Mon - 5 PM
Sat = 5:30 PM

Correlation

Correlation analysis is a statistical procedure by which we determine the degree of association or relationship between 2 or more variables. But correlation doesnot predicts anything about the cause & effect relationship.

Acc to Croxton & Cowden, "when the relationship is of quantitative nature & measuring the relationship & expressing it in brief formula is called co-relation.

The co-relation coefficient is denoted by 'r'. Karl Pearson developed the formula for co-relation co-efficient.

Uses of co-relation

$-1 < r$ or $R < +1$

-0.98 0.12 0.98

1) The correlation analysis helps us to measure the degree of relationship that exists between the variables
2) In business, forecasting is an imp phenomenon & correlation helps to make relatively more dependable forecaste.
3) Correlation is used to determind the regression co-efficient if S.D of 2 variables is known.
4) Co-relation is used in fact of reliability & validity of tests.

Properties of correlation coefficient:

1) It is rigidly defined
2) It is based on all the observations.
3) The correlation coefficient is a pure no. & has no unit of measurement.
4) It is independent of change of origin & scale.
5) It lies between $-1 < r < +1$
6) The sign of co-relation co-efficient depends on sign of covariance.
7) If the 2 variables are independent the co-relation coefficient bet'w them is '0'

Interpretation of correlation coefficient

1) $r = +1$ indicates perfect positive correlation, There is an equal proportional change in both the variables & in the same direction.

For example:- Temparature & reading in thermometer.

2) $r = -1$ indicates perfect negative correlation. i.e, There is an equal proportional change in both the variables & but in the opposite direction.

For ex- Pressure & volume of gas.

3) $r = 0$, implies variables are not related.

Eg. Results of B.Sc students & rainfall.

4) A value of 'r' very near to '0' means very little correlation between x & y i.e x & y are not linearly related variables

5) A value of 'r' near to +1 or -1 means y is highly dependent on x or x is highly dependent on y.

Limitation of correlation coefficient

1) Correlation coefficient is a measure of extent of linear relationship b/w 2 variables. However a small value of r indicates only a poor linear relationship b/w the variables. This doesn't rule out the possibility that the variables may be related in some other way.

2) A high value of r doesn't always imply that there is an direct cause & effect relationship b/w variables. High correlation may be due to the fact that both the variables depends on a third variable.

3) Some times it may happen that 2 series of observations show a high correlation coefficient even though there may not be any logical basis for relationship betw them. Such correlation betw 2 variables is known as Spurious or non sense co-relation.

4) It is sometimes difficult to interpret the significant of test correlation coefficient & often it is mis interpreted.

Formula for co-relation coefficient

Let $x_1, x_2, x_3 \ldots x_n$ be a set of observations of the variate X & $y_1, y_2, \ldots y_n$ be the corresponding values of Y. Then

$$r = \frac{\frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\frac{1}{n}\Sigma(x-\bar{x})^2}\cdot\sqrt{\frac{1}{n}\Sigma(y-\bar{y})^2}}$$

$$r = \frac{\text{Cov}(x,y)}{\sigma_x \cdot \sigma_y}$$

$$r_{xy} = \frac{\Sigma xy - n\bar{x}\bar{y}}{\sqrt{\Sigma x^2 - n\bar{x}^2}\sqrt{\Sigma y^2 - n\bar{y}^2}}$$

↗

11 Show that r (r_{xy}) is given by

$$r_{xy} = \frac{\Sigma(x_i y_i) - n\bar{x}\bar{y}}{\sqrt{\Sigma x_i^2 - n\bar{x}^2}\sqrt{\Sigma y_i^2 - n\bar{y}^2}}$$

P↑ S↑ →

P↑ D↓ →

Types of Correlation:

1) Positive or Negative correlation: Correlation is said to be positive if the variable move in same direction. It means that when the value of 1 variable increases, the value of other variable will also increase & vice versa.

Correlation is negative when variable of 1 variable increases, the value of other variable decreases & vice versa.

Eg of +ve correlation

(i) Price & supply of the commodity

(ii) Add Advertisement expenses & Sale of product

Eg of −ve correlation

(i) Price & Demand of a commodity

(ii) Interest rate on loan & loan taken by the public.

2) Simple, Partial & Multiple correlation

Simple correlation studies relationship between 2 variables. When relation is measured between 2 or more variable it is known as multiple co-relation. Multiple correlation studies the extent in which a variable is effected by the combined influence of a group of other variables. The study of relationship between 2 variables, keeping the effect of other variable constant is known as partial correlation. X D_{TV} P_{TV} Y_t

Eg- Simple

(i) Price & Supply / Demand of a comm.

Multiple

(i) Demand of a comm with the combined effect of its price, income of the consumer, prices of substitute product.

Partial correlation

① Demand of a commod & its price keeping the effects of income of the consumer, price of substitute goods as constant.

$r = +1$ $r = -1$ $r = 0.75$

3) ~~Linear~~ Linear and Non linear Correlation:

If the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable, then the co-relation is said to be linear.

