where, C = is the common class width.

Module 3

1 Bivariate Data :-

Data on two variables recorded simultaneously for a group of individuals are called "Bivariate data".

eg1. Height & weight of the students in a class

eg2. The marks obtained by a group of students in the test and the final exam.

eg3: The income and expenditure of a no. of family.

etc.

2 Scattered Diagram :-

The graphical representation of a bivariate data is called scattered diagram.

✡ 3 Correlation :-

By correlation we mean the association or inter-dependence between two variables. If one variable is found to increase, as the other increases, the variables are said to be positively correlated.

Again if one variable decreases as the other increases, they are said to be negatively correlated.

If one variable increases or decreases the other remain constant, the variables are said to be uncorrelated / zero correlated / independent.

4 Karl Pearson Correlation Coefficient :-

It is defined as $r_{xy} = \frac{cov(x, y)}{Sd(x) \times sd(y)}$

cov → co variance

$$\Rightarrow cov(x, y) = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - (\bar{x} \times \bar{y})$$

$$\Rightarrow Sd(x) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2 - (\bar{x})^2}$$

$$\Rightarrow Sd(y) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} y_i^2 - (\bar{y})^2}$$

Page No.
Date

Remark ①: $-1 \leq r_{xy} \leq 1$

q1 Using the following information calculate correlation coefficient b/w x & y.

$n = 5$
$\Sigma x = 30$
$\Sigma y = 67$
$\Sigma xy = 480$
$\Sigma x^2 = 220$
$\Sigma y^2 = 1059$

Ans

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

$$\bar{y} = \frac{\Sigma y}{n} = \frac{67}{5} = 13.4$$

$$cov(x, y) = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - (\bar{x} \times \bar{y})$$

$$= \frac{480}{5} - (6 \times 13.4)$$

$$= 15.6$$

$$sd(x) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2 - (\bar{x})^2}$$

$$= \sqrt{\frac{220}{5} - (6)^2}$$

$$= 2.82$$

$$sd(y) = \sqrt{\frac{1}{n}\Sigma y^2 - (\bar{y})^2}$$

$$= \sqrt{\frac{1059}{5} - (13.4)^2}$$

$$= 5.67$$

$$\approx 5.68$$

$$r_{xy} = \frac{cov(x, y)}{sd(x) \times sd(y)}$$

$$= \frac{15.6}{2.82 \times 5.68}$$

$$= 0.97$$

∴ x & y are highly positively correlated.

Q2 A computer while calculating correlation coefficient b/w x & y for 25 observations the following results are obtained.

$$n = 25$$
$$\Sigma x = 125$$
$$\Sigma y = 100$$
$$\Sigma x^2 = 650$$
$$\Sigma y^2 = 460$$
$$\Sigma xy = 508$$

It was, however later discovered at the time of checking that two pair of values (8, 12) & (6, 8) are wrongly copied as (6, 14) & (8, 6)

Calculate the current value of correlation coefficient.

Ans Correct $(\Sigma x) = 125 - (6+8) + (8+6)$

wrong value, right value

$= 125$

Correct $(\Sigma y) = 100 - (14+6) + (12+8)$

$= 100$

Correct $(\Sigma x^2) = 650 - (6^2 + 8^2) + (8^2 + 6^2)$

$= 650$

Correct $(\Sigma y^2) = 460 - (14^2 + 6^2) + (12^2 + 8^2)$

~~460~~

$= 460 - 232 + 208$

$= 436$

Correct $(\Sigma xy) = 508 - [(6 \times 14) + (8 \times 6)] + [(8 \times 12) + (6 \times 8)]$

$$= 508 - (84 + 48) + (96 + 48)$$

$$= 520$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{125}{25} = 5$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{100}{25} = 4$$

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

$$= \frac{520}{25} - (5 \times 4)$$

$$= 20.8 - 20$$

$$= 0.8$$

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

$$= \sqrt{\frac{650}{25} - (5)^2}$$

$$= \sqrt{26 - 25}$$

$$= \sqrt{1}$$

$$= 1$$

$$sd(y) = \sqrt{\frac{1}{n}\Sigma y^2 - (\bar{y})^2}$$

$$= \sqrt{\frac{436}{25} - (4)^2}$$

$$= \sqrt{17.44 - 16}$$

$$= \sqrt{1.44}$$

$$= 1.2$$

$$r_{xy} = \frac{cov(x, y)}{sd(x) \times sd(y)}$$

$$= \frac{0.8}{1 \times 1.2}$$

$$= 0.666$$

$$= 0.67$$

∴ x & y are highly positively correlated.

