Mod 5] Statistical Techniques

1] Correlation :-

The correlation is used to measure the association/relation b/w two or more variables.

eg. number of air conditioners sale & temperature in city.
income & expenditure of a person.

→ The variables may be related & sometimes there is no relation b/w them.

eg. The number of tourists travel are related to holidays.

The correlation is defined as the measure of strength of association b/w two variables.

* Types of correlation :-

1] Positive correlation :

If an increase (or decrease) in value of one variable correspond to an increase (or decrease) in the value of the other variable than the correlation is said to be positive correlation.

eg. radius of circle & area of circle are positively correlated.

2] Negative correlation :

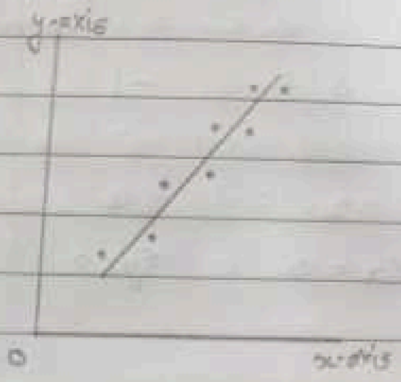
If an increase (or decrease) in the value of one variable correspond to decrease (or increase) in the value of the other variable than the correlation is said to be negative correlation.

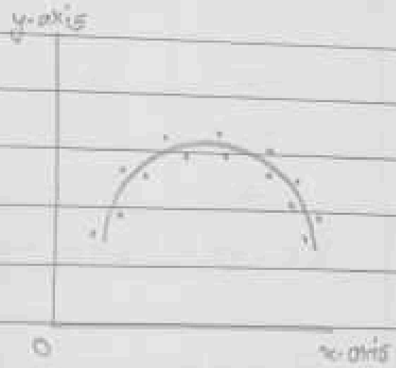
eg. The increase in commodity prices as supply decreases.

3) Linear & Non-linear Correlation:

If all the sets of points plotted in X-Y plane lie approximately on a straight line the correlation is linear correlation.

If all the sets of points plotted in X-Y plane lie approximately on a nonlinear curve then the correlation is said to be nonlinear correlation.





• Karl Pearson Coefficient of Correlation (r):

The Karl Pearson Coefficient of Correlation b/w variables X & Y can be calculated by relation -

$$\boxed{r = \frac{n\Sigma(xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma(x^2) - (\Sigma x)^2][n\Sigma(y^2) - (\Sigma y)^2]}}}$$

where, $x = X - \overline{X}$, $y = Y - \overline{Y}$

$\overline{X} = \frac{\Sigma X}{n}$, $\overline{Y} = \frac{\Sigma Y}{n}$

Q 1) Calculate the coefficient of correlation b/w X & Y from following data.

X	3	6	4	5	7
Y	2	4	5	3	6

⇒

X	Y	x $x = X - \bar{X}$	y $y = Y - \bar{Y}$	xy	x^2	y^2
3	2	-2	-2	4	4	4
6	4	1	0	0	1	0
4	5	-1	1	-1	1	1
5	3	0	-1	0	0	1
7	6	2	2	4	4	4
$\Sigma X = 25$	$\Sigma Y = 20$	$\Sigma x = 0$	$\Sigma y = 0$	$\Sigma xy = 7$	$\Sigma x^2 = 10$	$\Sigma y^2 = 10$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{25}{5} = 5$$

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{20}{5} = 4$$

$$r = \frac{n\Sigma(xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma(x^2) - (\Sigma x)^2][n\Sigma(y^2) - (\Sigma y)^2]}}$$

$$= \frac{5(7) - (0)(0)}{\sqrt{[5(10) - (0)^2][5(10) - (0)^2]}}$$

$$= \frac{35}{\sqrt{50 \times 50}}$$

$$= \frac{\cancel{35}\,7}{\cancel{50}\,10}$$

$$= 0.7$$

Q.2]

