UNIT-5 CORRELATION AND REGRESSION

1 (1 mark)

a) Write the formula for Karl pearson's coefficient of correlation.

There are several formulas to calculate Karl pearsons coefficient of correlation (r)

(1) $r = \frac{\text{co-variance of } xy}{\sigma_x \sigma_y}$ (2) $r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}}$ (3) $r = \frac{\Sigma XY}{N \sigma_x \sigma_y}$

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

where $\bar{x}, \bar{y}$ are means of series x & y

$\sigma_x \rightarrow$ S.D of series x

$\sigma_y \rightarrow$ S.D of series y

1.b) Define Regression.

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"

1c) Write the formula for rank correlation (Spearman's rank correlation)

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

$\rho \rightarrow$ rank coefficient of correlation

$D^2 \rightarrow$ sum of squares of differences of two ranks

$N \rightarrow$ Number of paired observations.

$$\rho = 1 - 6\left\{\frac{\Sigma D^2 + \frac{1}{12}(m^3-m) + \frac{1}{12}(m^3-3) + \cdots}{N^3 - N}\right\}$$

$m \rightarrow$ no. of items whose ranks are common.

1d) Write the applications of Regression.

* It is used to estimate the relation between two economic variables like Income and expenditure.
* It is highly valuable tool in Economics and Business.
* It is widely used for prediction purpose.
* We can calculate coefficient of correlation and coefficient to determination with the help of the regression coefficient.

1.e) Write the formula for the regression equation of x on y.

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

By solving we get $\boxed{X = a + bY}$

If the actual mean is fraction

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

$$r\frac{\sigma_x}{\sigma_y} = \frac{\Sigma dx\,dy - \frac{\Sigma dx \Sigma dy}{N}}{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}}$$

actual mean is not fraction

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

$$r \cdot \frac{\sigma_x}{\sigma_y} = \frac{\Sigma XY}{\Sigma Y^2}$$

(3 marks)

2.a) Find the coefficient of correlation between x and y for the following data.

X	10	12	18	24	23	27
Y	13	18	12	25	30	10

Let, x = X values

y = Y values

The actual means are

$$\bar{X} = \frac{\Sigma x_i}{N} = \frac{114}{6} = 19$$

$$\bar{Y} = \frac{\Sigma y_i}{N} = \frac{108}{6} = 18$$

x	y	$X = x - \bar{X}$	$Y = y - \bar{Y}$	x^2	y^2	XY
10	13	-9	-5	81	25	45
12	18	-7	0	49	0	0
18	12	-1	-6	1	36	6
24	25	5	7	25	49	35
23	30	4	12	16	144	48
27	10	8	-8	64	64	-64
114	108			236	318	70

The coefficient of correlation is

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}} = \frac{70}{\sqrt{236 \times 318}}$$

$$= \frac{70}{273.94} = 0.255 //$$

2.b) From the sample of 200 people 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 $Y = a + bX$

let,

the required straight line is

$$Y = a + bX$$

The two normal eqn's are $\Sigma Y = b\Sigma X + Na$

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

Substituting values,

$$20.72 = 11.34b + 200a \longrightarrow ①$$

$$22.13 = 12.16b + 11.34a \longrightarrow ②$$

From eq ①

$$a = \frac{20.72 - 11.34b}{200} \longrightarrow ③$$

Sub ③ in ②

$$\Rightarrow 22.13 = 12.16b + 11.34\left(\frac{20.72 - 11.34b}{200}\right)$$

$$\Rightarrow 22.13 = 12.16b + 11.34\,(0.1036 - 0.567b)$$

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

$$\Rightarrow 20.955 = 11.517b$$

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

$$a = \frac{20.72 - 11.34(1.82)}{200}$$

$$= 0.1036 - 0.0567(1.82)$$

$$\Rightarrow \boxed{a = 0.0005}$$

The equation of straight line is

$$\boxed{Y = 0.0005 + 1.82X}$$

2.9) Calculate the regression equations of Y on X from the data given below taking deviations from actual means of X on Y.

