Correlation

→ It deals with the relation between 2 or more variables
→ Helps in understanding nature and strength of the relationship b/w variables and can be used to make predictions value of one variable based on the other

Causation

→ Correlation does not mean causation
→ Causation means that one variable causes other
→ Causation implies that there is direct cause and effect relation b/w 2 variables but correlations shows there is some relation.

Simple Correlation

Study of relationship b/w 2 variables only.

Types

(i) Positive & Negative Correlation

if ↑ in a causes ↑ in b, positive
if ↑ in a causes ↓ in b, negative
if nothing happens then zero correlation

(ii) Linear and Non-linear

if there is a fixed pattern b/w 2 variable
eg: if increase in Revenue by 20% will result
in 2% salary hike, linear

if there is no fixed pattern, non-linear

(iii) Simple, Partial and Multiple

when correlation b/w 2 variables, simple

when variables are 3 but we study only 2, Partial

when we study more than 2 variables, multiple

Methods of Studying

- Scatter Diagram
- Karl Pearson's coefficient of correlation
- Spearman's Rank Correlation Coefficient

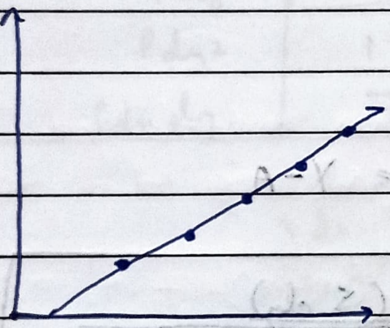
⇒ Correlation lies b/w -1 and 1
⇒ if correlation is -1, 1 then perfect correlation
⇒ if b/w 0.75 to 1 & -0.75 to -1 then high degree
⇒ if b/w 0.25 to 0.75 & -0.25 to -0.75 moderate
⇒ if b/w 0 to 0.25 & 0 to -0.25, low
⇒ if 0 then no correlation

Scatter Diagram

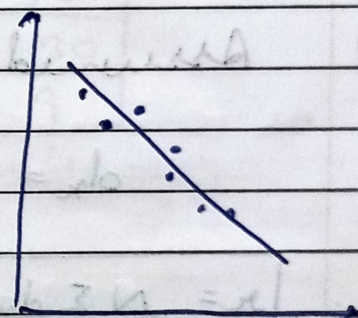
→ simple method of diagramatic representation to determine the nature of correlation b/w 2 variables only.

→ works only on individual and discrete for continuous to be converted into mid-pt.

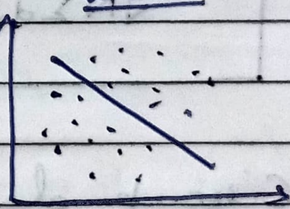
Perfect Positive



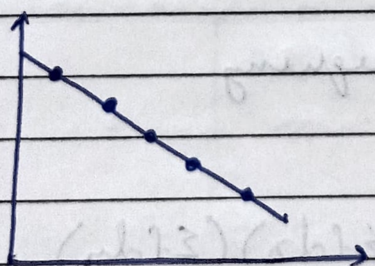
Negative



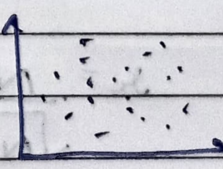
Weak



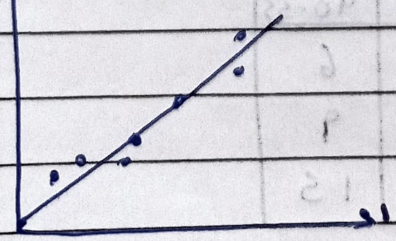
Perfect Negative



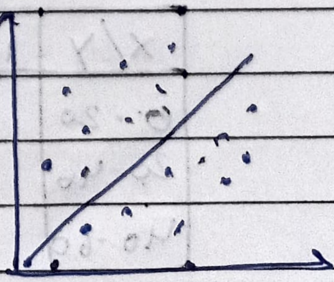
No Correlation



~~Perd~~ Positive (strong)



Weak



Karl Pearsons Coefficient

$$\text{Co-variance}(X, Y) = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{N}$$

① $r = \cancel{\frac{\text{cov}(X,Y)}{\sqrt{\ldots}}}$

$$\boxed{r = \frac{\text{cov}(X, Y)}{\sigma_X \cdot \sigma_Y} \qquad r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2} \cdot \sqrt{\Sigma(Y-\bar{Y})^2}}}$$

