Regression

The term regression was introduced by the English biometrician Sir Francis Galton (1822). Sir Francis describe a phenomenon which he described in analyzing the heights of children and their parents. He found that tall parents have tall children and short parents have shorts children. The average heights of children tends to step back or to regress toward the average heights of all men.

The tendency toward the average height of all men was called regression by Galton.

OR.

The interdependency between the dependent variable and one or more independent variable is called regression

Regression provides an equation to be used for estimating or predicting the average value of the dependent variable from the known values of the independent variable.

The dependent variable is assumed to be continuous/random variable whereas the independent variables are assumed to be fix values / not random.

- The study of the dependency of a variable on a single independent variable is called simple regression/two variable regression.
- The dependency of a variable on more than one/two independent variables is called multiple regression.
- When the dependency is represented by a straight line equation, the regression is said to be linear, otherwise is called curvilinear.
- The dependent variable is also called the regress and the predicate and the response or the explained variable.
- The independent variable is also9 called the non-random variable, the regress or the predictor, the regression variable or the explanatory variable.

Simple Linear Regression Model:

We assume that the linear relationship between dependent variable y_i and the independent variable X_i is,

$$y_i = \alpha + \beta X_i + \varepsilon_i \quad \text{(regression model)}$$

Where

y_i = dependent variable.

X_i = independent variable.

A & β = parameters.

ε_i = Residuals / error term.

Further more:

i. $E(\varepsilon_i) = 0$
ii. $\text{Var}(\varepsilon_i) = E(\varepsilon_i^2) = \boldsymbol{\Omega}^2$, for all i.
iii. $E(\varepsilon_i , \varepsilon_j) = 0$, for all $i \neq j$
iv. $E(X , \varepsilon_j) = 0$, X and ε are also independent of each other.
v. ε_i is normally distributed with a mean of zero and a constant variance $\boldsymbol{\Omega}^2$

<u>Least Square Estimates in Simple Linear Regression:</u>

$y_i = \alpha + \beta X_i + \varepsilon_i$ (for population Data)

$y_i = \alpha + \beta X_i + e_i$ (for simple Data)

"a" and "b" are the least squares estimates of α & β

- $E(a) = \alpha$
  $E(b) = \beta$

$a = \bar{y} - b\bar{x}$, $b = \dfrac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$

