Unit - III

STATISTICS

* Measures of Central Tendency:

1) Arithmetic Mean / Mean
2) Median
3) Mode

A) For ungrouped data :-

Suppose there are 'n' observations say $x_1, x_2, \cdots, x_n$ in data

1) A.M. ($\bar{x}$) :-

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x_i}{n}$$

2) Median :-

Arrange data in ascending or descending order

i) If n = odd,

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{observation}$$

ii) If n = even,

$$\text{Median} = \frac{(n/2)^{th}\text{obse} + (n/2+1)^{th}\text{obs.}}{2}$$

3) Mode :- Most repeated observation in the data

B) For discrete / ungrouped frequence distribution :-

x_i	x_1	x_2	$\cdots$	x_n
f_i	f_1	f_2	$\cdots$	f_n

1) A.M. ($\bar{x}$) :-

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_n}{f_1 + f_2 + \cdots + f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{N} \quad ; \quad N = \sum_{i=1}^{n} f_i$$

2) Median :-

x_i	f_i	l.c.f.
x_1	f_1	f_1
x_2	f_2	f_1+f_2
x_3	f_3	$f_1+f_2+f_3$
⋮	⋮	⋮
x_n	f_n	$f_1+\cdots+f_n=N$

Median = Value of x for which l.c.f is just greater that $\frac{N}{2}$

3) Mode :- Value of x is greatest frequency.

c) For continuous / grouped frequency distribution :-

Class	a_1-a_2	a_2-a_3	a_3-a_4	...	a_n-a_{n+1}
f_i	f_1	f_2	f_3	...	f_n
Midvalue x_i	$x_1=\frac{a_2+a_1}{2}$	$x_2=\frac{a_2+a_3}{2}$	x_3	...	x_n

1) A.M. ($\bar{x}$) :-

$$\bar{x} = \frac{x_1f_1+x_2f_2+\cdots+x_nf_n}{f_1+f_2+\cdots+f_n} = \frac{\sum_{i=1}^{n} f_ix_i}{N=\Sigma f_i}$$

2) Median :-

$$\text{Median} = L + \frac{h}{f}\left[\frac{N}{2} - \text{l.c.f.}\right]$$

where :-

Median Class = Class for which l.c.f is just greater $\frac{N}{2}$

L = Lower limit of median class

h = class width of median class

f = frequency of median class

$l.c.f.$ = l.c.f. of premedian class

3) Mode :-

$$\text{Mode} = l + h\left(\frac{f_m - f_1}{2f_m - f_1 - f_2}\right)$$

where

Modal class = class with highest frequency in the distribution

l = lower limit of modal class

h = class width of modal class

f_m = frequency of modal class

f_1 = frequency of premodal class

f_2 = frequency of postmodal class.

* Examples :-

1) The monthly income (in Rs.) of 10 families in a village is as follows

5200, 5000, 5100, 5250, 4950, 5300, 5350, 5150, 5200, 5110

find A.M, median, mode.

→ 1) $\bar{x} = \frac{\Sigma x_i}{10} = \frac{51610}{10} = 5161$

2) Median :-

4950, 5000, 5100, 5110, 5150, 5200, 5200, 5250, 5300, 5350

(5th and 6th observations)

here $n = 10$

$$\text{Median} = \frac{(n/2)^{th}\text{ obs} + (n/2+1)^{th}\text{ obs}}{2} = \frac{(5^{th}+6^{th})\text{ obs}}{2} = \frac{5150+5200}{2}$$

$= 5175$ //

3) Mode = 5200 //

2) Calculate A.M, Median, Mode for the following frequency distribution

x	5	6	7	8	9	10
f	8	10	9	6	5	4

→ i) A.M. :

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{(5\times8)+(6\times10)+(7\times9)+(8\times6)+(9\times5)+(10\times4)}{8+10+9+6+5+4}$$

$$= \frac{296}{42}$$

$$\bar{x} = 7.04$$

ii) Median:-

x_i	f_i	l.c.f.
5	8	8
6	10	18
7	9	27 ← $\frac{N}{2} = 21$
8	6	33
9	5	38
10	4	42 = N

here $\frac{N}{2} = 21$

∴ Median = 7 //

iii) Mode = 6 //

3) Compute A.M, Median & Mode for the frequency distribution

Profit (Rs) Per Shop	0-100	100-200	200-300	300-400	400-500
No. of shops	10	18	27	13	12