② Assumed Mean

$$d_x = X - A \qquad d_y = Y - A$$

$$\boxed{r = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}}}$$

③ Group data having frequency

$$r = \frac{N\Sigma f d_x d_y - (\Sigma f d_x)(\Sigma f d_y)}{\sqrt{N\Sigma f d_x^2 - (\Sigma f d_x)^2}\sqrt{N\Sigma f d_y^2 - (\Sigma f d_y)^2}}$$

X/Y	10-25	25-40	40-55
0-20	10	4	6
20-40	5	40	9
40-60	3	8	15

make standard table of 3×3

always 4

~~4 division due to 4 columns~~

				10-25	25-40	40-55	x			
x \ y			mid value →	17.5	32.5	47.5				
			diff of common value from →	-15	0	15	Sum of all f			
	mid value	value from 20	divide by common diff →	-1	0	1	f	fdx	fdx²	fdxdy
0-20	10	-20	-1	1 → 10 → 10	0 → 4 → 0	-1 → 6 → -6	20	-20	20	4
20-40	30	0	0	0 → 5 → 0	0 → 10 → 0	0 → 9 → 0	34	0	0	0
40-60	50	20	1	-1 → 3 → -3	0 → 8 → 0	1 → 15 → 15	26	26	26	16
			f	18	52	30				
			fdy	-18	0	30				
			fdy²	18	0	30				
			fdxdy	7	0	9				

always 4

same 3 division

3 division due to 3 intervals

~~4 division~~

Spearman's Rank Correlation

$$\boxed{p = 1 - \frac{6 \sum D^2}{n(n^2-1)}}$$

, D is the difference of ranks of 2 variables

⇒ **When Ranks are not Give**

Give ranks from 1 in the descending order and apply the formula

When Ranks are Repeated

X	Ranks
15	5
10	7.5
20	2
28	1
12	6
10	7.5
16	4
18	3

As 10 is repeating and we have to give 7 and 8th rank then we will give average of it to these

$$\rho = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \_\_\right]}{n(n^2-1)}$$

m_1 is number of times a rank is repeated
i.e. 10 is 2 times so $m_1 = 2$
It is done for each repeating number.

__/__/__

Regression

⇒ Statistical technique that expresses the relationship between 2 or more variables in the form of eq^n to estimate value of a variable on the basis of another variable.

⇒ Variable whose value is estimated is dependent and whose value is used to estimate is independent.

Regression Vs Correlation

⇒ Correlation measures strength of the relationship b/w two variables

⇒ Regression helps in predicting the value of dependent variable based on independent variable.

Assumptions

① There should be a linear relationship b/w both variables

② The error should be normally distributed

~~③ The independent variables should~~

③ The variance of error should be constant across all the values of independent variable

④ The observation should be independent of each other

Using Normal Eq^n (Least Square Method)

X on Y ($X = a + bY$)

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

Y on X ($Y = a + bX$)

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

used to find variable a and b

Regression Eq^n

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

$$b_{xy} = \frac{N\Sigma XY - \Sigma X\Sigma Y}{N\Sigma Y^2 - (\Sigma Y)^2}$$

Y on X

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

$$b_{yx} = \frac{N\Sigma XY - \Sigma X\Sigma Y}{N\Sigma X^2 - (\Sigma X)^2}$$

Deviations from Actual Means

X on Y

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

Y on X

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2}$$

if r (co-efficient of correlation is given)

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} \quad , \quad b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

if Cov is given

$$b_{xy} = \frac{Cov(X,Y)}{(\sigma_y)^2}$$

$$b_{yx} = \frac{Cov(X,Y)}{(\sigma_x)^2}$$

If regression eqⁿ of Y on X is given then number with X is b_{yx} and vice-versa

Intersection point of 2 regression eqⁿ gives us mean.

Co-efficient of Determination (R^2)

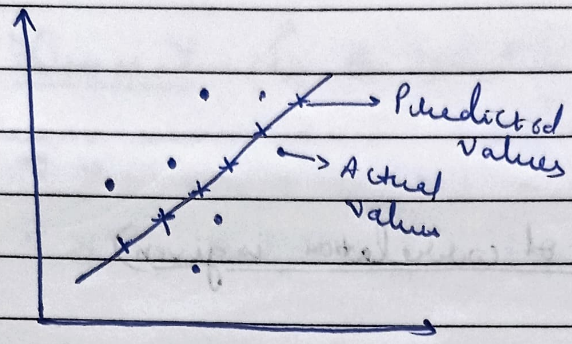
⇒ It is square of coefficient of correlation

→ always b/w 0 and 1

⇒ if closer to 0 less correlated and if closer to 1 more

⇒ if $r^2 = 1$, model is perfect fit

Mean Square Error (MSE)



$$MSE = \frac{\Sigma (Y_i - \hat{Y}_i)^2}{n}$$