Page No.
Date

- **Regression :-**

By regression of a variable (say y) on another variable (say x) we mean the dependence of y on x, on the average. In bivariate analysis one of the major problem is prediction of the value of the dependent variable when the independent variable is known. To solve this problem we established a functional relationship b/w x & y.

There are two types of regression equation :

i) When it is required to estimate the value of dependent variable(y) for a given value of independent variable (x). This is called regression equation of y on x. which is given by :

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

where, b_{yx} = regression coefficient of y on x.

$r_{xy} \Leftrightarrow r_{yx}$

Page No.
Date

(1) $b_{yx} = r_{xy} \times \frac{sd(y)}{sd(x)}$

$= \frac{cov(x,y)}{sd(x) \cdot sd(y)} \times \frac{sd(y)}{sd(x)}$

$= \frac{cov(x,y)}{(sd(x))^2}$

$= \frac{cov(x,y)}{var(x)}$ ——— (2)

ii) When it is required to estimate the value of a dependent variable (x) for a given value of independent variable (y) this is called regression equation of x on y which is given by:

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

(1) $b_{xy} = r_{xy} \times \frac{sd(x)}{sd(y)}$

$= \frac{cov(x,y)}{sd(x) \cdot sd(y)} \times \frac{sd(x)}{sd(y)}$

$= \frac{cov(x,y)}{(sd(y))^2}$

Remark ② If θ be the acute angle b/w two regression lines then $\theta = \tan^{-1}\left[\frac{1-r^2}{r} \times \frac{S_x \cdot S_y}{S_x^2 + S_y^2}\right]$

S_x is the Sd of x & S_y is the Sd of y.

$$= \frac{Cov(x,y)}{Var(y)} \quad \text{---(2)}$$

Remark ① $b_{xy} \times b_{yx}$

$$= \left[r \times \frac{Sd(x)}{Sd(y)}\right] \times \left[r \times \frac{Sd(y)}{Sd(x)}\right]$$

$$= r^2$$

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

$$r = \pm\sqrt{b_{yx} \times b_{xy}}$$

if b_{yx} & b_{xy} → +ve then $r \to$ +ve

" " " → -ve " $r \to$ -ve

Q If the lines of regression of y on x is $3x + 2y = 26$ and that of x on y is respectively $6x + y = 31$ Find ~~corelb~~ correlation coefficient (r).

Ans y on x

$$3x + 2y = 26$$

$$\Rightarrow 2y = -3x + 26$$

$$\Rightarrow y = \underbrace{\frac{-3}{2}}_{b_{yx}} x + \frac{26}{2}$$

If $r = 0 \Rightarrow \theta = 90°$ in that case the two regression lines are $\perp$ to each other

If $r = \pm 1 \Rightarrow \theta = 0°$ or $180°$ in that case the regression lines will coincide ie. that will be a perfect co-relation

x on y

$6x + y = 31$

$6x = -y + 31$

$x = -\frac{(-1)}{6}y + \frac{31}{6}$

$\rightarrow b_{xy}$

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

$$= -\sqrt{\left(-\frac{3}{2} \times -\frac{1}{6_2}\right)}$$

$$= -\sqrt{\frac{1}{4}} = -\frac{1}{2}$$

Q You are given the following data

variable	X	Y
mean	47	96
sd	8	9

$r = 0.36$

Determine the value of y when $x = 50$

$y = ?$

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

$$b_{yx} = r \times \frac{Sd(y)}{Sd(x)}$$

$$= 0.36 \times \frac{9}{7}$$

$$= 0.4$$

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

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

$$= 96 + 0.4 (50 - 47)$$

$$= 97.2$$

7/9/23

- **Principle of Least Squares :-**

The least square method is a crucial statistical method to find a regression line or a best fit for the given data. This method is used is evolution and regression.

The method of least square defines the solution for the minimisation of the sum of squares of deviations or the error in the result of evolution.

Let us assume that the given points of data are $(x_1, y_1), (x_2, y_2), \ldots$

(x_n, y_n) in which all x are independent and all y are dependent variable. Suppose that "$f(x)$" is the fitting curve and "d" represent error or deviation from each given points now we can write $d_1 = y_1 - f(x_1)$

$$d_2 = y_2 - f(x_2)$$

$$\vdots$$

$$d_n = y_n - f(x_n)$$

The least square explain that the curve that best fits is represented by the property of that the sum of square (S.S.) must be minimum.

$$S = \sum_{i=1}^{n} di^2$$

$$= d_1^2 + d_2^2 + \cdots + d_n^2$$

is minimum.