Price	10	12	13	12	16	15
Amount demanded	10	22	24	27	29	33

let, x = price

y = amount demanded

The actual means are

$$\bar{x} = \frac{\Sigma x_i}{N} = \frac{78}{6} = 13$$

$$\bar{y} = \frac{\Sigma y_i}{N} = \frac{145}{6} = 24.16 \approx 24$$

x	y	$dx = x - \bar{x}$	$dy = y - \bar{y}$	dx^2	dy^2	$dxdy$
10	10	-3	-14	9	196	42
12	22	-1	-2	1	4	2
13	24	0	0	0	0	0
12	27	-1	3	1	9	-3
16	29	+3	5	9	25	15
15	33	2	9	4	81	18
78	145	0	1	24	315	74

Regression eqn of Y on X is

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$b_{yx} = \frac{\Sigma dxdy - \frac{\Sigma dx \Sigma dy}{N}}{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}} = \frac{74 - \left(\frac{0 \times 1}{6}\right)}{24 - \frac{(0)}{6}}$$

$$= \frac{74}{24} = 3.084$$

Regression equation is

$$Y - 24 = 3.084(X - 13)$$

$$\Rightarrow Y = 24 + 3.084X - 40.092$$

$$\Rightarrow \boxed{Y = 3.084X - 16.092}$$

Regression eqn of X on Y is

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$b_{xy} = \frac{\Sigma dx dy - \frac{\Sigma dx \Sigma dy}{N}}{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}} = \frac{74 - \left(\frac{0 \times 1}{6}\right)}{314 - \frac{(1)^2}{6}}$$

$$= \frac{74}{\frac{1883}{6}} = \frac{74 \times 6}{1883} = 0.235$$

Regression equation is

$$X - 13 = 0.235(Y - 24)$$

$$\Rightarrow X = 13 + 0.235Y - 5.64$$

$$\Rightarrow \boxed{X = 0.235Y + 7.36}$$

2d) Difference between correlation and Regression.

Correlation	Regression
* Measures the strength and direction of a linear relation-ship between two variables.	* Predicts the value of one variable based on another variable.
* The relationship is symmetric the correlation is same whether you treat one variable as independent / dependent variable.	* The relationship is assymetric one variable is considered the idependent and the other is dependent.
* Results in correlation coefficient (r) between -1 and +1, indicating strength of direction of the relationship.	* Provides an equation ($y = mx + c$) that describes the relationship, allowing for predictions of dependent variable based on independent variable.

2.e) The rank of 16 students in Mathematics and statistics are as follows (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). Calculate the rank correlation coefficient for proficiencies of this group in Mathematics and the statistics.

let, x = rank of mathematics

y = rank of statistics.

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
			136

The rank correlation is

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

$$= 1 - \frac{\cancel{100}\ \overset{3}{\cancel{8}}\ (\overset{17}{\cancel{136}})}{\underset{8}{\cancel{16}}\ (256-1)} = 1 - \frac{\cancel{3} \times 17}{\cancel{255}_{85}}$$

$$= \frac{85-17}{85} = \frac{68}{85} = 0.8$$

∴ The rank correlation is 0.8 //

(5 marks)

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

Age of cars	2	4	6	7	8	10	12
Annual cost	1600	1500	1800	1900	1700	2100	2000

let, x = age of cars

y = Annual cost

the actual means are

$$\bar{x} = \frac{\Sigma x_i}{N} = \frac{49}{7} = 7$$

$$\bar{y} = \frac{\Sigma y_i}{N} = \frac{12600}{7} = 1800$$

x	y	$X = x - \bar{x}$	$Y = \frac{(y-\bar{y})}{100}$	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
				70	28	37

The Karl pearson's coefficient of correlation is

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

$$= \frac{37}{44.27} = 0.836 //$$

3.b) Find Karl pearson's coefficient of correlation from the following data.