X	12	17	22	27	32
Y	113	119	117	115	121

$\Rightarrow \bar{X} = \frac{\Sigma X}{n} = \frac{110}{5} = 22$

$\bar{Y} = \frac{\Sigma Y}{n} = \frac{585}{5} = 117$

X	Y	x	y	xy	x^2	y^2
12	113	-10	-4	40	100	16
17	119	-5	2	-10	25	4
22	117	0	0	0	0	0
27	115	5	-2	-10	25	4
32	121	10	4	40	100	16
$\Sigma X = 110$	$\Sigma Y = 585$	$\Sigma x = 0$	$\Sigma y = 0$	$\Sigma xy = 60$	$\Sigma x^2 = 250$	$\Sigma y^2 = 40$

$$\therefore r = \frac{n\Sigma(xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma(x^2) - (\Sigma x)^2][n\Sigma(y^2) - (\Sigma y)^2]}}$$

$$= \frac{5(60) - (0)(0)}{\sqrt{[5(250) - (0)^2][5(40) - (0)^2]}}$$

$$= \frac{300}{\sqrt{5 \times 250 \times 5 \times 40}}$$

$$= 0.6$$

* Karl Pearson's coefficient of correlation.

Important properties:

1] $-1 \leq r \leq 1$

2] Correlation coefficient is independent of change of origin & change of scale i.e. If $x = au + b$, $y = cv + d$ where a, b, c, d are constants then, $r_{xy} = r_{uv}$.

9] Interpretation of coefficient correlation.

1] $r > 0.95 \Rightarrow$ High degree of correlation.
The value of variable can be estimated accurately.

2] $0.75 < r < 0.95 \Rightarrow$ The value of variable can be calculated roughly from value of other variable.

3] $0.40 < r < 0.60 \Rightarrow$ Somewhat related.
Value of one variable cannot be calculated.

4] $r < 0.35 \Rightarrow$ Poor correlation.
One variable cannot be estimated from other.

5] $r \sim 0 \Rightarrow$ No relation : Independent variable.

H.W

Q. Calculate the coefficient of correlation.

X	100	98	85	92	90	84	88	90	93	95
Y	500	610	700	630	670	800	800	750	700	690

* Spearman's Rank correlation coefficient (R) :

Type 1 : When ranks are given for 2 variables (R_1 & R_2)

$$\boxed{R = 1 - \frac{6 \Sigma D^2}{N^3 - N}}$$

where $D = R_1 - R_2$
N = No. of observations

Type 2 : When ranks are not given, values are given.

1] Arrange X in ascending or descending order
2] Arrange Y in ascending or descending order
3] Create column of rank R_1 for X & R_2 for Y values.
4] Create column $D = R_1 - R_2$
5] Calculate R by above formula.

Note:

X	3	8	9	10	8	3
Y	14	10	2	3	10	10

→

X	Y	R_1	R_2
3	14	1.5	6
8	10	3.5	4
9	2	5	1
10	3	6	2
8	10	3.5	4
3	10	1.5	4

	1	2	3	4	5	6
$R_1 \Rightarrow$	3	3	8	8	9	10
	1.5 $\frac{1+2}{2}$	1.5	3.5 $\frac{3+4}{2}$	3.5	5	6
			4	4	4	
$R_2 \Rightarrow$	2	3	10	10	10	14
	1	2		$\frac{3+4+5}{3}$		6

→ If two or more members have same ranks, then, once ranks are assigned, if m is the number of members having equal ranks then factor $\frac{1}{12}(m^3-m)$ is added to ΣD^2

$$\therefore \boxed{R = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}(m^3-m)\right]}{N^3-N}}$$

→ If there are more than one such cases then this factor is added corresponding to each case.

$$\therefore \boxed{R = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}(m_1^3-m_1) + \frac{1}{12}(m_2^3-m_2) + \dots\right]}{N^3-N}}$$

For above ex ⇒ $R = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}(2^3-2) + \frac{1}{12}(2^3-2) + \frac{1}{12}(3^3-3)\right]}{N^3-N}$

3 8 10

Q.1] compute the spearman's Rank correlation coefficient.