- **Fitting of Linear Curve :-**

The equation of least square line is given by $\boxed{y = a + bx}$.

The normal equation for "a"

$$\Rightarrow \quad \Sigma y = n.a + b \Sigma x \quad \text{---------} (1)$$

n = no. of pair of observation.

The normal equation for "b"

$$\Rightarrow \quad \Sigma xy = a\Sigma x + b\Sigma x^2 \quad \text{——— (2)}$$

solving these two normal equations we can get the required fitted line.

Q ~~Ex.~~ Consider the following data

x	y	x^2	y^2	xy
8	4	64	16	32
3	12	9	144	36
2	1	4	1	2
10	12	100	144	120
11	9	121	81	99
3	4	9	16	12
6	9	36	81	54
5	6	25	36	30
6	1	36	1	6
8	14	64	196	112
62	72	468		503

use the least square method to determine the equation of a linear curve.

Ans Let the linear curve is to be fitted is given by $y = a + bx$.

The normal equations are given by

substituting these values in the normal equation we get.

$$72 = 10 \times a + b \times 62$$
$$72 = 10a + 62b \quad \text{——} (1)$$

$$503 = 62a + 468b \quad \text{——} (2)$$

$$4464 = 620a + 3844b \rightarrow (1) \times 62$$
$$5030 = 620a + 4680b \rightarrow (2) \times 10$$

$$-566 = -836b$$
$$b = 0.677$$

and putting in (1)

$$72 = 10a + 62 \times 0.677$$
$$72 = 10a + 41.974$$
$$a = 3.0026$$

So, the required straight line equation $y = a + bx$

$$\Rightarrow y = 3.0026 + 0.677x.$$

- **Exponential curve fitting :-**

The equation of curve is given by

$$y = a \cdot b^x$$

taking log in both sides

$$\log y = \log a + x \log b$$

$$Y = A + xB$$

$Y = \log y$
$A = \log a$
$B = \log b$

The normal equations :-

$$\Sigma y = nA + B\Sigma x$$
$$\Sigma xy = A\Sigma x + B\Sigma x^2$$

$A = \log a$ | $B = \log b$
$a = A\text{-antilog}(A)$ | $b = A\text{-antilog}(B)$

Q Fit an exponential curve $y = ab^x$ by the method of least square for the following data.

XY	X	Y	$X = \log(x)$	$Y = \log(Y)$	X^2
0.434	2	27.8	0.301	1.444	0.090
0.855	3	62.1	0.477	1.793	0.227
1.229	4	110	0.602	2.041	0.362
1.542	5	161	0.699	2.206	0.488
4.05	14		2.079	7.484	1.169

$$\Sigma y = nA + B\Sigma x$$

$$7.484 = 4A + B\,2.079 \quad \text{——} (1)$$

$$\Sigma xy = A\Sigma x + B\Sigma x^2$$

$$4.05 = 2.079A + 1.169B \quad \text{——} (2)$$

$$A = 0.868$$
$$B = 0.285$$

So, $a = \text{Antilog}\,(0.868)$
$= 7.379$

$b = \text{Antilog}\,(0.285)$
$= 1.927.$

So, $y = a \cdot b^x$

$y = 7.379 \cdot (1.927)^x$

- **Spearman Rank correlation coefficient:**

Case I: **Non-tie case:**

The Spearman rank correlation coefficient for non-tie case is given by

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

where, d_i = difference b/w two series of ranks

n = no. of Observation.

Q The marks obtained by 8 students in maths and physics in a test are given below.

Ans

Marks in maths	Rank in maths (x_i)	Marks in phy	Rank in phy (y_i)	$d_i = x_i - y_i$	d_i^2
43	7	36	8	-1	1
77	3	68	3	0	0
64	5	49	6	-1	1
96	1	79	2	-1	1
48	6	50	5	1	1
35	8	41	7	1	1
86	2	82	1	1	1
71	4	65	4	0	0
					6

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

$$R = 1 - \frac{6 \times 6}{8(8^2-1)}$$

$$= 1 - \left(\frac{36}{8 \times 63}\right)$$

$$= 0.928$$

Case 2: Tie Case

The rank correlation is given by

$$R = 1 - \frac{6\left[\sum d_i^2 + \sum \frac{t(t^2-1)}{12}\right]}{n(n^2-1)}$$

where, t = length of tie

= no. of tie.

9/9/23

Q1 In the following table are recorded data showing the test scores made by 10 salesmen on an intelegent test and their weekly sales.