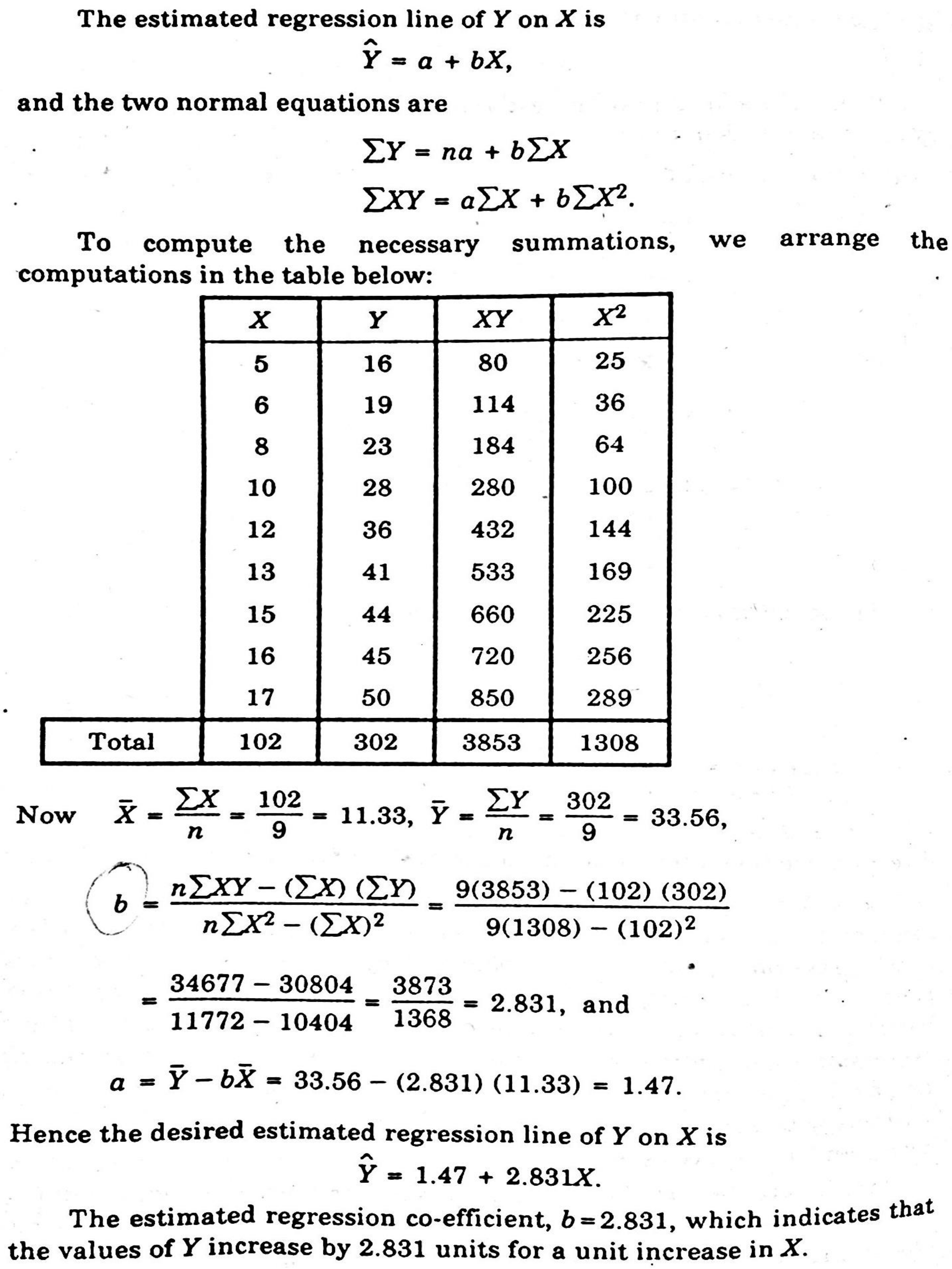
$\bar{y} = \dfrac{\Sigma y}{n}$

$\bar{x} = \dfrac{\Sigma x}{n}$

Example:

Compute the least squares regression equation of Y on X for the following data. What is the regression co-efficient and what does it mean?

X	5	6	8	10	12	13	15	16	17
Y	16	19	23	28	36	41	44	45	50



The estimated regression line of Y on X is

$$\hat{Y} = a + bX,$$

and the two normal equations are

$$\sum Y = na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2.$$

To compute the necessary summations, we arrange the computations in the table below:

	X	Y	XY	X^2
	5	16	80	25
	6	19	114	36
	8	23	184	64
	10	28	280	100
	12	36	432	144
	13	41	533	169
	15	44	660	225
	16	45	720	256
	17	50	850	289
Total	102	302	3853	1308

Now $\bar{X} = \dfrac{\sum X}{n} = \dfrac{102}{9} = 11.33,\ \bar{Y} = \dfrac{\sum Y}{n} = \dfrac{302}{9} = 33.56,$

$$b = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2} = \frac{9(3853) - (102)(302)}{9(1308) - (102)^2}$$

$$= \frac{34677 - 30804}{11772 - 10404} = \frac{3873}{1368} = 2.831, \text{ and}$$

$$a = \bar{Y} - b\bar{X} = 33.56 - (2.831)(11.33) = 1.47.$$

Hence the desired estimated regression line of Y on X is

$$\hat{Y} = 1.47 + 2.831X.$$

The estimated regression co-efficient, $b = 2.831$, which indicates that the values of Y increase by 2.831 units for a unit increase in X.