X	18	20	34	52	12
Y	39	23	35	18	46

⇒

X = 12 18 20 34 52
1 2 3 4 5

Y = 18 23 35 39 46
1 2 3 4 5

X	Y	R_1	R_2	$D = R_1 - R_2$	D^2
18	39	2	4	-2	4
20	23	3	2	1	1
34	35	4	3	1	1
52	18	5	1	4	16
12	46	1	5	-4	16
					$\Sigma D^2 = 38$

N = 5

$$R = 1 - \frac{6\left[\Sigma D^2\right]}{N^3 - N}$$

$$= 1 - \frac{6\,[38]}{(5)^3 - 5}$$

$$= 1 - \frac{\cancel{6} \times \cancel{38}}{\cancel{120}}$$

$$= 1 - 1.9$$

$$= -0.9$$

Q.7 Compute the spearman's Rank correlation coefficient.

X	82	71	82	46	62	74	71	56	71	85
Y	75	86	75	55	57	68	78	52	65	67

→

X = 46 56 62 71 71 71 74 82 82 85
1 2 3 5 5 5 7 8.5 8.5 10

$\therefore \frac{4+5+6}{3} = 5$

Y = 52 55 57 65 67 68 75 75 78 86
1 2 3 4 5 6 7.5 7.5 9 10

$\therefore \frac{8+9}{2} = 8.5$

$\therefore \frac{7+8}{2} = 7.5$

X	Y	R_1	R_2	$D = R_1 - R_2$	D^2	
82	75	8.5	7.5	1	1	
71	86	5	10	-5	25	
82	75	8.5	7.5	1	1	
46	55	1	2	-1	1	
62	57	3	3	0	0	
74	68	7	6	1	1	
71	78	5	9	-4	16	
56	52	2	1	1	1	
71	65	5	4	1	1	
85	67	10	5	5	25	
					$\Sigma D^2 = 72$	$N = 10$

$$R = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}(m^3 - m) + \dots\right]}{N^3 - N}$$

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

$$= 1 - \frac{6\left[72 + 2 + 1/2 + 1/2\right]}{(10)^3 - 10}$$

$= 0.5455$

Q.3] Find R.

X	60	30	37	30	42	37	55	45
Y	50	25	33	27	40	33	50	42

1.5 1.5 3.5 3.5 5 6 7 8

⇒ X = 30 30 37 37 42 45 55 60

Y = 25 27 33 33 40 42 50 50

1 2 3.5 3.5 5 6 7.5 7.5

X	Y	R_1	R_2	$D = R_1 - R_2$	D^2
60	50	1.5	1	0.5	0.25
30	25	1.5	2	-0.5	0.25
37	33	3.5	3.5	0	0
30	27	3.5	3.5	0	0
42	40	5	5	0	0
37	33	6	6	0	0
55	50	7	7.5	-0.5	0.25
45	42	8	7.5	0.5	0.25
					$\Sigma D^2 = 1$

$N = 8$

$$R = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}(m^3 - m) + \dots\right]}{N^3 - N}$$

$$= 1 - \frac{6\left[1 + \frac{4}{12}(2^3 - 2)\right]}{(8)^3 - 8}$$

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

$$= 0.964$$

* Regression :-

Regression is a method of estimating the value of one variable when that of the other variable is known & when the variables are correlated.

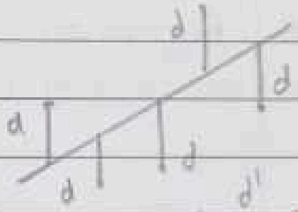
* Lines of Regression :

When two variables are highly correlated in the graph the points lie in a narrow strip.

