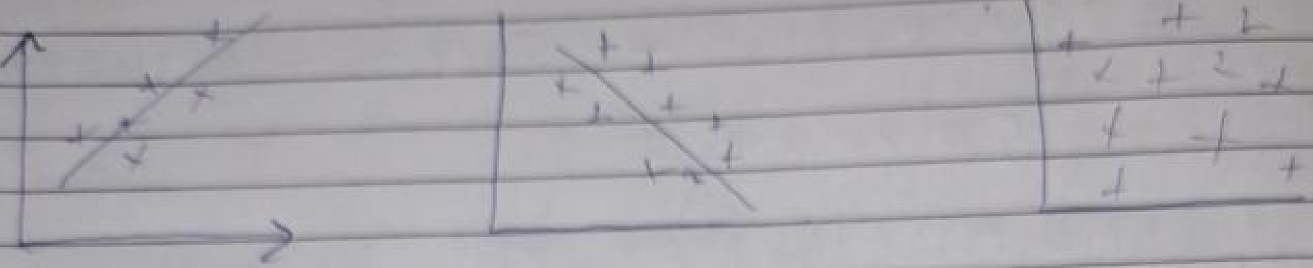
Date :

Correlation
Statistical tool to study relationship between two variables.

→ Positive Correlation   Negative Correlation   No correlation



Methods of Studying Correlation

1] Scatter diagram :- Graphical method, does not show degree of correlation

2] Karl Pearson's correlation coff - measure nature and strength. Value ranges b/w -1 to 1. Denotes strength of association - quantitative data only linearly related

3] Spearson's rank correlation :- uses 2 sets of rank used when both variables are quantitative, qualitative ordinal, one is quantitative and other is qualitative ordinal, not linearly related

Karl Pearson's coefficient
It is the measure of the strength and direction of a linear relationship b/w two variables.

$$\rho = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$$

Date :

n = no of pair of observation

Q]

x	y	xy	x^2	y^2
1	-3	-3	1	9
2	-1	-2	4	1
3	0	0	9	0
4	1	4	16	1
5	2	10	25	4
15	-1	9	55	15

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

$$= \frac{5(9) - (15)(-1)}{\sqrt{5(55) - 15^2}\sqrt{5(15) - (-1)^2}}$$

$$= \frac{60}{\sqrt{50}\sqrt{74}} \approx 0.986$$

coeff = 0.986 shows strong correlation

Q]

x	y	xy	x^2	y^2
64	66	4224	4096	4356
65	67	4355	4225	4489
66	65	4290	4356	4225
67	68	4556	4489	4624
68	70	4760	4624	4900
69	68	4692	4761	4624
70	72	5040	4900	5184
469	476	31917	31917	32402

$n = 7$

$\rho = 0.810$ } strong +ve correlation b/w 2 variables

Spearman's Rank Correlation Coefficient

↳ Both variables are quantitative
↳ Both are qualitative ordinal
↳ One variable is quantitative other is qualitative ordinal

$$\rho(x, y) = 1 - \frac{6\sum D^2}{N(N^2-1)}$$

$D = X - Y$ diff b/w rank of individuals

Case 1 ⟹ When ranks are given

Q]

Ranking acc (R_1) to length of service	Rank acc (R_2) to efficiency	difference ($R_1 - R_2$)	D^2
1	2	-1	1
2	3	-1	1
3	5	-2	4
4	1	3	9
5	9	-4	16
6	10	-4	16
7	11	-4	16
8	12	-4	16
9	8	1	1
10	7	3	9
11	6	~~8~~ 5	~~64~~ 25
12	4	8	64
			$178 = D^2$

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

$$= 1 - \frac{6 \times 178}{12(144-1)} = 1 - \frac{1068}{1716} = 0.378$$

low degree positive correlation b/w length of service & efficiency

Case 2 ⇒ Quantitative Data

Q)

x	y	Rx	Rxy	D	D^2
60	59	2	3	-1	1
59	51	3	5	-2	4
61	63	1	2	-1	1
25	67	5	1	+4	16
48	52	4	4	0	0
					$\Sigma = 22$

$n = 5$

1) Assign ranks

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

$$= 1 - \frac{6 \times 22}{5(24)}$$

$$= -0.1$$

what if Ranks are equal?

x	y	Rx	Ry	D	D^2
10	26	6.5	3.5	3	9
15	25	3	6	-3	9
14	24	5	8	-3	9
10	25	6.5	6	0.5	0.25
9	28	8	2	6	36
15	26	3	3.5	0.5	0.25
16	25	1	6	-5	25
15	30	3	1	2	4
					67.5

$n = 8$

possible position for 15 is 2,3,4 avg $\frac{2+3+4}{3} =$

$$\rho = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}(3^2-3) + \frac{1}{12}(2^2-2) + \frac{1}{12}(2^2-2) + \frac{1}{12}(3^2-3)\right]}{n(n^2-1)}$$

$$= 1 - \frac{6[67.5 + 0.5 + 0.166 + 0.166 + 0.5]}{8 \times 63}$$

$= 0.1807$ not strongly related

Q] Find Rank correl coeff

Inventory	Earning	R_x	R_y	D	D^2
4	11	6	3	2	4
5	9	4.5	4	0.5	0.25
7	13	2	1.5	0.5	0.25
8	7	1	7	-6	36
6	13	3	1.5	1.5	2.25
3	8	7	5.5	1.5	2.25
5	8	4.5	5.5	1	1
					46

$$\rho = 1 - \frac{6\left[46 + \frac{1}{12}(2^2-2) + \frac{1}{12}(2^2-2) + \frac{1}{12}(2^2-2)\right]}{7(7^2-1)}$$

$$= 1 - \frac{6[46 + 0.5)}{7 \times 48}$$

$= 0.169$

Regression

It is used to predict value of one variable based on value of other variable.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

dependent variable, intercept, slope coeff, random error / noise

We need to estimate β_0 & β_1 to find the line that fits all the points well.

least Squares method

$$E = \sum_{i=1}^{n} (y_i - \hat{y}^i)^2$$ y^i = actual $\hat{y}$ = predicted

$$b_0 = \bar{y} - b_1 \bar{x} \qquad b_1 = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$$

$$= \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2}$$

To find b_0 & b_1 we solve these 2 equations

$$\sum y = n b_0 + b_1 \sum x$$

$$\sum yx = b_0 \sum x + b_1 \sum x^2$$