Example 10.2 In an experiment to measure the stiffness of a spring, the length of the spring under different loads was measured as follows:

X=Loads (1b)	3	5	6	9	10	12	15	20	22	28
Y=length (in)	10	12	15	18	20	22	27	30	32	34

Find the regression equations appropriate for predicting

i) the length, given the weight on the spring;

ii) the weight, given the length of the spring. (W.P.C.S., 1964)

The data come from a bivariate population, *i.e.* both X and Y are random, therefore there are two regression lines. To find the regression equation for predicting length (Y), we take Y as dependent variable and treat X as independent variable (*i.e.* non-random). For the second regression, the choice of the variables is reversed.

The computations needed for the regression lines are given in the following table:

	X	Y	X^2	Y^2	XY
	3	10	9	100	30
	5	12	25	.144	60
	6	15	36	225	90
	9	18	81	324	162
	10	20	100	400	200
	12	22	144	484	264
	15	27	225	729	405
	20	30	400	900	600
	22	32	484	1024	704
	28	34	784	1156	932
Total	130	220	2288	5486	3467

i) The estimated regression equation appropriate for predicting the length, Y, given the weight X, is

$$\hat{Y} = a_0 + b_{yx} X,$$

where $$b_{YX} = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2} = \frac{(10)(3467) - (130)(220)}{(10)(2288) - (130)^2}$$

$$= \frac{6070}{5980} = 1.02, \text{ and}$$

$$a_0 = \bar{Y} - b_{yx}\bar{X} = 22 - (1.02)(13) = 8.74$$

Hence the desired estimated regression equation is

$$\hat{Y} = 8.74 + 1.02\, X$$

ii) The estimated regression equation appropriate for predicting the weight, X, given the length is

$$\hat{X} = a_1 + b_{xy} Y,$$

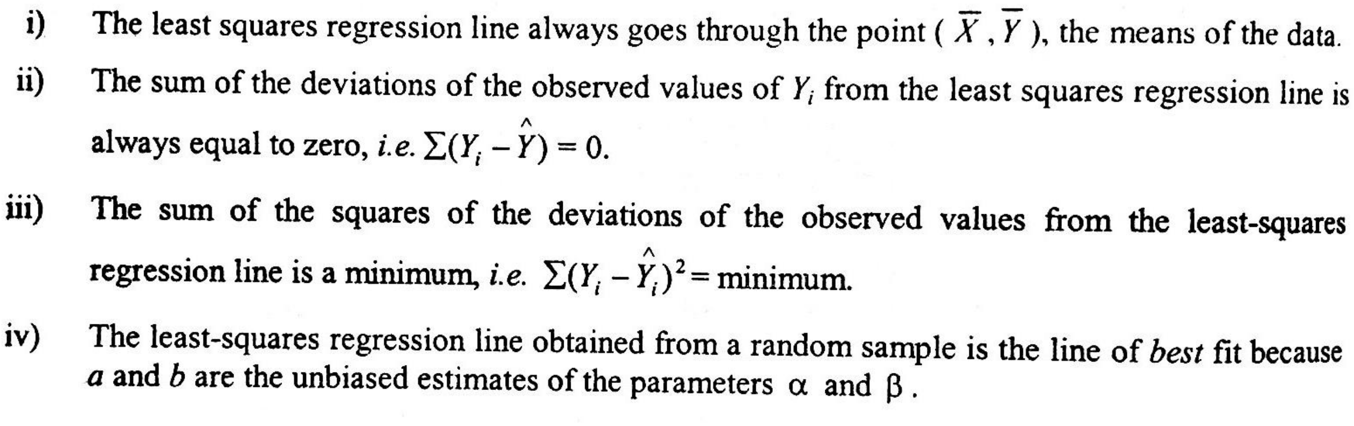
where $$b_{XY} = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma Y^2 - (\Sigma Y)^2} = \frac{(10)(3467) - (130)(220)}{(10)(5486) - (220)^2}$$

$$= \frac{6070}{6460} = 0.94, \text{ and}$$

$$a_1 = \bar{X} - b_{xy}\bar{Y} = 13 - (0.94)(22) = -7.68$$

Hence $\hat{X} = 0.94Y - 7.68$ is the estimated regression equation appropriate for predicting the weight (X), given the length (Y).