→

Class	f_i	x_i (midvalue)	l.c.f.	
0-100	10	50	10	
100-200	18	150	28	← Premedian & Premodal class
200-300	27	250	55	← Median class & Modal class
300-400	13	350	68	← Post modal class
400-500	12	450	80 = N	

$$\bar{x} = \frac{\Sigma f_i x_i}{N} = 248.75$$

2) Median :-

Median class = 200 - 300

$L = 200$

$f = 27$

$h = 100$

$lc.f = 28$

$$\text{Median} = L + \frac{h}{f}\left[\frac{N}{2} - lc.f\right]$$

$$= 200 + \frac{100}{27}\left[40 - 28\right]$$

$$\text{Median} = 244.45$$

3) Mode :- $L = 200$, $f_m = 27$, $f_1 = 18$, $f_2 = 13$, $h = 100$.

$$\text{Mode} = l + h\left[\frac{f_m - f_1}{2f_m - f_1 - f_2}\right] = 200 + 100\left[\frac{27 - 18}{54 - 18 - 13}\right] = 239.13$$

* Remark :-

1) (Mean - Mode) = 3 (Mean - Median)

2) Combined Mean :-

Let $x_1, x_2, \cdots, x_{n_1}$ & $y_1, y_2, \cdots, y_{n_2}$ be two set of data then their combined mean is,

$$\bar{z} = \frac{\sum_{i=1}^{n_1} x_i + \sum_{i=1}^{n_2} y_i}{n_1 + n_2} = \frac{n_1 \bar{x} + n_2 \bar{y}}{n_1 + n_2}$$

* Measures of Dispersion :-

Standard Deviation (σ) :-

1) For ungrouped data :-

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

or

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

2) For discrete or continuous frequency distribution :-

$$\sigma = + \sqrt{\frac{\sum f_i (x_i - \bar{x})^2}{N}} \quad ; \quad N = \sum f_i$$

OR

$$\sigma = + \sqrt{\frac{\sum f_i x_i^2}{N} - (\bar{x})^2} \quad ; \quad \bar{x} = \frac{\sum f_i x_i}{N}$$

* Remarks :-

1) Variance = $(S.D.)^2 = \sigma^2$

2) Short cut method :-

A) Arithmetic Mean

i) For ungrouped data :-

Put $d = x_i - A$; A = assumed mean

$\Rightarrow \bar{d} = \bar{x} - A$ ✓

$\Rightarrow \bar{x} = \bar{d} + A$

ii) For grouped data :-

Put $u = \frac{x_i - A}{h}$

$\Rightarrow \bar{x} = A + h\,\bar{u}$

B) Standard Deviation :-

i) $d = x_i \pm A$

$\Rightarrow S.D.[x_i] = S.D.[d]$

ii) & $S.D.[kx] = k\, S.D.[x]$

Put $u = \frac{x - A}{h} = \frac{d}{h}$

$\Rightarrow d = uh$

$\therefore S.D.[d] = S.D.[uh]$

$= \sqrt{\frac{\sum f h^2 u^2}{N} - (h\bar{u})^2}$

$= h\ S.D.[u]$

Ex:- Compute S.D. for the following data.

15, 18, 22, 25, 10

$\rightarrow$ $\bar{x} = \frac{\Sigma x_i}{5} = \frac{90}{5} = 18.$

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

$= \sqrt{\frac{(15^2+18^2+22^2+25^2+10^2)}{5} - (18)^2}$

S.D. $= 5.25$ //

2) Calculate S.D. for, the following data.

class	0-10	10-20	20-30	30-40	40-50	50-60
f_i	5	9	15	12	10	3

$\rightarrow$

Class	f_i	x_i	$u_i = \frac{x_i - A}{h} = \frac{x_i - 25}{10}$	$f_i u_i$	$f_i u_i^2$
0-10	5	5	-2	-10	20
10-20	9	15	-1	-9	9
20-30	15	25	0	0	0
30-40	12	35	1	12	12
40-50	10	45	2	20	40
50-60	3	55	3	9	27
	54			22	108

$\therefore$ S.D.[x] $= h$ S.D.[u]

$= 10 \sqrt{\frac{\Sigma f_i u_i^2}{N} - \left(\frac{\Sigma f_i u_i}{N}\right)^2} = 10 \sqrt{\frac{108}{54} - \left(\frac{22}{54}\right)^2}$

$= 13.54$ //

$$\bar{x} = A + h\bar{u}$$

$$= 25 + 10\left(\frac{22}{54}\right)$$

$$\bar{x} = 29.07$$ //

* Coefficient of Variation :-

$$C.V. = \frac{\sigma}{A.M.} \times 100 = \frac{S.D.}{A.M.} \times 100$$

Generally it is used to define consistency of the data or variability of the data.

