UNIT - 5 Correlation & Regression

[1 marks]

1 (a) Write the formula for Karl Pearson's coefficient of correlation

Sol: $$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}}$$

where $X = (x - \bar{x})$

$Y = (y - \bar{y})$

$\bar{x}$ & $\bar{y}$ are means of the series x and y

(b) Define Regression.

Sol: The statistical method which helps us to estimate the unknown value of one variable from the known value of the related variable is called regression

---

(c) Write the formula for Rank Correlation (Spearman's rank correlation)

Sol:

$$\rho = 1 - \frac{6 \Sigma D^2}{N(N^2 - 1)}$$

where $\rho \rightarrow$ Rank of coeff of correlation

$D^2 \rightarrow$ Sum of the squares of the difference of two ranks

$N \rightarrow$ Number of paired observations

(d) Write the applications of regression

Sol: (1) It is widely used for prediction purpose

(2) It is a highly valuable tool in Economics & Business

(3) We can calculate coefficient of correlation and coefficient of determination with the help of the regression coefficient.

(e) Write the formula for the regression equation of X on Y Normal equations are (2)

Sol:

$$\Sigma X = Na + b\Sigma Y$$

$$\Sigma XY = a\Sigma Y + b\Sigma Y^2$$

[3 marks]

(2) (a) Define Correlation and Types.

Sol: Correlation is a statistical analysis which measures and analyses the degree (or) extent to which two variables fluctuate with reference to each other.

Types: Correlation is classified into many types.

(1) positive & negative
(2) simple & multiple
(3) partial & total
(4) linear & non-linear

(b) Given $n=10$, $\sigma_x = 5.4$, $\sigma_y = 6.2$ and sum the product of deviation from the mean of X and Y is 66 find the correlation co-efficient

Sol: given $n=10$, $\sigma_x = 5.4$, $\sigma_y = 6.2$

$$\sigma_x^2 = \frac{1}{n}\Sigma(x-\bar{x})^2 \quad \text{---------- (1)}$$

$$(5.4)^2 10 = \Sigma(x-\bar{x})^2$$

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

$$(6.2)^2 10 = \Sigma(y-\bar{y})^2 \quad \text{---------- (2)}$$

$$\Sigma(x-\bar{x})(y-\bar{y}) = 66 \quad \text{---------- (3)}$$

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2 \Sigma(y-\bar{y})^2}} = \frac{66}{(5.4)(6.2)}$$

$$\boxed{r = 0.1971}$$

(c) Write the properties for rank Correlation coefficient. (3)

Sol: (i) The value of ρ lies between +1 and -1

(ii) If $\rho = 1$, there is complete agreement in the order if the ranks and the directions of the rank is same

(iii) If $\rho = -1$ then there is complete disgreement in the order of the ranks and they are in opposite directions

d) Difference between correlation & regression

Sol: The correlation coeff is a measure of degree of covariability between two variables, while the regression establishment a functional relation between dependent and independent variables. So that the former can be predected for a given value of the later.

In a corelation, both the variables x & y are random variables, whereas in regression, x is a random variable and y is fixed variable. The coefficient of correlation is a relative measure where as Regression coefficient is an absolute figure.

(e) (2) Following are rank obtained by 16 students in two subjects statistics and math [3 marks]

(1,1) (2,10) (3,3) (4,4) (5,5) (6,7) (7,2) (8,6) (9,8) (10,11) (11,15) (12,9) (13,14) (14,12) (15,16) (16,13)

To what extent the knowledge of students is related?