If the strip is nearly a straight line, we may draw a line such that all the points are close to that line from both the sides. A line drawn such that the square root sum of square of the the distances of the points from line is minimum is called the line of regression.

There are two lines of regression :

1) The line of regression of Y on X.
This line is of form $Y = a + bX$.



2) The line of regression of X on Y
This line is of form $X = a + bY$.

* The equations of line of regression of Y on X is

$$\boxed{Y - \bar{Y} = b_{yx}(X - \bar{X}) \qquad \text{where, } b_{yx} = \frac{N\Sigma xy - \Sigma x \Sigma y}{N\Sigma x^2 - (\Sigma x)^2}}$$

, where $x = X - \bar{X}$, $y = Y - \bar{Y}$

where, $\bar{X} = \frac{\Sigma X}{N}$, $\bar{Y} = \frac{\Sigma Y}{N}$

<u>Note</u> : If $\bar{x}$ & $\bar{Y}$ are not integers value then round off to integer values & use that new value <u>only</u> for calculating the table values.

* The equ. of line of regression of X on Y is

$$\boxed{X - \bar{X} = b_{xy}(Y - \bar{Y}) \quad \text{where, } b_{xy} = \frac{N\Sigma xy - \Sigma x \Sigma y}{N\Sigma y^2 - (\Sigma y)^2}}$$

where, $x = X - \bar{X}$ & $y = Y - \bar{Y}$

where, $\bar{X} = \frac{\Sigma X}{N}$ & $\bar{Y} = \frac{\Sigma Y}{N}$

Note :- 1] If $\bar{X}$ & $\bar{Y}$ are not integer values then round off to integer values & use that new values only for calculating the table values.

2] Here, b_{xy} & b_{yx} are called as coefficients of regression.

* Regression coefficients :-

a] Coefficient of regression of y on x :

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

where, $\sigma_x = \sqrt{\frac{(X-\bar{X})^2}{N}}$ & $\sigma_y = \sqrt{\frac{(Y-\bar{Y})^2}{N}}$

$\bar{X} = \frac{\Sigma X}{N}$, $\bar{Y} = \frac{\Sigma Y}{N}$

b] Coefficient of regression of x on y :

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

where, $\sigma_x = \sqrt{\frac{(X-\bar{X})^2}{N}}$ & $\sigma_y = \sqrt{\frac{(Y-\bar{Y})^2}{N}}$

$\bar{X} = \frac{\Sigma X}{N}$, $\bar{Y} = \frac{\Sigma Y}{N}$

* Properties of coefficients of regression :-

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

2) Both the coefficients of regression always have same sign. ($\because$ their product is always equal to r^2 which will be a positive number)

3) If one coefficient of regression is greater than one then, the other must be less than 1.
[Proof imp]

4) Arithmetic mean of the coefficients of regression is greater than or equal to the coefficients of correlation r.
i.e. $\frac{b_{yx} + b_{xy}}{2} \geq r$.

5) Coefficients of regression are independent of change of origin but not of change of scale.

6) If correlation is perfect (i.e. $r = \pm 1$) then, two coefficients of regression are reciprocals of each other.

Q.1)

X	5	6	7	8	9	10	11
Y	11	14	14	15	12	17	16

Find equ. of line of regression of Y on X for the table.

$\Rightarrow$

we know, equ. is $Y - \bar{Y} = b_{yx}(X - \bar{X})$

where, $b_{yx} = \frac{N\Sigma xy - \Sigma x \Sigma y}{N\Sigma x^2 - (\Sigma x)^2}$

here, $N = 7$ $\quad \bar{X} = \frac{\Sigma X}{N} = \frac{56}{7} = 8$ $\quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{99}{7} = 14.14 \approx 14$