Ex. If A & B are two set of data with

$$(C.V.)(A) < (C.V.)(B)$$

$\Rightarrow$ set A is more consistent than B

or set B shows more variability than A.

* Examples :-

1) The scores of two cricketers A & B for 6 matches each are given below, find who is more consistent & whose performance is better.

P1 A :	58	59	60	54	65	66
P1 B :	84	56	92	65	86	44

$\frac{354+66}{2}$ $\frac{120}{2}$

→ i) For player A :

x_i	$d_i = x_i - 60$	d_i^2
58	-2	4
59	-1	1
60	0	0
54	-6	36
65	5	25
66	6	36
	2	102

$$\therefore \bar{x} = A + \bar{d} = A + \frac{\Sigma d_i}{n}$$

$$= 60 + \frac{2}{6} = 60 + 0.3333$$

$$\bar{x}_A = 60.3333$$

$$\sigma_A = +\sqrt{\frac{\Sigma d_i^2}{n} - \left(\frac{\Sigma d_i}{n}\right)^2} = +\sqrt{\frac{102}{6} - \left(\frac{2}{6}\right)^2}$$

$$\sigma_A = 4.11$$

$$\therefore (C.V.)_A = \frac{\sigma_A}{\bar{x}_A} \times 100 = \frac{4.11}{60.3333} \times 100$$

$$(C.V.)_A = 6.8122$$

2) For player B.

y_i	$d_i' = y_i - 65$	$(d_i')^2$
84	19	361
56	-9	81
92	27	729
65	0	0
86	21	441

44	$\frac{-21}{37}$	$\frac{441}{2053}$

$$\therefore \bar{y}_B = 65 + \frac{37}{6} = 71.1667$$

$$\sigma_B = +\sqrt{\frac{\Sigma d_i^2}{n} - \left(\frac{\Sigma d_i}{n}\right)^2}$$

$$= +\sqrt{\frac{2053}{6} - \left(\frac{37}{6}\right)^2}$$

$$\sigma_B = 17.4395$$

$$\therefore (C.V.)_B = \frac{\sigma_B}{\bar{y}_B} \times 100 = \frac{17.4395}{71.1667}$$

$$(C.V.)_B = 24.50$$

$$\Rightarrow \quad (C.V.)_A < (C.V.)_B$$

$\Rightarrow$ Player A is more consistent than B //

And as

$$\bar{x}_A < \bar{x}_B$$

$\Rightarrow$ Player B, it's performance is better than Player A.

Correlation :

- simultaneous variation

i.e. change in one variable causes change in another variable.

Types :-

i) +ve correlation : increase (/decrease) ⟶ increase (/ decrease)

ii) -ve correlation : increase (/ decrease) ⟶ decrease (/ increase)

Examples :-

i) Rainfall → crop output ⇒ +ve correlation / -ve correlation

ii) income → expenditure ⇒ +ve correlation

iii) price → demand of commodity ⇒ -ve correlation

Measures of Correlation :-

* Karl Pearson's Coefficient of Correlation (r) :-

Correlation coefficient between two variables x & y is denoted by $r(x,y)$ & defined as,

$$r = \frac{cov(x,y)}{\sigma_x \, \sigma_y}$$

where $cov(x,y)$ = covariance of x & y

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

(Discrete data)

$$= \frac{1}{N} \sum f_i (x_i - \bar{x})(y_i - \bar{y})$$

(Frequency distribution)

Deviation Method

Put $u_i = x_i - A$ or $\frac{x_i - A}{h}$

If $v_i = y_i - B$ or $\frac{y_i - B}{h}$

then

$$\boxed{r(x,y) = r(u,v)}$$

i.e. $r(x,y)$ is invariant under change of origin or change of scale property

* Correlation coefficient in simplified form :-

i) For ungrouped data :-

$$r(x,y) = r(u,v) = \frac{n\sum u_i v_i - \sum u_i \sum v_i}{\sqrt{n\sum u_i^2 - (\sum u_i)^2}\sqrt{n\sum v_i^2 - (\sum v_i)^2}}$$

