

(i) **Sampling** :- Sampling is simply the process of learning about the population on the basis of a sample drawn from it.

→ The theory of sampling is a study of relationship existing between population and samples drawn from population.

→ The fundamental object of sampling is to get as much information as possible about the population by examining only a part of it.

Parameter and Statistic :- The statistical constants of the population such as mean (μ), standard deviation (σ) are called the **parameters**.

Similarly, these statistical constants for the sample drawn from the given population such as mean ($\bar{x}$) and standard deviation (s) are called **statistic**.

→ The aim of sampling is to gather the maximum information about the population.

NOTE : The population may be finite / infinite.

Estimate :- The statistic whose distribution concentrates as closely as possible near the true value of the parameter regarded as the estimate

(OR)

An estimate is a statement made to find an unknown population parameter.

Estimator :- The procedure ~~to d~~ or rule to determine an unknown population parameter is called an estimator.

Example :- Sample mean ($\bar{x}$) is an estimator of population mean (μ), because sample mean is a method of determinig the population mean.

NOTE :- A parameter can have one / ~~mor~~ two / many estimators.

Basically, there are two kinds of estimates to determine the statistic of the population parameter namely

(i) POINT ESTIMATION
(ii) INTERVAL ESTIMATION

(i) Point Estimation :- If an estimate of a population parameter is given by a single value, then the estimate is called point estimate of parameter.

(ii) Interval Estimation :- If an estimate of population parameter is given by two different values between which, the parameter may be considered to lie, then the estimate called an interval estimate of the parameter.

Example :- If the height of student is measured as 162 cm, then measurement gives a point estimate. But if the height is given as 163 ± 3.5 cm i.e (159.5, 166.5) then the height lies between 159.5 cm and 166.5cm and the measurement gives an interval estimate.

23/8/17

STATISTICAL INFERENCE :- Statistical Inference is the process by which we draw a conclusion about some measure of a population based on a sample value.

→ The measure might be a variable such as mean, standard deviation etc.

→ There are two types of problems under statistical inference :-

(i) Hypothesis Testing :- To test some hypothesis about parent population from which the sample is drawn.

(ii) Estimation :- To use the statistics obtained from sample, as estimate of unknown parameter of the population from which sample is drawn.

* Types of Sampling :- Some of the commonly known and frequently used types of sampling are :-

(a) PURPOSIVE SAMPLING
(b) RANDOM SAMPLING
(c) SIMPLE SAMPLING
(d) STRATIFIED SAMPLING

PURPOSIVE SAMPLING :- Its one in which, the sample units are selected with definite purpose in view.

Example :- If we want to give the picture that, the standard of living has increased in the city of Hyderabad, we may take individuals in sample from rich & posh localities like Banjara Hills, Jubilee Hills, Gachibouli etc and ignore the localities where low income troop live.

RANDOM SAMPLING :- The sample units are selected at random.

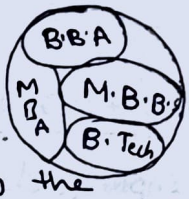
NOTE : Random sampling is applicable when characteristic of each sample are similar.

Example :- We can measure some quantity such as mean, standard deviation etc of B.Tech students but we can't measure some quantity of M.B.B.S and B.Tech students because both characteristics are heterogenous (different).

SIMPLE SAMPLING :- Its a RANDOM SAMPLING in which each unit of the population has an equal chance.

NOTE :- RANDOM SAMPLING doesn't imply SIMPLE SAMPLING but converse is true.

STRATIFIED SAMPLING :- Here, the entire heterogenous population is divided into a number of homogenous groups, termed as 'strata', which differ from one another but each of these groups is homogenous within itself.



→ The sample which is the aggregate of the sampled units of each of the stratum, is termed as stratified sample and technique of drawing this sample is known as stratified sample.

8/1/17

SAMPLING DISTRIBUTION OF STATISTIC :- If we draw a sample, of size 'n' from a given finite population of size 'N', then the total number of possible ~~n~~ samples are

$$N_{C_n} = \frac{N!}{(N-n)!\,n!} = K \text{ (say)}$$

→ For each of these ~~case~~ samples, we can compute some statistic $t = t(x_1, x_2, x_3, \cdots x_K)$; $\bar{x}$, s^2 in particular the mean $\bar{x}$, the variance s^2 etc as given below:

Sample Number	Statistics		
	t	$\bar{x}$	s^2
1	t_1	$\bar{x}_1$	s_1^2
2	t_2	$\bar{x}_2$	s_2^2
3	t_3	$\bar{x}_3$	s_3^2
⋮	⋮	⋮	⋮
K	t_K	$\bar{x}_K$	s_K^2

The set of the values of the statistic, so obtained, one for each sample constitutes, which is called the sampling distribution of the sample.

STANDARD ERROR:- The standard deviation of the sampling distribution of a statistic is known as its standard error (S.E).

→ The standard errors of the well-known statistics for large samples are given below, where

n - sample size

σ^2 - population variance

P - population proportion

$Q = 1 - P$

A part of a population with a particular attribute, expressed as a fraction, decimal or percentage of the whole population.

Ex:- For example, let's say you had 1,000 people in the population and 237 of those people have blue eyes. The fraction of people who have blue eyes is 237 out of 1000. i.e $P = \frac{237}{1000} = 0.237$

STATISTIC	S.E
1. Sample mean ($\bar{x}$)	$(\sigma/\sqrt{n})$
2. Sample S.D (s)	$\sqrt{\frac{\sigma^2}{2n}}$
3. Sample variance (s^2)	$\sigma^2\sqrt{\frac{2}{n}}$

REMARK:- Standard error plays a very important role in large sample theory and forms the basis of testing of hypothesis.