Wages	100	101	102	102	100	99	97	98	96	95
Cost of living	98	99	99	97	95	92	95	94	90	91

Let, x = wages

y = cost of living

The actual means are

$$\bar{x} = \frac{\Sigma x_i}{N} = \frac{990}{10} = 99$$

$$\bar{y} = \frac{\Sigma y_i}{N} = \frac{950}{10} = 95$$

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	X^2	Y^2	XY
100	98	1	3	1	9	3
101	99	2	4	4	16	8
102	99	3	4	9	16	12
102	97	3	2	9	4	6
100	95	1	0	1	0	0
99	92	0	-3	0	9	0
97	95	-2	0	4	0	0
98	94	-1	-1	1	1	1
96	90	-3	-5	9	25	15
95	91	-4	-4	4	16	16
990	950			54	96	61

The Karl pearson's coefficient of correlation is

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}} = \frac{61}{\sqrt{54 \times 96}}$$

$$= \frac{61}{72} = 0.847 //$$

3.5) 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

Let, the required straight line is

$$Y = a + bX$$

The two normal equations are $\Sigma Y = b\Sigma X + Na$

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

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

Substituting the values

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

$\Sigma Y = 182$, $\Sigma X = 113$, $N = 7$

$$113b + 7a = 182 \longrightarrow (1)$$

$\Sigma XY = 3186$, $\Sigma X^2 = 1983$, $\Sigma X = 113$

$$1983b + 113a = 3186 \longrightarrow (2)$$

eq① × 113

$$12769b + 791a = 20566 \longrightarrow ③$$

eq② × 7

$$13881b + 791a = 22302 \longrightarrow ④$$

Eq ③ − ④

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

$$\Rightarrow a = 0.82$$

Eqn of straight line is $Y = a + bX$

$a = 0.82$

$b = 1.56$

$$\boxed{Y = 0.82 + 1.56X}$$

3d) Find the most likely production corresponding to a rainfall 40 from the following data.

	Rainfall	Production
Average	30	500 kgs
Standard deviation	5	100 kgs.
coefficient of correlation	0.8	

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

So, we should find regression equation of Y on X

mean of X-series $(\bar{X}) = 30$

mean of Y-series $(\bar{Y}) = 500$

S.D of X-series $\sigma_x = 5$

S.D of Y-series $\sigma_y = 100$

Regression of Y on X

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

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

when, $X = 40$,

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

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

$$\Rightarrow Y - 500 = \frac{4}{10}$$

$$\Rightarrow Y = 500 + 0.4$$

$$\Rightarrow \boxed{Y = 500.4}$$

∴ The expected value of Y is 500.4kg.

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

let, x = rank of statistics

y = rank of mathematics

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
			40

The rank correlation is

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

$$= 1 - \frac{6 \times 40}{10(100-1)}$$

$$= 1 - \frac{8^2 \times 4}{99_{33}} = 1 - \frac{8}{33}$$

$$= \frac{25}{33} = 0.76$$

(10 marks)

4.a) ~~Following~~ 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

$$\bar{x} = \frac{\Sigma x_i}{N} = \frac{356}{10} = 35.6 \approx 36$$

$$\bar{y} = \frac{\Sigma y_i}{N} = \frac{313}{10} = 31.3 \approx 31$$

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	X^2	Y^2	XY
28	23	-8	-8	64	64	64
41	34	5	3	25	9	15
40	33	4	2	16	4	8
38	34	2	3	4	9	6
35	30	-1	-1	1	1	1
33	26	-3	-5	9	25	15
40	28	4	-3	16	9	-12
32	31	-4	0	16	0	0
36	36	0	5	0	25	0
33	38	-3	7	9	49	-21
356	313	-4	3	160	195	76