OR

$$r(u,v) = \frac{\sum u_i v_i - n\bar{u}\bar{v}}{\sqrt{\sum u_i^2 - n(\bar{u})^2}\sqrt{\sum v_i^2 - n(\bar{v})^2}}$$

ii) For Discrete / Continuous Frequency Distribution :-

$$r(x,y) = r(u,v) = \frac{N\sum f_i u_i v_i - (\sum f_i u_i)(\sum f_i v_i)}{\sqrt{N\sum f_i u_i^2 - (\sum f_i u_i)^2}\sqrt{N\sum f_i v_i^2 - (\sum f_i v_i)^2}}$$

OR

$$r(u,v) = \frac{\sum f_i u_i v_i - N\bar{u}\bar{v}}{\sqrt{\sum f_i u_i^2 - N(\bar{u})^2}\sqrt{\sum f_i v_i^2 - N(\bar{v})^2}}$$

IMP

* Remarks :-

$$-1 \le r(x,y) \le 1$$

* Examples :-

1) Compute correlation coefficient between supply & price of commodity using following data.

Supply	152	158	169	182	160	166	182
Price	198	178	167	152	180	170	162

→ Correlation coefficient is given by,

$$r(x,y) = r(u,v) = \frac{n\sum u_i v_i - \sum u_i \sum v_i}{\sqrt{n\sum u_i^2 - (\sum u_i)^2}\sqrt{n\sum v_i^2 - (\sum v_i)^2}}$$

Table :

	Supply(x)	Price(y)	x*x	y*y	x*y	u=x-166(A)	v=y-170(B)	u*u	v*v	u*v
	152	198	23104	39204	30096	-14	28	196	784	-392
	158	178	24964	31684	28124	-8	8	64	64	-64
	169	167	28561	27889	28223	3	-3	9	9	-9
	182	152	33124	23104	27664	16	-18	256	324	-288
	160	180	25600	32400	28800	-6	10	36	100	-60
	166	170	27556	28900	28220	0	0	0	0	0
	182	162	33124	26244	29484	16	-8	256	64	-128
Total	1169	1207	196033	209425	200611	7	17	817	1345	-941

here $n = 7$.

$r(x,y) = r(u,v)$

$$\therefore \quad r = \frac{7(-941) - (7)(17)}{\sqrt{7(817) - (7)^2}\sqrt{7(1345) - (17)^2}}$$

$$\boxed{r = -0.9322}$$

H.W.

1) Calculate correlation coefficient for following distribution

x	5	9	15	19	24	28	32
y	7	9	14	21	23	29	30
f	6	9	13	20	16	11	7

Ans: $r = 0.9825$

* Examples on Correlation :-

1) Given: $n = 6, \sum(x - 18.5) = -3, \sum(y - 50) = 20, \sum(x - 18.5)(y - 50) = -120, \sum(x - 18.5)^2 = 19, \sum(y - 50)^2 = 850$. Calculate coefficient of correlation.

→ Suppose u & v are deviations of x & y values from 18.5 & 50 respectively. $[u = x - 18.5,\ v = y - 50]$

∴ Given: $n = 6,\ \sum u_i = -3,\ \sum v_i = 20,\ \sum u_i v_i = -120$

$$\sum u_i^2 = 19,\ \sum v_i^2 = 850$$

Since.

$$r(x,y) = r(u,v)$$

$$= \frac{n\sum u_i v_i - \sum u_i \sum v_i}{\sqrt{n\sum u_i^2 - (\sum u_i)^2}\ \sqrt{n\sum v_i^2 - (\sum v_i)^2}}$$

$$= \frac{6(-120) - (-3)(20)}{\sqrt{6(19) - (-3)^2}\ \sqrt{6(850) - (20)^2}}$$

$$r = -0.9395 //$$

$r = \dfrac{\sum u_i v_i - n\bar{u}\bar{v}}{\sqrt{\sum u_i^2 - n(\bar{u})^2}\ \sqrt{\sum v_i^2 - n(\bar{v})^2}}$

$\bar{u} = \dfrac{\sum u_i}{n} = \dfrac{-3}{6} = -0.5$

$\bar{v} = \dfrac{\sum v_i}{n} = \dfrac{20}{6} = 3.3333$

2) Given: $r = 0.9, \sum xy = 60, \sigma_x = 3, \sum y^2 = 100$. Find the number of items if x and y are deviations of u and v from arithmetic mean.