Properties Of The Least – Square Regression Line:



i) The least squares regression line always goes through the point ($\overline{X}$, $\overline{Y}$), the means of the data.

ii) The sum of the deviations of the observed values of Y_i from the least squares regression line is always equal to zero, *i.e.* $\Sigma(Y_i - \hat{Y}) = 0$.

iii) The sum of the squares of the deviations of the observed values from the least-squares regression line is a minimum, *i.e.* $\Sigma(Y_i - \hat{Y}_i)^2 =$ minimum.

iv) The least-squares regression line obtained from a random sample is the line of *best* fit because *a* and *b* are the unbiased estimates of the parameters α and β .

CORRELATION

Correlation, like covariance, is a measure of the degree to which any tow variables vary together. In other words, two variables are said to be correlated if they tend to simultaneously vary in same direction. If both the variables tend to increase (or decrease) together, the correlation is aid to be direct or positive, e.g the length of an iron bar will increase as the temperature increase. If one variable tends to increase as the other variable decreases, the correlation is said to be negative or inverse, e.g the volume of gas will decreases as pressure increases. It is worth remarking that in correlation, we assess the strength of the relationship (or interdependence) between two variables; both the variables are random variables, and they are treated symmetrically, i.e there is no distinction between dependent and independent variable. In regression, by contrast, we are interested in determining the dependence of one variable that is random, upon the other variable that is non- random or fixed, and in predicting the average value of the dependent variable by using the known values of other variable.

By correlation analysis we measure the strength of relationship between two variables by finding a single number called a correlation coefficient. The primary objective is to measure the strength or degree of linear association between two variables.

Pearson Product Moment Correlation Co-efficient:

A numerical measure of strength in the linear relationship between any two variables is called the Pearson product moment correlating co-efficient or sometimes, the coefficient of simple correlation or total correlation. The simple linear correlation coefficient for n pairs of observation (Xi, Yi) usually denoted by letter “r” , is defined by

$$r = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{\sqrt{\Sigma(X - \overline{X})^2 \, \Sigma(Y - \overline{Y})^2}}$$

The population correlation co-efficient for a bivariate distribution, denoted by p, has already been defined as,

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var(X)Var(Y)}}}$$

For computational purpose, we have an alternative form of r as;

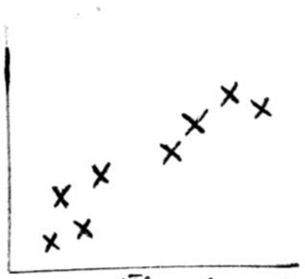
$$r = \frac{\Sigma XY - (\Sigma X)\,(\Sigma Y)/n}{\sqrt{[\Sigma X^2 - (\Sigma X)^2/n][(\Sigma Y^2 - (\Sigma Y)^2/n)]}}$$

$$= \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[n\Sigma X^2 - (\Sigma X)^2][n\Sigma Y^2 - (\Sigma Y)^2]}}$$

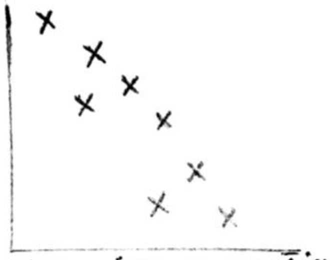
Correlation is concerned with whether or not there is any association between two variables. If two variables are related to any extent, then changes in the value of one are associated with changes in the value of the other e.g an increase in the sales of a company may show a strong association with increase in the money spent on the advertising of its products.