Solution:

X	Y	D = X − Y	D^2
1	1	0	0
2	10	−8	64
3	3	0	0
4	4	0	0
5	5	0	0
6	7	−1	1
7	2	5	25
8	6	2	4
9	8	1	1
10	11	−1	1
11	15	−4	16
12	9	+3	9
13	14	−1	1
14	12	2	4
15	16	−1	1
16	13	3	9

Rank correlation coefficient formula

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2-1)}$$

$$= 1 - \frac{6(136)}{16(256-1)}$$

$$= 1 - \frac{816}{4080}$$

$$= 1 - 0.2$$

$$= 0.8$$

$$\boxed{\rho = 0.8}$$

[5 marks]

3) (a) Calculate the coefficient of correlation between age of cars (X) and annual maintenance cost (Y) and comment:

X	2	4	6	7	8	10	12
Y	1600	1500	1800	1900	1700	2100	2000

Sol:

let age of cars = x

Annual maintenance = y

Computation of coefficient of correlation

x	y	$X = x - \bar{x}$, $\bar{x} = 7$	$Y = (y - \bar{y})/100$, $\bar{y} = 1800$	X^2	Y^2	XY
2	1600	-5	-2	25	4	10
4	1500	-3	-3	9	9	9
6	1800	-1	0	1	0	0
7	1900	0	1	0	1	0
8	1700	1	-1	1	1	-1
10	2100	3	3	9	9	9
12	2000	5	2	25	4	10
$\Sigma x = 49$	$\Sigma y = 12600$	$\Sigma X = 0$	$\Sigma Y = 0$	$\Sigma X^2 = 70$	$\Sigma Y^2 = 28$	$\Sigma XY = 37$

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}} = \frac{37}{\sqrt{70 \times 28}}$$

$$\boxed{r = 0.836}$$

Comment: we can observe that there is a high degree of positive correlation between age of cars and annual maintenance cost.

(b) From sample of 200 pairs observation the following quantities were calculated:

$\Sigma X = 11.34$ $\Sigma Y = 20.72$ $\Sigma X^2 = 12.16$ $\Sigma Y^2 = 84.96$

$\Sigma XY = 22.13$. From the above data show how to compute the coefficient of the equations $Y = a + bX$

Solution:

Sol: We can compute the coefficients of the eqt $Y = a + bX$ by solving the normal eqt's:

$$\Sigma Y = na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

now, substituting the values

$\Sigma Y = na + b\Sigma X \longrightarrow 20.72 = 200a + 11.34b$

$\Sigma XY = a\Sigma X + b\Sigma X^2 \longrightarrow 22.13 = 11.34a + 12.16b$

now solving,

$$20.72 = 200a + 11.34b$$

$$\frac{20.72 - 11.34b}{200} = a$$

$$a = 0.1036 - 0.0567b$$

$$= 0.1036 - 0.0567(1.82)$$

$$\boxed{a = 0.0005}$$

now, $22.13 = 11.34a + 12.16b$

$$22.13 = 11.34(0.1036 - 0.0567b) + 12.16b$$

$$22.13 = 1.175 - 0.643b + 12.16b$$

$$20.955 = 11.517b$$

~~$b = 11.51$~~

$$b = \frac{20.955}{11.517} = \boxed{b = 1.82}$$

(c) Determine the equation of a straight line which best fits the data.

X	10	12	13	16	17	20	25
Y	10	22	24	27	29	33	37

Solution:

Let the required straight lines $Y = a + bX$

the two normal equations are

$$\Sigma Y = Na + b\Sigma X \quad \text{——} (1)$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 \quad \text{——} (2)$$

X	X^2	Y	XY
10	100	10	100
12	144	22	264
13	169	24	312
16	256	27	432
17	289	29	493
20	400	33	660
25	625	37	925
$\Sigma X = 113$	$\Sigma X^2 = 1983$	$\Sigma Y = 182$	$\Sigma XY = 3186$