→ Given

$$r = 0.9,\ \sum xy = 60,\ \sigma_x = 3,\ \sum y^2 = 100$$

also $x = u - \bar{u}$ & $y = v - \bar{v}$

$$r(u,v) = \frac{cov(u,v)}{\sigma_u \sigma_v} = \frac{\sum (u - \bar{u})(v - \bar{v})}{n\,\sigma_u \sigma_v}$$

$$r = \frac{\sum xy}{n\,\sigma_x \sigma_y}$$

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

$= \sqrt{\frac{\Sigma y_i^2}{n}}$

$\sigma_y^2 = \frac{\Sigma y_i^2}{n} = \frac{100}{n}$

$\therefore \quad r^2 = \frac{(\Sigma xy)^2}{n^2 \sigma_x^2 \sigma_y^2}$

$(0.9)^2 = \frac{(60)^2}{n^2 (3)^2 \left(\frac{100}{n}\right)}$

$n = \frac{(60)^2}{(0.9)^2 (3)^2 (100)} =$

$= 4.938$

$n \approx 5$ //

* Regression Line :-

A) Regression line of y on x :-

Consider the set of values (x_i, y_i) ; $i = 1, 2, \cdots, n$ & let the line of regression of y on x be $y = mx + c$

Then it is given by,

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

OR

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

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

= regression coefficient of y on x

OR

$$(y_i - \bar{y}) = \frac{cov(x,y)}{\sigma_x^2} (x_i - \bar{x})$$

$\because r = \frac{cov(x,y)}{\sigma_x \sigma_y}$

Note:- Slope of the regression line of y on x is,

$$m = b_{yx} = r \frac{\sigma_y}{\sigma_x} = \frac{cov(x,y)}{\sigma_x^2}$$

B) Regression Line of x on y :-

Consider the set of values (x_i, y_i) ; $i = 1, 2, \cdots, n$ & let the line of regression of x on y be $x = ny + b$

Then it is given by,

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

or

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

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

= regression coefficient of x on y

or

$$(x_i - \bar{x}) = \frac{cov(x,y)}{\sigma_y^2} (y_i - \bar{y})$$

$\because \quad r = \frac{cov(x,y)}{\sigma_x \sigma_y}$

Note:- Slope of the line x on y is

$$n = b_{xy} = r \frac{\sigma_x}{\sigma_y} = \frac{cov(x,y)}{\sigma_y^2}$$

* Remark:-

1) As $b_{yx} = r \frac{\sigma_y}{\sigma_x}$, $b_{xy} = r \frac{\sigma_x}{\sigma_y}$

$\Rightarrow \quad r^2 = b_{xy} b_{yx} \leqslant 1$ Imp to decide regression lines

$$\Rightarrow \quad r = \begin{cases} +\sqrt{b_{yx}\, b_{xy}} & ; \text{ when both } b_{yx}, b_{xy} > 0 \\ -\sqrt{b_{yx}\, b_{xy}} & ; \text{ when both } b_{yx}, b_{xy} < 0 \end{cases}$$

2) The intersection of the two regression lines is the point

$$(\bar{x}, \bar{y})$$

3) If θ is the angle between two regression lines then

$$m_1 = b_{yx} \quad , \quad m_2 = \frac{1}{b_{xy}}$$

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

$$\tan\theta = \frac{(1 - r^2)\, \sigma_x \sigma_y}{|r|\, (\sigma_x^2 + \sigma_y^2)}$$

* Examples :-

1) obtain regression lines for the following data

x	6	2	10	4	8
y	9	11	5	8	7

hence find $y(5)$.

→ Table :-

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
6	9	36	81	54
2	11	4	121	22
10	5	100	25	50
4	8	16	64	32
8	7	64	49	56
30	40	220	340	214

$n=5$, $\bar{x} = \frac{\Sigma x_i}{n} = \frac{30}{5} = 6$

$\bar{y} = \frac{\Sigma y_i}{n} = \frac{40}{5} = 8$

$\sigma_x = \sqrt{\frac{\Sigma x_i^2}{n} - \left(\frac{\Sigma x_i}{n}\right)^2} = \sqrt{\frac{220}{5} - 36} = 2.83$

$\sigma_y = \sqrt{\frac{\Sigma y_i^2}{n} - \left(\frac{\Sigma y_i}{n}\right)^2} = \sqrt{\frac{340}{5} - 64} = 2$

$r = \frac{COV(x,y)}{\sigma_x \sigma_y} = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{n \sigma_x \sigma_y}$

$= \frac{\frac{\Sigma [x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y}]}{n}}{\sigma_x \sigma_y}$