Scatter Diagrams:

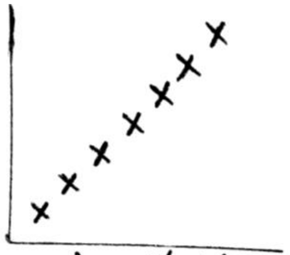
Scatter diagrams (or scatter graphs) provides a useful means of deciding whether or not there is association between variables. The construction of a scatter diagram is by drawing a graph so that the scale for one variables (the independent variable if this can be determined) lies along the horizontal axis, and the other variable (the dependent variable) on the vertical axis . each pair of figure is then plotted as a single point on the graph.



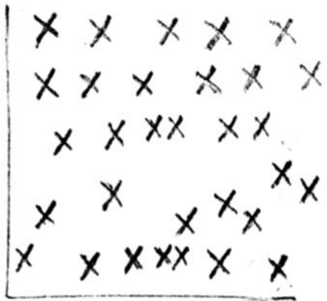
a: positive correlation



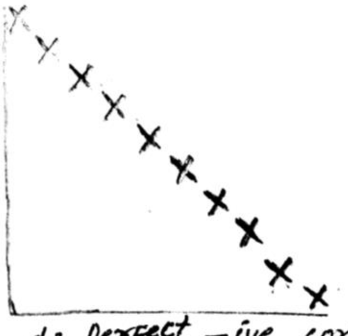
b: -ive correlation



c: perfect +ive correlation



e: NO correlation



d: perfect -ive correlation

Figure (a) : indicates positive correlation so that as the variable “X” increases so “y” will increase

Figure (b) : indicates negative correlation so that as the variable “X” increases so “y” will decrease.

Figure (c) : indicates perfect positive correlation between the two variables so that they both increases in same proportion.

Figure (d) : indicates perfect negative correlation between the two variables so that they both decreases in same proportion.

Figure (e): indicated that there is no correlation between the two variables. There are a number of lines of best fit which could be drawn with equal validity.

What is meant by and what is the function of a scatter diagram?

A Scatter diagram is a figure in which each pair of independent – dependent observation is plotted as a point in the “xy” plane.

Difference between correlation and association:

Association gives the relationship between attributes, while the correlation gives the relationship between variables.

Regress VS Correlation:

Regression and correlation have same fundamental difference that is worth mentioning. In regression analysis there is an asymmetry in the way the dependent and explanatory variables are treated. The dependent variables is assumed to be statistical, random or stochastic that is to have a probability distribution. The explanatory variables, on the other hand, are assumed to have fixed values. In correlation analysis, on the other hand, we treat any (two) variables symmetrically, i.e there is no distinction between the dependent explanatory variables.

Properties of Correlation Co-efficient:

i. The value of “r” does not depend on the unit of measurement for either variable.

ii. The value of “r” is symmetrical with respect to “x” and “y” i.e

$$r_{xy} = r_{yx}$$

iii. The value of “r” is between -1 and +1 i.e

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

r = -1 indicates a perfect negative relationship and

r = +1 indicates a substantial positive relationship

r = 0 indicated a no relationship between variables.

iv. “ r “ is the geometric mean of two regression coefficient

$$r = \pm \sqrt{b_{xy} \times b_{yx}}$$

v. “r” is independent of change of origin and scale i.e if we make transformation.

$$u = \frac{x-a}{h} \quad \text{and} \quad v = \frac{y-b}{k}$$

$$\text{then} \quad r_{uv} = r_{xy}$$

<u>Simple Correlation:</u>

It is defined as the degree of relationship existing between two variables. For example, the relationship between smoking and lungs cancer, between scores on statistics and mathematics examinations and so on and its limit are

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

<u>Curvilinear Correlation:</u>

Correlation may be curvilinear, when all points (x,y) on scatter diagram are seem to lie near a curve.

<u>Linear Correlation:</u>

Correlation may be linear, when all points (x,y) on a scatter diagram seem to cluster near a straight line.

<u>Multiple Correlation:</u>

It is independence between a variable and a group of other variables its coefficient is denoted by R_{xyz} etc and its limits are.

$$-1 \leq R_{xyz} \leq +1$$

Or

The association or interdependence between a variable and a group of other variables is called multiple correlation.