N = no. of items = 10

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

$$= \frac{76 - \left(\frac{-4 \times 3}{10}\right)}{\sqrt{\left(160 - \frac{(-4)^2}{10}\right)\left(195 - \frac{(3)^2}{10}\right)}}$$

$$= \frac{77.2}{175.23} = 0.440 //$$

4b) A sample of 12 fathers (F) and their elder sons gave the following data about their elder sons (S). calculate the coeffient 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

positions	1st	2nd	3rd	4rt	5th	6th	7th	8th	9th	10th	11th	12th
Decreasing order of X	70	70	69	68	68	67	67	66	65	64	62	60
ranks	1	2	3	4.5	4.5	6.5	6.5	8	9	10	11	12

positions	1st	2nd	3rd	4rt	5th	6th	7th	8th	9th	10th	11th	12th
Decreasing order of Y	71	70	69	68	68	68	68	67	66	66	65	65
ranks	1	2	3	5.5	5.5	5.5	5.5	8	9.5	9.5	11.5	11.5

Fathers(X)	sons(Y)	x	y	$D = x - y$	D^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
					72.5

In X-series for 68, $m=2$

for 67, $m=2$

The correlation is $\frac{(2^3-2)}{12} + \frac{(2^3-2)}{12} = \frac{6}{12} + \frac{6}{12} = \frac{12}{12} = 1$

In Y-series for 68, $m=4$

for 66, $m=2$

for 65, $m=2$

The correlation is $\frac{(4^3-4)}{12} + \frac{(2^3-2)}{12} + \frac{(2^3-2)}{12}$

$= \frac{60}{12} + \frac{6}{12} + \frac{6}{12} = \frac{72}{12} = 6$

The rank correlation coefficient is

$$\rho = 1 - 6\left[\frac{\Sigma D^2 + \frac{m^3-m}{12} + \frac{m^3-m}{12} + \cdots}{N^3-N}\right]$$

$$= 1 - \frac{6(72.5+1+6)}{12^3-12} = 1 - \frac{6(79.5)}{1716}$$

286

$$= \frac{286 - 79.5}{286} = \frac{206.5}{285}$$

$$= 0.722 //$$

4c) Given bivariate 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 is Y = 2.5.

The actual mean are

$$\bar{x} = \frac{\Sigma x_i}{N} = 2.87 \approx 3$$

$$\bar{y} = \frac{\Sigma y_i}{N} = 2$$

X	Y	$dx = x - \bar{x}$	$dy = y - \bar{y}$	dx^2	dy^2	dxdy
1	6	-2	4	4	16	-8
5	1	2	-1	4	1	-2
3	0	0	2	0	4	0
2	0	-1	-2	1	4	2
1	1	-2	-1	4	1	2
1	2	-2	0	4	0	0
7	1	4	-1	16	1	-4
3	5	0	3	0	9	0
23	16	-1	0	33	36	-10

The regression eqn of Y on X is

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

$$b_{yx} = \frac{\Sigma dx\,dy - \frac{\Sigma dx \Sigma dy}{N}}{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}}$$

$$= \frac{-10 - \left(\frac{-1 \times (0)}{8}\right)}{33 - \frac{(-1)^2}{8}}$$

$$= \frac{-10}{33 - \frac{1}{8}} = -0.304$$

Regression eqn Y on X is

$$(Y-2) = -0.304\,(X-3)$$

when $X = 10$

$$Y-2 = -0.304\,(10-3)$$

$$Y-2 = -0.304\,(7)$$

$$Y = 2 - 2.128$$

$$\boxed{Y = -0.128}$$

Regression eqn of X on Y is

$$X - \bar{X} = b_{xy}\,(Y - \bar{Y})$$

$$b_{xy} = \frac{\Sigma dx\,dy - \frac{\Sigma dx\,\Sigma dy}{N}}{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}}$$

$$= \frac{-10 - \left(\frac{-1 \times 0}{8}\right)}{36 - \frac{(0)^2}{8}} = \frac{-10}{36}$$

$$= -0.27$$

Regression eqn of X on Y is

$$(X-3) = -0.27\,(Y-2)$$

when $Y = 2.5$

$$X-3 = -0.27\,(2.5-2)$$

$$X-3 = -0.27\,(0.5)$$

$$X = 3 - 0.135$$

$$\boxed{X = 2.865}$$