$= \frac{\frac{\Sigma x_i y_i}{n} - \bar{x}\bar{y}}{\sigma_x \sigma_y} = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{n \sigma_x \sigma_y}$

$$r = \frac{214 - (5)(6)(8)}{5(2.83)(2)}$$

$$r = -0.9187 \text{ //}$$

i) Regression line of y on x is

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

$$y - 8 = (-0.9187) \frac{2}{2.83} (x - 6)$$

$$y = -0.6493\, x + 3.8958 + 8$$

$$\boxed{y = -0.6493\, x + 11.8958}$$

2) Regression Line of x on y :—

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

$$x - 6 = (-0.9187) \left(\frac{2.83}{2}\right) (y - 8)$$

$$\boxed{x = -1.291\, y + 16.39}$$

2) The regression equations are $8x - 10y + 66 = 0$ & $40x - 18y - 214 = 0$. The value variance of x is 9.

Find

i) The mean values of x & y

ii) The correlation between x & y

iii) The standard deviation of y.

$8x = 10y - 66$

$x = \frac{10}{8}y - \frac{66}{8} \Rightarrow b_{xy} = \frac{10}{8}$

$18y = 40x - 214$

$y = \frac{40}{18}x - \frac{214}{18} \Rightarrow b_{yx} = \frac{40}{18}$

$b_{xy} \cdot b_{yx} = \frac{10}{8} \cdot \frac{40}{18} > 1$ —

→ Let regression line of y on x be

$8x - 10y + 66 = 0$

$\Rightarrow 10y = 8x + 66$

$y = \frac{8}{10}x + \frac{66}{10} = 0.8x + 6.6$

$\Rightarrow b_{yx} = 0.8$ //

And let the regression line of x on y be

$40x - 18y - 214 = 0$

$\Rightarrow 40x = 18y + 214$

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

$x = 0.45y + 5.35$

$\Rightarrow b_{xy} = 0.45$ //

i) Mean value of x & y

Point of intersection of two regression lines is $(\bar{x}, \bar{y})$

Solving two equations simultaneously, we get

$$\boxed{\bar{x} = 13 \quad , \quad \bar{y} = 17}$$

ii) The correlation bet^n x & y

We know,

$$r = \begin{cases} +\sqrt{b_{yx}\, b_{xy}} & ; \text{ if both } b_{xy}, b_{yx} > 0 \\ -\sqrt{b_{xy}\, b_{yx}} & ; \text{ if both } b_{xy}, b_{yx} < 0 \end{cases}$$

$$r = +\sqrt{b_{xy}\, b_{yx}} \quad ; \quad \because b_{yx} = 0.45 > 0,\ b_{xy} = 0.8 > 0$$

$$= +\sqrt{(0.8)(0.45)}$$

$$= +\sqrt{0.36}$$

$$\boxed{r = 0.6} \qquad (-1 \le r \le 1)$$

iii) To find σ_y :-

Given variance of x is 9

i.e. $\sigma_x^2 = 9 \Rightarrow \sigma_x = 3$

As we know,

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

$$\Rightarrow \sigma_y = \frac{\sigma_x\, b_{yx}}{r} = \frac{(3)(0.8)}{0.6} = 4$$

2) If the two lines of regression are $9x+y-\lambda=0$ & $4x+y=\mu$. Mean of x & y are 2 & -3 respectively, find the values of λ, μ & coefficient of correlation between x & y.

→ i) To find λ & μ.

As $(\bar{x}=2, \bar{y}=-3)$ is the pt of intersection of two regression lines

$\Rightarrow$ $(\bar{x}, \bar{y})$ satisfies both the eq's.

$\Rightarrow$ $\lambda=15$, $\mu=5$.

∴ Regression lines are

$9x+y=15$ & $4x+y=5$

↓ ↓

x on y y on x

↓ ↓

$b_{xy} = -\frac{1}{9}$ $b_{yx} = -4$

$$r = -\sqrt{b_{xy}\, b_{yx}}$$

$$= -\sqrt{\frac{4}{9}}$$

$$\boxed{r = -0.663}$$

* Curve Fitting by Least Squares Criteria :-

A) Polynomial Regression :-

Let $y = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$

represents polynomial of deg 'n'.