Partial Correlation:

It is the interdependence between the two variables ignoring the effect of other variables. Its coefficient is denoted by r_{xyz} etc and its limits are.

$$-1 \leq r_{xyz} \leq +1$$

Or

Partial correlation measures the degree of association between two variables ignoring the effect of a set of other controlling variables.

Positive Correlation:

When the moment of the variable is in the same direction, it is called positive correlation e.g when price increases the quantity supplied increases and when price decreases the quantity supplied also decreases.

OR

When both variables moves in the same direction then correlation is positive.

Negative Correlation:

When the moments of the variables are in the inverse direction, it is called negative correlation e.g when price increases demand for commodity decreases, when price falls demand increases.

Or

When the variables move in opposite direction then correlation is negative.

No correlation or Zero Correlation:

When two variables are independent then there is correlation e.g there is no correlation between weights of students and the colour of their hair.

Co-efficient of Correlation:

It is the numerical value which shows the degree of interdependence or association between two or more than two variables. The simple correlation is defined as

$$r = \frac{Cov(x,y)}{\sqrt{Var(x).Var(y)}} = \frac{Cov(x,y)}{\sigma x . \sigma y}$$

"r" thus defined is measure of linear association between two variables and its limits are $-1 \leq r_{xyz} \leq +1$, -1 and +1 indicating perfect negative and positive association respectively.

Co-efficient of Association:

The strength of associating between two attributes A and B is measured by Yule’s Coefficient of association, which is denoted by Q and defined as:

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

It’s value always lies between -1 and +1.

When the attributes are independent then Q = 0.

When the attributes are completely associated then Q = +1.

When the attributes are completely disassociated then Q= -1.

Association:

If the two attributes A and B are not independent then they are said to be associated i.e

$$(AB) \neq \frac{(A)(B)}{n}$$

Then A and B are associated.

A and B are said to be positively associated if

$$(AB) > \frac{(A)(B)}{n}$$

On the other hand , A and B are said to be negatively associated if.

$$(AB) < \frac{(A)(B)}{n}$$

Attributes:

A Characteristics which varies only in quality from individual to individual and cannot be measure in quantity, is called on attribute. The example of attributes are:

Marital Status of man, the color of car, education level, richness etc.

The attributes cannot be numerically expressed but only their presence or absence can be described.

Co-efficient of determination:

The co-efficient of determination, R^2 is defined as the proportion of the total variation in y explained by the regression of y on x. the total variation in y or total sum of squares.

$$TSS = \sum (Y_i - \bar{Y})^2 = \sum y_i^2$$

The explained variation in y or error sum of squares.

$$ESS = \sum (Y_i - \hat{Y}_i)^2 = \sum e_i^2$$

Then

$$R^2 = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = 1 - \frac{\sum e_i^2}{\sum y_i^2}$$

R^2 is the most commonly used measure of the goodness of fit of a regression line. It a nonnative quantity and its limits are.

$$-1 \leq R^2 \leq +1$$

An R^2 of 1 means a perfect fit, where as an R^2 of zero means no relationship between the dependent and explanatory variable (s).

Dichotomy:

The process of diving the objects into two mutually exclusive classes is called dichotomy. E.g division of a population according to sex into two classes as males and females.

Positive and Negative Classes:

The capital Latin Letters A,B,C… are usually used to denote the attributes. The letters A,B,C,…. Are designated to the individuals possessing the attributes A,B,C … while the Greek letters α, β,γ, are designated to the individuals do not possessing the attributes A,B,C,…. The attributes denoted by A,B,C…. are called positive attributes while the attributes denoted by α, β,γ .. are called negative attributes.

Consistence:

The class frequencies observed in the same population are said to be consistent, if they conform with one other. For a consistent data, any class frequency can never be negative. If any class frequency is negative then the data are in consistent.