X	Y	x	y	xy	x^2
5	11	-3	-3	9	9
6	14	-2	0	0	4
7	14	-1	0	0	1
8	15	0	1	0	0
9	12	1	-2	-2	1
10	17	2	3	6	4
11	16	3	2	6	9
		$\Sigma x = 0$	$\Sigma y = 1$	$\Sigma xy = 19$	$\Sigma x^2 = 28$

H.W

$$b_{yx} = \frac{7(19) - (0)(\cancel{1})}{7(28) - (0)^2} = \frac{133}{196} = 0.6786$$

$$\therefore \; Y - 14.1429 = (0.6786)(X - 8)$$

$$Y = 0.6786X + 8.7141$$

Q.7

X	100	110	120	130	140	150	160	170	180	190
Y	45	51	54	61	56	70	74	78	85	89

Find b_{xy}, b_{yx} & r.

⇒

X	Y	x	y	xy	x^2	y^2
100	45	-45	-21	945	2025	441
110	51	-35	-15	525	1225	225
120	54	-25	-12	300	625	144
130	61	-15	-5	75	225	25
140	56	-5	-10	50	25	100
150	70	5	4	20	25	16
160	74	15	8	120	225	64
170	78	25	12	300	625	144
180	85	35	19	665	1225	361
190	89	45	23	103	2025	529
		$\Sigma x = 0$	$\Sigma y = 3$	$\Sigma xy = 3103$	$\Sigma x^2 = 8250$	$\Sigma y^2 = 2049$

$\bar{X} = \frac{1450}{10}$

$= 145$

$\bar{Y} = \frac{663}{10}$

$= 66.3$

≈ 66

Note :- if both b_{yx} & b_{xy} are +ve
then $r \to$ +ve
if both b_{yx} & b_{xy} are -ve
then $r \to$ -ve

$$b_{yx} = \frac{N\Sigma xy - \Sigma x \Sigma y}{N\Sigma x^2 - (\Sigma x)^2} = \frac{10(3103) - (0)(3)}{10(8250) - (0)^2} = 0.3761$$

$$b_{xy} = \frac{N\Sigma xy - \Sigma x \Sigma y}{N\Sigma y^2 - (\Sigma y)^2} = \frac{10(3103) - (0)(3)}{10(2049) - (3)^2} = 1.5150$$

$$\text{and } r = \sqrt{b_{yx} \cdot b_{xy}} = \sqrt{0.3761 \times 1.5150} = 0.7548.$$

H.W. Q.3] Find equ. of lines of regression for data. Also Find r and estimate Y when $X = 15$

X	7	8	9	10	11	12	13
Y	13	16	16	17	14	19	18

Note : If X values is known & we want to calculate corresponding Y value, then equ. of line of regression of Y on X is obtained and from that required value is calculated.

Q.4] Given $y = 30 - 6x$ & $2x = -y + 12$, $\sigma_x^2 = 16$.
Find 1] $\bar{X}$ & $\bar{Y}$ 2] r 3] σ_y^2

$\Rightarrow$

here, $y = 30 - 6x$ & $2x = -y + 12$

solving these two equ. $\Rightarrow (x, y) = 4.5, 3$

$\therefore \bar{X} = 4.5$ & $\bar{Y} = 3$

here, equ. are $y = 30 - 6x \Rightarrow y = a + bx \quad \therefore b_{yx} = -6$

$x = -\frac{y}{2} + 6 \Rightarrow x = a + by \quad \therefore b_{xy} = -1/2$

$\therefore r = \sqrt{b_{yx} \cdot b_{xy}} = \sqrt{-6 \times -1/2} = \sqrt{3} = -1.732$ ✗ $\because [-1 < r < 1]$

$\therefore$ Equ's are $x = 5 - y/6 \quad \therefore b_{xy} = -1/6 \qquad \therefore r = \sqrt{-1/6 \times -2}$

$y = 12 - 2x \quad \therefore b_{yx} = -2 \qquad = \sqrt{1/3} = -0.5773$

Page No. 120

now, $b_{yx} = r\frac{\sigma_y}{\sigma_x} \Rightarrow \sigma_y = \frac{\sigma_x b_{yx}}{r}$

