CHAP 17

Correlation and Regression

Types of Data:-

① Univariate Data:-

- one variable at a time.

② Bivariate Data

- When data are collected and analysed on two variables, simultaneously, they are known as bivariate.

The corresponding frequency distribution, derived from it, is known as bivariate frequency distribution.

Marginal Distribution | Conditional Distribution

Correlation:-

- In a bivariate data, if change in one variable causes change in another variable either directly or inversely, then the two variable are known to be associated or correlated. P ↑ → Demand ↓

Types of Correlation

- Positive
  - Perfect positive → $r = 1$
  - Positive → $0 < r < 1$
- Zero $r = 0$
- Negative
  - perfect Negative → $r = -1$
  - Negative → $-1 < r < 0$

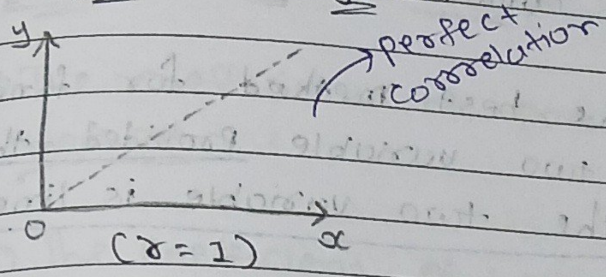
- Correlation = r
- The value of correlation range from -1 to 1,

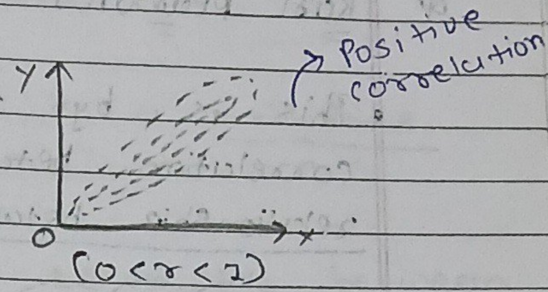
$$-1 \leq r \leq 1$$

~~Mes~~Measures of Correlation

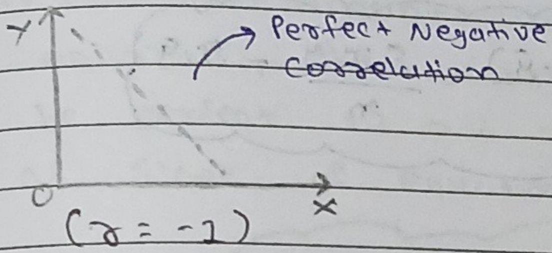
- Scatter Diagram
  - linear
  - Non. ~~linear~~ / Curve linear
- Karl pearson's Product Moment correlation coefficient
- Spearman's Rank correlation coefficient
- Coefficient ~~of~~ Concurrent Deviations.

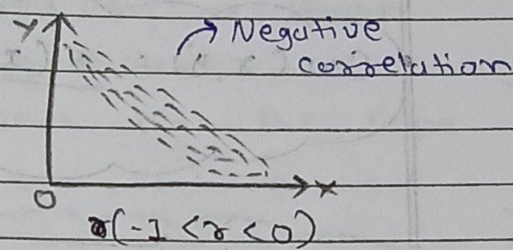
→ Scatter Diagram :-



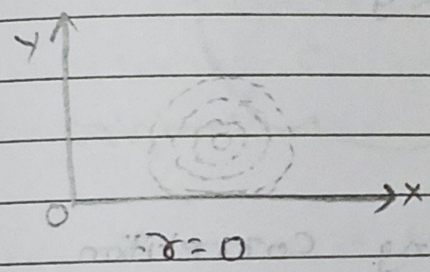


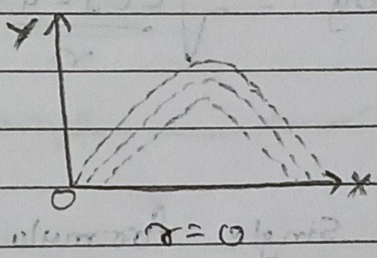
- The plotted points lie from lower left corner to upper right corner.





- The plotted points concentrate from upper left to lower right.





- The plotted points could be equally distributed without depicting any particular pattern.

Karl pearson's Product Moment Correlation Coefficient:-

- This is by far the best method for finding correlation between two variable provided the relationship between the two variable is linear.