If the amount of change in one variable do not tend ~~to~~ to bear a constant ratio to the amount of change in other variable, then the co-relation is said to be non-linear.

$y = 2y$

Eg: Linear

① Amount of rainfall & reading in rain gauge

Non-linear:

① Amount of rainfall & crop yield.

4) Spurious & Non-sense Correlation:

It is sometime seen that though there is no cause & effect relationship betw 2 variables but still the value of the 'r' on numerical basis may prove to be significant. Such type of correlation is ~~sup~~ Spurious or

~~no~~ non sense correlation.

Eg Non sense correlation

① No. of snake bites & no. of fresh graduates.

Probable error & coefficient of correlation

Probable error of the coefficient of correlation helps in interpreting its value. With the help of probable error it is possible to determine the reliability of the value of coefficient of correlation. The probable error of r is given by:

$$P.E_r = 0.6745\frac{(1-r^2)}{\sqrt{N}}$$

where r = co-efficient of correlation

N = no. of pairs of observations

1) If the value of r is less than probable error the value of r is not at all significant
2) If the value of r is more than 6 times the probable error, the value of r is significant
3) By adding & substracting the value of probable error from the coefficient of correlation we get respectively the upper & lower limits within which coefficient of co-relation can be expected to lie.

Standard error of r =

$$S.E_r = \frac{1-r^2}{\sqrt{N}}$$

RANK Correlation

Spearman's Rank correlation coefficient is defined as

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

or

$$R = 1 - \frac{6 \Sigma D^2}{N^3 - N}$$

where

R = Rank coefficient of correlation
D = Difference of Rank between paired items in 2 series
N = No. Pairs of items

Features

① The sum of the differences of ranks between 2 variables shall be zero '0'. Symbolically

$\Sigma D = 0$

② Spearman's correlation coefficient is distribution free or non parametric.

③ The spearman's correlation coefficient is nothing but Karl pearson's correlation coefficient between ranks

Scatter Diagram

The Scatter diagram is the simplest method of studying relationship betw 2 variables. The simplest device for ascertaing ~~whe~~ whether scatter variables are related is to prepare a dot chart where the horizontal or x axis represents one (1) variable & vertical axis representing the other. The diagram so obtained is known as Scattered diagram or dot diagram. From the scatter diagram I can have a good idea about the relationship between variables. The following points may be borne in mind in interpretation of a scattered diagram. Regarding the correlation between 2 variables:

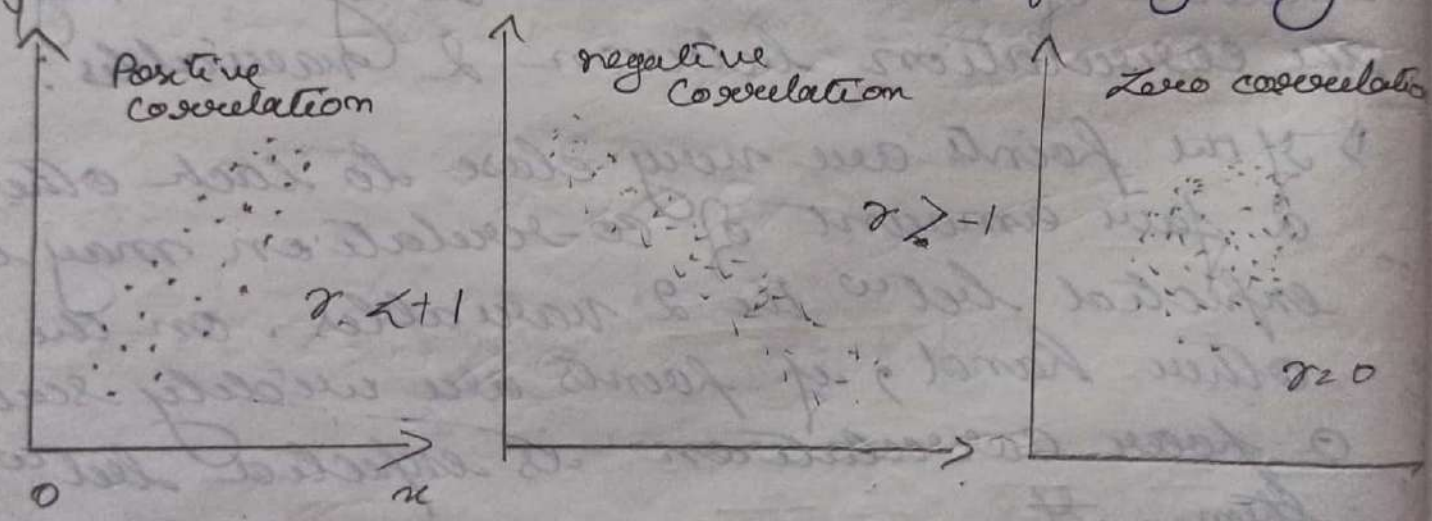
1) If the points are very close to each other, a fair amount of co-relation may be expected betw the 2 variables. On the other hand, if points are widely scattered a poor correlation is expected betw them.