$= \frac{(-2)(4)}{-0.5774}$

$= 13.855$

H.W Q.] If two lines of regression are $4x-5y+33=0$ and $20x-9y-107=0$
Find the value of
1] $\bar{X}$ & $\bar{Y}$
2] r
3] σ_y if $\sigma_x = 3$

* Curve Fitting :-

Fitting a straight line or 1° curve $y = a+bx$.

x & y are given in data, we need to fit the straight line which best fits the given data. This can be done by least square method

$$\boxed{\begin{aligned} \Sigma y &= Na + b\Sigma x \\ \Sigma xy &= a\Sigma x + b\Sigma x^2 \end{aligned}}$$

These are normal equations for the straight line obtained using least square method.

Q.1] Fit a straigh line to the following data.

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

$\Rightarrow$

Let $y = a+bx$ be the required straight line ... ①

To fit a straight line to given data, the normal equs are

$\Sigma y = Na + b\Sigma x$ ... ②

$\Sigma xy = a\Sigma x + b\Sigma x^2$ ... ③

here, $N = 5$

x	y	xy	x^2
10	14	140	100
12	17	204	144
15	23	345	225
23	25	575	529
20	21	420	400
$\Sigma x = 80$	$\Sigma y = 100$	$\Sigma xy = 1684$	$\Sigma x^2 = 1398$

Equ. ② $\Rightarrow$ $\Sigma y = Na + b\Sigma x$

$100 = 5a + 80b \Rightarrow 1600 = 80a + 1280b$

Equ. ③ $\Rightarrow$ $\Sigma xy = a\Sigma x + b\Sigma x^2$

$1684 = 80a + 1398b$

solving these two equ,

$84 = 0 + 118b$

$b = \frac{84}{118} = 0.7118$

$\therefore 1600 - 1280(0.7118) = 80a$

$\therefore 80a = 688.896$

$a = 8.6112 \Rightarrow \underline{y = 8.6112 + 0.7118x}$

* Fitting a Parabola / second degree curve to the given data:

$$y = a + bx + cx^2$$

Let x & y be the variables in the given data, we need to fit a parabola which best fits the given data. This can be done by least square method.

The normal equa to fit the given data are as follows:

$$\boxed{\begin{aligned} \Sigma y &= Na + b\Sigma x + c\Sigma x^2 \\ \Sigma xy &= a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \\ \Sigma x^2 y &= a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 \end{aligned}}$$

Q.] Using least square method, fit a parabola to the given data.

x	-2	-1	0	1	2
y	-3.150	-1.390	0.620	2.880	5.378

$\Rightarrow$

x	y	x^2	x^3	x^4	xy	x^2y
-2	-3.150	4	-8	16	6.3	-12.6
-1	-1.390	1	-1	1	1.39	-1.39
0	0.620	0	0	0	0	0
1	2.880	1	1	1	2.88	2.88
2	5.378	4	8	16	10.756	21.512
$\Sigma x = 0$	$\Sigma y = 4.338$	$\Sigma x^2 = 10$	$\Sigma x^3 = 0$	$\Sigma x^4 = 34$	$\Sigma xy = 21.326$	$\Sigma x^2y = 10.402$

Equ. are $\Rightarrow$ $\Sigma y = Na + b\Sigma x + c\Sigma x^2$

$4.338 = 5a + 0b + 10c \quad \ldots \text{(I)}$

$N = 5$

$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$

$21.326 = 0a + 10b + 0c \quad \ldots \text{(II)}$

$\Sigma x^2y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$

$10.402 = 10a + 0b + 34c \quad \ldots \text{(III)}$

Solv. (I), (II) & (III), we get,

$a = 0.621, \ b = 2.1326, \ c = 0.1232$

$\therefore \quad \underline{y = 0.621 + 2.1326x + 0.1232x^2}$