$N = 7$

$\Sigma X = 113$

$\Sigma X^2 = 1983$

$\Sigma Y = 182$

$\Sigma XY = 3186$

from ①

$182 = 113b + 7a$ —— ③

from ②

$3186 = 1983b + 113a$ —— ④

113 multiply in equ ③ $\Rightarrow 12769b + 791a = 20566$ —— ⑤

7 multiply in equ ④ $\Rightarrow 13881b + 791a = 22302$ —— ⑥

equ ⑥ - ⑤ $\Rightarrow 1112b = 1736$

$b = \frac{1736}{1112}$ $\boxed{b = 1.56}$

from ③ $\Rightarrow 182 = 113b + 7a$

$182 = 113(1.56) + 7a$

$a = \frac{5.72}{7}$

$\boxed{a = 0.817}$

the equ of straight lines is $Y = a + bX$

$\boxed{Y = 0.817 + 1.56X}$

This equ is required straight line. This is called regression equation of Y on X.

(d) Find the most likely production corresponding to a rainfall for the following data. [5 marks]

	Rainfall (X)	Production (Y)
Average	30	500 kgs
Standard deviation	5	100 kgs
Coefficient of correlation	0.8	

Solution:

Sol: We have to calculate the value of Y when X = 40

So we have to find the regression equation of Y on X

Mean of X series, $\bar{X} = 30$

Mean of Y series, $\bar{Y} = 500$

σ of X series, $\sigma_x = 5$

σ of Y series $\sigma_y = 100$

Regression of Y on X

$$Y - \bar{Y} = r \cdot \frac{\sigma_x}{\sigma_y} (X - \bar{X})$$

$$Y - 500 = (0.8)\frac{5}{100}(X - 30)$$

when $X = 40$,

$$Y - 500 = \frac{4}{100}(40 - 30) = \frac{40}{100}$$

$$Y = 500 + \frac{4}{10} = 500.4$$

Hence the expected value of Y is 500.4 kg

(3) (e) Following are the rank obtained by 10 students in two subjects, Statistics (X) and Mathematics (Y). To what extent the knowledge of the students in two subjects is related.

X	1	2	3	4	5	6	7	8	9	10
Y	2	4	1	5	3	9	7	10	6	8

Solution:

X	Y	D = X - Y	D^2
1	2	-1	1
2	4	-2	4
3	1	2	4
4	5	-1	1
5	3	2	4
6	9	-3	9
7	7	0	0
8	10	2	4
9	6	3	9
10	8	2	4

Rank correlation coefficient formula:

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2-1)}$$

$$= 1 - \frac{6(40)}{10(10^2-1)}$$

$$= 1 - \frac{160}{10(100-1)} = 1 - \frac{160}{10(99)}$$

$$= 1 - \frac{160}{990}$$

$$= \boxed{0.76}$$

(4) (a) Calculate the Karl Pearson's coefficient of correlation for the following paired data. What inference would you draw from the estimate?

X	28	41	40	38	35	33	40	32	36	33
Y	23	34	33	34	30	26	28	31	36	38

Solution:

computation of correlation coefficient.

Computation of correlation coefficient

x	$X = x - 35$ $X = x - \bar{x}$	X^2	y	$Y = y - 31$ $Y = y - \bar{y}$	Y^2	XY
28	−7	49	23	−8	64	56
41	6	36	34	3	9	18
40	5	25	33	2	4	10
38	3	9	34	3	9	9
35	0	0	30	−1	1	0
33	−2	4	26	−5	25	10
40	5	25	28	−3	9	−15
32	−3	9	31	0	0	0
35	1	1	36	5	25	5
33	2	4	38	7	49	−14
$=355$	$\Sigma X = 6$	$\Sigma X^2 = 162$	$\Sigma Y = 313$	$\Sigma Y = 3$	$\Sigma Y^2 = 195$	$\Sigma XY = 8$

Here $N = 10$, ~~take~~ $\bar{x} = 35$ & $\bar{y} = 31$

WKT: $$r = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{N}}{\sqrt{\left(\Sigma X^2 - \frac{(\Sigma X)^2}{N}\right) \times \left(\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}\right)}}$$

$$= \frac{89 - \frac{6(3)}{10}}{\sqrt{162 - \frac{6^2}{10}} \sqrt{195 - \frac{3^2}{10}}} = \frac{89 - \frac{6(3)}{10}}{\sqrt{162 - \frac{36}{10}} \sqrt{195 - \frac{9}{10}}}$$

$$\boxed{r = 0.45}$$

(b) A sample of 12 fathers (F) and their elder sons gave the following data about their elder sons (S). Calculate the coefficient of correlation