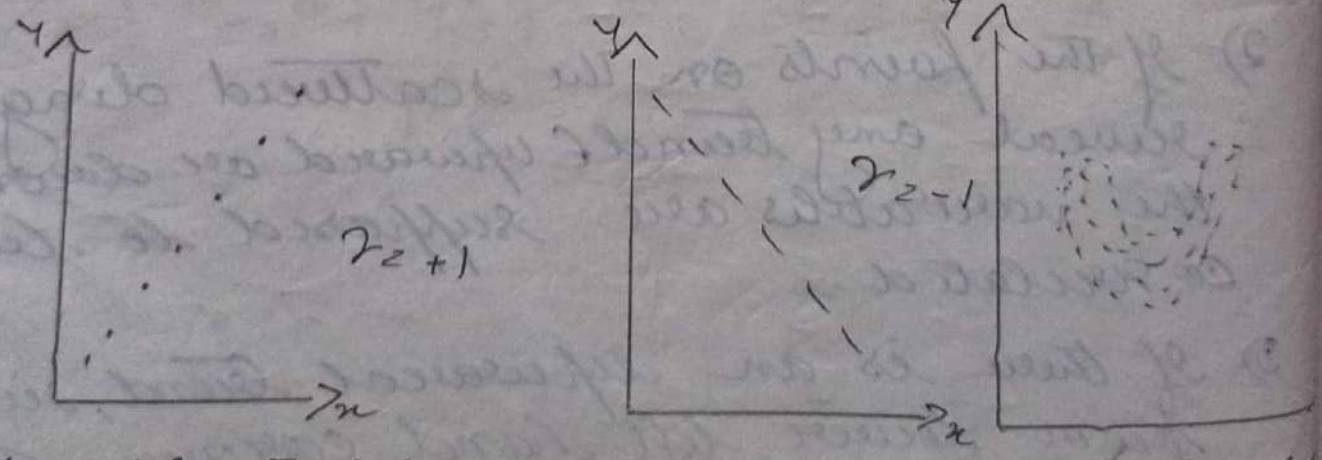
2) If the points on the scattered diagram reveal any trend (upward or downward), the variables are supposed to be co-related.

3) If there is an upward trend, rising from lower left hand corner to upper right hand corner. The co-

is said to be positive correlation. On the other hand if there is an downward trend from upper left hand corner to lower right hand corner, the correlation is said to be negative.

iii) If all the points lie on a straight line starting from left bottom & going upwards to the right top, then the correlation is perfectly positive. On the other hand if all the points lie on a straight line starting from the left top & going downwards to the right bottom, is then the correlation is perfectly neg.





Perfectly positive 'r' perfectly neg 'r' zero correla

Limitation of correlation coefficient:

Regression (unit-IV)

Regression analysis reveals average relation-ship betⁿ 2 variables, & thus makes possible estimation or prediction. Regression analysis is used for estimating or predicting the unknown value of a variable for a ~~given~~ ~~value~~ known value of another variable. More precisely if x & y are 2 related variables then regression analysis helps us to estimates the value of y for given value of x & vice versa. A

According to prof. Blair, "Regression is a mathematical measure of the average relationship between 2 or more variables in terms of original units of the data"

Defendent & Independent variable

The regressional analysis is used for estimating the unknown value of 1 variable from the known value of the other variable. The variable whose value is to be predicted is called as the dependent variable or explained variable. On the other hand, the variable with the help of which prediction is done is called

the independent variable or explanatory variable

Types of Regression

If Regression analysis is confined to the study of 2 variables only then it is known as simple regression.

On the other hand the regression analysis that studies more than 2 variables at the same time is called multiple regression.

In case of simple regression if the relation bet^n the dependent & independent variable follows a straight line pattern then the regression is called as linear regression. However if the relation is expressed in the form of a curve, then that type of regression is called ~~curvilin~~ curvilinear regression.

Simple linear regression / Regression lines

In case of 2 variables, X & Y, we shall have 2 regression lines, i.e regression on X on Y & regression of Y on X. The regression line of X on Y gives the most probable value of X for given value of Y. Similarly the regression line of Y on X gives the most probable value of Y for given value of X. However, when there is a

perfect correlation between x & y ($r = \pm 1$) then the regression lines will co-incides, i.e we have only one line. Otherwise, the regression lines intersect each other at point $(\bar{x}, \bar{y})$.

The regression eqn of x on y is given by

~~x_{xy}~~

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

Here $\bar{X}$ & $\bar{Y}$ are the A.M. of x & y respectively

b_{xy} = is the regression coefficient of x on y.

& is given by

$$b_{xy} = r \frac{\sigma_x}{\sigma_y}$$

Similarly

The regression eqn of y on x is given by

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

where,

$$b_{yx} = r \frac{\sigma_y}{\sigma_x}$$