For given set of n pairs of observations (x_i, y_i) the unknowns $a_0, a_1, \cdots, a_n$ are estimated by least squares method of minimizing.

$$S = d_i^2 = \sum_{i=1}^{n} [y_i - (a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_n x_i^n)]^2$$

This results in following n+1 normal equations.

$$\frac{\partial S}{\partial a_0} = 0 \Rightarrow 2 \sum_{i=1}^{n} [y_i - (a_0 + a_1 x_i + \cdots + a_n x_i^n)] = 0$$

$$\Rightarrow \sum y_i = n a_0 + a_1 \sum x_i + \cdots + a_n \sum x_i^n \quad \text{— (1)}$$

$$\frac{\partial S}{\partial a_1} = 0 \Rightarrow 2 \sum_{i=1}^{n} (-x_i)[y_i - (a_0 + a_1 x_i + \cdots + a_n x_i^n)] = 0$$

$$\Rightarrow \sum x_i y_i = a_0 \sum x_i + a_1 \sum x_i^2 + \cdots + a_n \sum x_i^{n+1} \quad \text{— (2)}$$

$$\vdots$$

$$\frac{\partial S}{\partial a_n} = 0 \Rightarrow 2 \sum_{i=1}^{n} (-x_i^n)[y_i - (a_0 + a_1 x_i + \cdots + a_n x_i^n)] = 0$$

$$\Rightarrow \sum x_i^n y_i = a_0 \sum x_i^n + a_1 \sum x_i^{n+1} + \cdots + a_n \sum x_i^{2n} \quad \text{— (n+1)}$$

Solving these (n+1) normal eq's simultaneously, we get values of $a_0, a_1, \cdots, a_n$ & hence best fitting curve.

25/05/2021

A) Fitting of a Straight Line :-

Least squares straight line

To fit a straight line

$$y = ax + b \quad ; \quad (\text{where } a, b \text{ constants to be determined})$$

to (x_i, y_i) ; $i = 1, 2, \cdots, n$ points by least squares criteria has normal eq^{n}s.

$$\sum y_i = na + b \sum x_i$$

$$\sum x_i y_i = a \sum x_i + b \sum x_i^2$$

Solving these two eq^{n}s we can find a & b.

B) Fitting of a Parabola (Quadratic Curve) :-

To fit the parabola

$$y = a_0 + a_1 x + a_2 x^2$$

where a_0, a_1, a_2 are constants to be determined, to (x_i, y_i) ; $i = 1, 2, \cdots, n$ points has normal equations.

$$\sum_{i=1}^{n} y_i = n a_0 + a_1 \sum x_i + a_2 \sum x_i^2$$

$$\sum_{i=1}^{n} x_i y_i = a_0 \sum x_i + a_1 \sum x_i^2 + a_2 \sum x_i^3$$

$$\sum_{i=1}^{n} x_i^2 y_i = a_0 \sum x_i^2 + a_1 \sum x_i^3 + a_2 \sum x_i^4$$

Solving these simultaneously we get a_0, a_1, a_2

c) Fitting of Exponential Curve :-

i) Growth curve :- $y = A B^x$

$$\ln y = \ln A + x \ln B$$

$$y^* = a^* + b^* x$$

... straight line

Normal eq's are.

$$\sum y_i^* = n a^* + b^* \sum x_i$$

$$\sum x_i y_i^* = a^* \sum x_i + b^* \sum x_i^2$$

Solving these two, we can find a^* & b^* hence A & B.

2) Decay Curve :- $y = A B^{-x}$

d) Fitting of Geometric Curve :-

$$y = A x^B$$

$$\ln y = \ln A + B \ln x$$

$$y^* = a^* + B x^*$$

$$y^* = \ln y, \quad a^* = \ln A, \quad x^* = \ln x$$

∴ Normal Eqⁿ are.

$$\sum y_i^* = n a^* + B \sum x^*$$

$$\sum x_i^* y_i^* = a^* \sum x_i^* + B \sum (x_i^*)^2$$

Solving these two eq^{n}s we can find a^* & B hence A & B.

E) Fitting of Reciprocal Curve :-

$$y = \frac{1}{A+Bx}$$

Put $y^* = A + Bx$

$$\Rightarrow \quad y = \frac{1}{y^*}$$

$\therefore$ Normal eq^{n}s are

$$\sum y_i^* = n A + B \sum x_i$$

$$\sum x_i y_i^* = A \sum x_i + B \sum x_i^2$$

Solving these two, we can find A & B.

* Examples:-

1) Fit a straight line to the following data by least squares method

x	0	2	4	6	8	12	20
y	10	12	18	22	20	30	30

Hence find $y(22)$.