F	65	63	67	64	68	62	70	66	68	67	69	71
S	68	66	68	65	69	66	68	65	71	67	68	70

Solution: We write the given values,

f	S	Rank	Rank	$d_i = x_i - y_i$	d_i^2
65	68	9	5.5	3.5	12.25
63	66	11	9.5	1.5	2.25
67	68	6.5	5.5	1.0	1
64	65	10	11.5	-1.5	2.25
68	69	4.5	3	1.5	2.25
62	66	12	9.5	2.5	6.25
70	68	2	5.5	-3.5	12.25
66	65	8	11.5	3.5	12.25
68	71	4.5	1	-3.5	12.25
67	67	6.5	8	-1.5	2.25
69	68	3	5.5	-2.5	6.25
71	70	1	2	-1	1

$\Sigma d_i^2 = 72.5$

Repeated values are given common rank, which is the mean of the ranks then would have got. In X 68 and 67 appear twice.

Correction factor in X series,

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

In Y 68 appears 4 times, 66 appears twice and 65 appears twice.

Correction factor in Y series,

$$\frac{4(4^2-1)}{12} + \frac{2(2^2-1)}{12} + \frac{2(2^2-1)}{12} \Rightarrow 6$$

Rank correlation coefficient

$$r = 1 - \frac{6(72.5+7)}{12(12^2-1)}$$

$$\boxed{r = 0.722}$$

(c) Given the bi-variate data:

X	1	5	3	2	1	1	7	3
Y	6	1	0	0	1	2	1	5

I) find the regression line of Y on X and hence predict Y if X = 10

II) fit a regression line of X on Y and hence predict X if Y = 2.5

Solution:

X	Y	X^2	X^2	XY
1	6	36	1	6
5	1	1	25	5
3	0	0	9	0
2	0	0	4	0
1	1	1	1	1
1	2	4	1	2
7	1	1	49	7
3	5	25	9	15
$\Sigma X = 23$	$\Sigma Y = 16$	$\Sigma Y^2 = 68$	$\Sigma X^2 = 99$	$\Sigma XY = 36$

$\Sigma X = 23$; $\Sigma Y = 16$; $\Sigma X^2 = 99$, $\Sigma Y^2 = 68$

$\Sigma XY = 36$

(i) Y on X

$Y = a + bX$

given $\boxed{X = 10}$

WKT

$$\Sigma Y = Na + b\Sigma X \quad \text{——} \textcircled{1}$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 \quad \text{——} \textcircled{2}$$

we get

$16 = 8a + 23b$ (from eqt ①)

$36 = 23a + 99b$ (from eqt ②)

Solving the above equations we get,

$\boxed{b = -0.304}$ $\boxed{a = 2.874}$

now,

$Y = a + bX$ ($X = 10$ given)

$Y = 2.874 - 0.304(10)$

$Y = 2.874 - 3.04$

$\boxed{Y = -0.166}$

(ii) X on Y

$X = a + bY$

given $\boxed{Y = 2.5}$

WKT

$\Sigma X = Na + b\Sigma Y \longrightarrow 23 = 8a + 16b$ — (3)

$\Sigma XY = a\Sigma Y + b\Sigma Y^2 \longrightarrow 36 = 16a + 68b$ — (4)

solving eqt (3) & (4) we get

$\boxed{a = 3.430}$ $\boxed{b = -0.277}$

now,

$X = a + bY$ ($Y = 2.5$ given)

$X = 3.430 + (-0.277)(2.5)$

$= 3.430 - 0.692$

$\boxed{X = 2.738}$