- $$r = \frac{Cov(x, y)}{S_x \times S_y}$$

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

$$\rightarrow S_x = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n}} = \sqrt{\frac{\Sigma x^2}{n} - \bar{x}^2}$$

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

- A single formula for computing correlation coefficient is given by

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

Page

Properties of Correlation Coefficient:

i) The coefficient of correlation is a unit-free measure.

ii) The Coefficient of correlation always lies between -1 and 1, including both the limiting values.

$$-1 \le r \le 1$$

iii) If two variables are related by a linear equation, then correlation coefficient will always be perfect +1 or -1 depends on the sign of slope of equation.

iv) Change of origin = No Impact

v) Change of scale = No impact of value but affected by sign.

Original		Change
x	$\longrightarrow$	u
y	$\longrightarrow$	v
r_{xy}	$\longrightarrow$	r_{uv}

- If sign of both change of scale are same.
  $\therefore r_{uv} = r_{xy}$

- If sign of both change of scale are different
  $\therefore r_{uv} = -r_{xy}$

Spearman's Rank Correlation coefficient :-

- When we need finding correlation between two qualitative characteristics, say, beauty and intelligence, we take ~~resort res~~ recourse to using rank correlation coefficient.

- Rank correlation can also be applied to find the level of agreement (or disagreement) between two ~~variable~~ judges so far as assessing a qualitative characteristics is concerned.

- S R CC is given by

$$r_R = 1 - \frac{6 \Sigma d^2}{n(n^2-1)}$$

- Where r_R denotes Rank correlation coefficient and it lies between -1 and 1 inclusive of these two values.

- $d_i = x_i - y_i$ represents the difference in ranks for the i-th individual and n denotes the number of individuals.

Coefficient of Concurrent Deviation:-

- A very simple & casual method of finding correlation when we are not serious about the ~~magna~~ magnitude of the two variables.

$$r_c = \pm\sqrt{\pm\frac{(2c-m)}{m}}$$

- Where c is the num. of concurrent deviation (same direction) m is num. of pairs compared,

$m = n-1$

Regression Analysis:-

- In regression analysis, we are concerned with the estimation of one variable for a given value of another variable on the basis of an average mathematical relationship between the two variable.

Estimation of Y when X is given $\longrightarrow$ Y on X

Y Dependent

X Independent

$y = a + bx$

Estimation of x when y is given $\longrightarrow$ X on Y

X Dependent

Y Independent

$x = a + by$

- Estimation of Y when x is given :-
  Regression line of Y on X
  $$\boxed{Y - \bar{Y} = b_{yx}(x - \bar{x})}$$

- Estimation of x when Y is given :-
  Regression line of X on y
  $$\boxed{x - \bar{x} = b_{xy}(y - \bar{y})}$$

Method of Least Squares

- Regression coefficient :→

i) $b_{yx} = \dfrac{cov(x,y)}{var\ of\ x}$

ii) $b_{yx} = r \cdot \dfrac{SD_y}{SD_x}$

↓ y on x

i) $b_{xy} = \dfrac{cov(x,y)}{var\ of\ y}$

ii) $b_{xy} = r \cdot \dfrac{SD_x}{SD_y}$

↓ x on Y

- Properties of Regression line / coefficients :-

  - The two lines of regression intersect at the point $(\bar{x}, \bar{y})$ mean where x & y are the variable under consideration.

  - According to this property, the point of intersection of the regression line of y on x and the regression line x on y is $(\bar{x}, \bar{y})$ i.e. the solution of the simultaneous equations in x and Y.

iii) $r = \sqrt{b_{yx} \times b_{xy}}$

- Product of the regression coefficient must be numerically less than unity.

- This can be applied, unlike correlation for any type of relationship linear as well as curvilinear.

- The two lines of regression coincide i.e. become identical when $r = -1$ or 1 or in other words, there is a perfect negative or positive correlation between the two variable under discussion.

- If $r = 0$ Regression lines are perpendicular to each other.

- Coefficient of Determination / Explained variance / Accounted Variance:-

- Correlation Coefficient measuring a linear relationship between the two variables indicates the amt. of variation of one variable accounted for by the other variable. r^2

- Coefficient of non-determination / unexplained variance / unaccounted variance:-

- The coefficient of non-determination is given by $(1-r^2)$ and can be interpreted as the ratio of unexplained variance to the total variance.

The end