→ To fit the straight line

$$y = a + bx$$

to the $n = 7$ points

∴ Normal eq^{n}s are,

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

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

Table:-

x_i	y_i	x_i^2	$x_i y_i$
0	10	0	0
2	12	4	24
4	18	16	72
6	22	36	132
8	20	64	160
12	30	144	360
20	30	400	600
Total 52	142	664	1348

eq^n (1) & (2) becomes,

$$142 = 7a + 52b \quad \text{—— (3)}$$

$$1348 = 52a + 664b \quad \text{—— (4)}$$

Solving (3) & (4) simultaneously we get

$$b = 1.0556, \quad a = 12.4444$$

∴ Eq^n of line is $y = 1.0556x + 12.4444$ & $y(22) = 35.66$ //

2) Fit a parabola (second deg. poly) to the following data.

x	-2	-1	0	1	2
y	4	1	2	7	15

→ Let the eq^n be

$$y = a_0 + a_1 x + a_2 x^2$$

to fit to $n = 5$ points.

∴ Normal eq^{n}s are

$$\sum y_i = n a_0 + a_1 \sum x_i + a_2 \sum x_i^2$$

$$\sum x_i y_i = a_0 \sum x_i + a_1 \sum x_i^2 + a_2 \sum x_i^3$$

$$\sum x_i^2 y_i = a_0 \sum x_i^2 + a_1 \sum x_i^3 + a_2 \sum x_i^4$$

Table :

x_i	y_i	x_i^2	x_i^3	x_i^4	$x_i y_i$	$x_i^2 y_i$
-2	4	4	-8	16	-8	16
-1	1	1	-1	1	-1	1
0	2	0	0	0	0	0
1	7	1	1	1	7	7
2	15	4	8	16	30	60
Total 0	29	10	0	34	28	84

Normal eq^{n}s becomes.

$$29 = 5a_0 + 0 + 10a_2 \quad \text{—— ①}$$

$$28 = 0 + 10a_1 + 0 \quad \text{—— ②}$$

$$84 = 10a_0 + 0 + 34a_2 \quad \text{—— ③}$$

Solving ①, ② & ③ simultaneously, we get

$$a_0 = 2.085, \quad a_1 = 2.8, \quad a_2 = 1.857$$

$$\therefore \quad \boxed{y = 2.085 + 2.8x + 1.857x^2}$$

3) Fit an exponential curve $y = ab^x$ to the following data

x	2	4	6	8	10	12
y	1.8	1.5	1.4	1.1	1.1	0.9

→ Given curve is,

$$y = ab^x$$

$$\ln y = \ln a + \ln b\, x$$

Put $y^* = \ln y$, $a^* = \ln a$, $b^* = \ln b$

$$\therefore \quad y^* = a^* + b^* x$$

which is linear in x.

Hence normal eqns are

$$\sum y_i^* = n a^* + b^* \sum x_i$$

$$\sum x_i y_i^* = a^* \sum x_i + b^* \sum x_i^2$$

Table

x_i	y_i	$y_i^* = \ln y$	x_i^2	$x_i y_i^*$
2	1.8	0.5878	4	1.1756
4	1.5	0.4055	16	1.622
6	1.4	0.3365	36	2.019

8	1.1	0.0953	64	0.7624
10	1.1	0.0953	100	0.953
12	0.9	−0.1054	144	−1.2648
Total 42		1.415	364	5.2684

∴ Normal eq^{n}s becomes.

$$1.415 = 6a^* + 42b^*$$

$$5.2684 = 42a^* + 364\,b^*$$

Solving these two eq^{n}s we get

$$a^* = 0.6995 \quad , \quad b^* = -0.0662$$

$$\ln a = a^* \Rightarrow a = e^{a^*} = e^{0.6995} = 2.012$$

$$\ln b = b^* \Rightarrow b = e^{b^*} = e^{-0.0662} = 0.936$$

∴ The required least squares exponential curve is,

$$\boxed{y = (2.012)\,(0.936)^x}$$

H.W.:

1) Fit an exponential curve of the form $y = ae^{bx}$ for the following data

x:	1	2	3	4
y:	7	11	17	27

Ans: $y = 4.48e^{0.45x}$

2) Fit a power function (geometric curve) of the form $y = ax^b$ to the following data and estimate y at $x = 12$

Price x	20	16	10	11	14
Demand y	22	41	120	89	56

Ans: $y = 28491\ x^{-2.38}$

$y(x=12) \approx 77$