Independence:

The independence mean that there is no relationship between attributes A and B. two attributes A and B are said to be independent if.

$$(AB) = \frac{(A).(B)}{n}$$

Contingency Table:

A Contingency table is a tabular representation of categorical data. A contingency table usually shows frequencies. For particular combination of r categories, A_1, A_2,….., A_r of attributes A and C categories, A_1, A_2,….., B_c of attribute B. the no. of individuals belonging to A_i and B_j is represented by O_{ij} where

$$\sum_i \sum_j O_{ij} = n$$

Rank Correlation:

The correlation between two set of ranking for the variables X and Y is called rank correlation. Sometimes, the accurate assessment is not possible then the objects are arranged in order according to some characteristics of intereset. The order given to an individual is called the rank.

Suppose that we have n pairs of observations from a bivariate population as (a_1,b_1), (a_2,b_2) ,……., (a_n,b_n). the values of the set a_i be ranked as X_1, X_2,….., X_n and the values of set b_i, be ranked as Y_1, Y_2,….., Y_n .

We assume that there are no some ranks given to two or more objects. Then the coefficient of rank correlation is

$$\gamma_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Where d= x – y

r_s always lies between -1 and +1.

Or

Rank Correlation:

Sometimes it is possible to arrange various items of a series in serial order with respect to some characteristics, if the numerical measurement of this value is difficult e.g a teacher can arrange the students in his class in ascending or descending order of intelligence or a sales manager can arrange a group of salesman in ascending or descending order by efficiency where as quantities measurement of these characteristic (i.e intelligence and efficiency) is not possible directly. In such cases the product movement method inappropriate because the data are not form a measuring device, then we sue Auxiliary correlation procedure to hand such problems.

To Calculate the coefficient of rank correlation:

a. Rank each respective variable in order.
b. Put the corresponding rank of one variable against the rank of the other variable.
c. Shows the difference between the rankings in each case.
d. Square the difference and find the sum of the squares.
e. Apply the formula:

$$\gamma = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)}$$

Permutations:

It is an arrangement of all or part of a set of objects is called permutation.

Or

A group of items with a certain order (arrangements) is called permutation.

For example,

A , B, C can be written as.

ABC, ACB, BAC, BCA, CAB, CBA. These are the different permutation.

Fundamental Principles:

- **a.** If one operation can be performed in “m” ways and other performed in “n” ways, then 0. If performing the two operation will be m x n.
- **b.** Permutation of n objects = n!.
- **c.** Permutation of n distinct objects arranged in circle = (n – 1)!
- **d.** Permutation of n objects of which n_1 are of one kind, n_2 are of other kind.

$$= \left(\overset{n}{n_1, n_2, \ldots, n_r}\right) = \frac{\underline{|n}}{\underline{|n_1}, \underline{|n_2}, \ldots, \underline{|n_r}}$$

When

$$n_1 + \quad + \cdots + n_r = n$$

Permutation of “n” dissimilar things taken “r” at a time.

$$^{n}P_{r} = \frac{n!}{(n-r)!}$$

Example:

How many numbers of two different digits can performed with figures 1,2,3,4,5,6.

Solution:

For first digit there are 6 cases because any no. can be chosen. For 2nd there are only 5 way then the total way = 6x5 = 30.

2nd method:

$$^{6}P_{2} = \frac{6!}{(6-2)!} = 30$$

Example: There are 10 baby taxis running between Jhelum and Mangla. In how many ways a man can go from Jhelum to Mangla and return by different baby Taxis.

Solution:

There are 10 ways of making the first passages, then are 9 chosen to return.

Total ways = 9 x 10 = 90.

2nd method.

$$^{10}P_2 = \frac{10!}{(10-2)!} = 90$$

Combination:

It is group or selection made by taking all or part of a set is called combination e.g

a. Combination of made by letters a,b,c and d.
b. Number of combination of n dissimilar things taken r at a time.

$$^{n}C_r = \frac{n!}{r!(n-r)!}$$