Calculate the rank correlation coefficient b/w the intelegence & efficiency & in salesmanship.

Serial no.	Intelegence Test score	Rank(x_i)	Sales amount	Rank(y_i)	$d_i = x_i - y_i$	d_i^2
1	50	8.5	25	9	-0.5	0.25
2	70	3	60	1	2	4
3	50	8.5	45	5	3.5	12.25
4	60	5	50	3	2	4
5	80	2	45	5	-3	9
6	50	8.5	20	10	-1.5	2.25
7	90	1	55	2	-1	1
8	50	8.5	30	7.5	1	1
9	60	5	45	5	0	0
10	60	5	30	7.5	-2.5	6.25

Page No.
Date

ang for 60 $= \frac{4+5+6}{3}$

$= \frac{15}{3} = 5$

ang for 50 $= \frac{7+8+9+10}{4}$

$= \frac{34}{4} = 8.5$

ang for 45 $= \frac{4+5+6}{3}$

$= \frac{15}{3} = 5$

$\Sigma d_i^2 = 40$

Rank	length
5	3
8.5	4
5	3
7.5	2

$\Sigma \frac{t(t^2-1)}{12}$

$= \frac{3(3^2-1)}{12} + \frac{4(4^2-1)}{12} + \frac{3(3^2-1)}{12} + \frac{2(2^2-1)}{12}$

$= 2 + 5 + 2 + \frac{1}{2}$

$$= \frac{4+10+4+1}{2} = \frac{+19.5}{2} = 9.5.$$

$$R = 1 - \frac{6\left[\Sigma d_i^2 + \frac{\Sigma t(t^2-1)}{12}\right]}{n(n^2-1)}$$

$$= 1 - \frac{6[40+9.5]}{10(100-1)}$$

$$= 1 - 0.3$$
$$= 0.7$$

Q2 10 competitors in a musical contest were ranked by 3 ~~judges~~ judges A, B, C

Rank by A (x_i)	Rank by B (y_i)	Rank by C (z_i)	$u_i = x_i - y_i$	u_i^2	$v_i = y_i - z_i$	v_i^2	$w_i = x_i - z_i$	w_i^2
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

Using rank correlation method discuss which pair of judges

Page No.
Date

Spy the nearest approach to common likings in music.

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

$$= 1 - \frac{6 \times 20\cancel{0}}{10 \times 99}$$

$$= 1 - \frac{120}{99}$$

$$= \cancel{-0.33} - 0.21$$

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

$$= 1 - \frac{6 \times 214}{10 \times 99}$$

$$= 1 - \frac{1284}{990}$$

$$= -0.29$$

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

$$= 1 - \frac{6 \times 60}{10 \times 99}$$

$$= 1 - \frac{36}{99} = 0.636 = 0.64$$

Since 0.64 is the largest; judges A & C has the nearest approach to common ~~Scale~~ likings in music.

Q3 For some bivariate data the following results were obtained $\bar{x} = 53.2$, $\bar{y} = 27.9$, the regression coefficient of y on x $= -1.5$ & that of x on y $= -0.2$

Calculate i) correlation coefficient b/w x & y

ii) find the most probable value of y when $x = 60$.

Ans i) $b_{yx} = -1.5$

$b_{xy} = -0.2$

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

$$= -\sqrt{(-1.5) \times (-0.2)}$$

$$= -\sqrt{0.3}$$

$$= -0.54$$

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

$y - 27.9 = -1.5(60 - 53.2)$

$y = -1.5(60 - 53.2) + 27.9$

$y = 17.7$

Q4 The regression line of y on x is given by $y = 32 - x$ & the regression line of x on y is given by $x = 13 - 0.25y$

Find correlation coefficient.

Ans

y on x	x on y
$b_{yx} = -1$	$b_{xy} = -0.25$

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

$r = \pm\sqrt{b_{yx} \times b_{xy}}$

$= -\sqrt{(-1) \times (-0.25)}$

$= -0.5$

Q5 The two regression lines are given by $8x - 10y + 66 = 0$ & $40x - 18y = 214$

Find correlation coefficient.

Ans Let $8x - 10y + 66 = 0$ y on x

$-10y = -66 - 8x$ $(y = mx + c)$

$10y = 66 + 8x$

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

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

& $40x - 18y = 214$

$40x = 214 + 18y$ $(x = my + c)$

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

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

$r = +\sqrt{\frac{8}{10} \times \frac{18}{40}}$

$= \sqrt{0.36}$

$= 0.6$