STATISTICS

Classification of Data

① Grouped Data – which can be organised.

eg. 0-10 – 3
10-20 – 4
20-30 – 5

② Continuous data – upper limit of 1st interval is equal to the lower limit of second ~~data~~ interval.

③ Discontinuous data – Not equal.

eg. 0-9 – 3
10-19 – 2
20-29 – 1

② Ungrouped data : Cannot be organised into classes or it is just a list of nos.

eg. 3, 5, 11, 15, ...

Frequency Distribution

① Discrete : if data is represented in such a way that exact measurements of the units are clearly shown.

(ii) Continuous :- in which data are arranged in classes groups which are not exactly measurable.

Cumulative frequency Distribution :-

Sum of all the frequencies i.e, first class + second class + third class and so, on is so obtained is known as cf (cumulative frequency).

* two types :- (i) less than
(ii) greater than.

(i) less than :- For less than cumulative frequency, we add up the frequency from above.

(ii) greater than :- For greater than, cumulative frequency, we add up frequencies from below.

wages	cf (no. of workers)
less than 1099.5	125
" " 1199.5	275
" " 1299.5	475
" " 1399.5	725
" " 1499.5	925
" " 1599.5	1000

wages	cf (no. of workers)
greater than 999.5	1000
" " 1099.5	875
" " 1199.5	725
" " 1299.5	525
" " 1399.5	275
" " 1499.5	100

Relative frequency Distribution :-

very useful for the comparison of two or more frequency distribution.

$$\text{Relative frequency \%} = \left[\frac{\text{class frequency}}{\text{Total frequency}} \times 100\right]$$

#

(i) Pie diagrams :-

$$\text{Central angle} = \left[\frac{\text{frequency} \times 360^\circ}{\text{Total frequency}}\right]$$

(ii) Histogram :-

Set of adjacent rectangles whose area is proportional to the frequencies of a given continuous frequency distribution.

no gap b/w any two successive rectangles.

(iii) frequency polygon :- To draw the frequency polygon of an ungrouped frequency distribution, we plot the points with abscissae as the variate values and the ordinate as the corresponding frequencies.

These plotted points are joined by st. lines to obtain the frequency polygon.

(10) Ogive :- (Of curve) :- When we plot the upper class limits along X-axis and cumulative frequencies along Y-axis, And on joining them, we get a curve called an Ogive.

* two types :-

(i) less than ogive ; the rising curve.

(ii) More than ogive ; falling curve.

(i) less than type :-

the upper limits along the X-axis. and the corresponding cummulative frequencies along y-axis.

To get a rising curve.

(ii) More than type :-

Subtract the frequency of each class Now, mark the lower class limits along X-axis and the corresponding Y-axis (Of).

To get a declining curve.

Measures of Central Tendency.

following are the 5 measures of central tendency :-

(1) Mathematical Averages :-

(a) Arithmetic Mean or Mean
(b) Geometric Mean
(c) Harmonic Mean.

} Also known as Measures of location.

(2) positional Averages :-

(a) Median
(b) Mode.

ARITHMETIC MEAN (AM) :-

Sum of all the numbers in the series is divided by the total no. of series.

ARITHMETIC MEAN OF UNGROUPED OR INDIVIDUAL OBSERVATIONS :-

If $x_1, x_2, x_3, \ldots, x_4, x_n$ are n observations, of a variable x, then the AM is

(i) Direct Method :-

$$\bar{X} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} \quad \text{or} \quad \bar{X} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)$$

(ii) Shortcut Method :- (Assumed Mean Method)

$$\bar{X} = A + \frac{1}{n}\sum_{i=1}^{n} d_i$$

where,

A = assumed mean

d_i = $x_i - A$

WEIGHTED ARITHMETIC MEAN :-

If $w_1, w_2, w_3, \ldots, w_n$ are the weights assigned to the n values of $x_1, x_2, \ldots, x_n$, respectively, then the weighted average of AM is given by

$$\bar{X} = \frac{w_1x_1 + w_2x_2 + \cdots + w_nx_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

OR

$$\bar{X} = \frac{\Sigma wx}{\Sigma w}$$

ARITHMETIC MEAN OF A DISCRETE FREQUENCY DISTRIBUTION :-

① **Direct Method :-** If a variable X takes values $x_1, x_2, x_3, \ldots, x_n$ which corresponding frequencies $f_1, f_2, f_3, \ldots, f_n$, respectively, then the arithmetic mean of the values is

$$\bar{X} = \frac{f_1x_1 + f_2x_2 + \ldots + f_nx_n}{f_1 + f_2 + f_3 + \ldots + f_n}$$

or $$\bar{X} = \frac{\sum_{i=1}^{n} f_i x_i}{N}$$

where, $N = f_1 + f_2 + \ldots + f_n$.

$$\therefore N = \sum_{i=1}^{n} f_i$$

(ii) Shortcut Method :- If the values of x_i or (and) f_i is large, the calculation of AM by the formula used above is quite tedious and time consuming.

In such a case, we take the deviation from assumed mean A which is in the middle or just close to it in the data.

Then,

$$\bar{X} = A + \frac{1}{N} \sum_{i=1}^{n} f_i d_i$$

where, $d_i = x_i - A$

A = Assumed Mean.

=

Note

This method is nothing but shifting of origin from zero to the Assumed Mean A on the no. line.

(iii) Step deviation Method :- Sometimes, during the application of shortcut method of finding the mean, the deviations d_i are divisible by a common number b (say).

In such a case, the

arithmetic mean is reduced to a great extent by taking.

$$u_i = \frac{x_i - A}{b}, \quad i = 1, 2, \ldots, n$$

$$\bar{X} = A + b\left(\frac{1}{N}\sum_{i=1}^{n} f_i u_i\right)$$

NOTE: This process is called change of scale on the no. line.

ARITHMETIC MEAN OF A GROUPED OR CONTINUOUS FREQUENCY DISTRIBUTION :-

We need to compute the mid-point of class intervals (x_i).

The mid-points are multiplied by the corresponding frequencies (f_i).

The sum of this product is obtained and is divided by the sum of frequencies.

The AM may be compu-ted by applying any of the methods used in a discrete frequency.

COMBINED ARITHMETIC MEAN :-

If we are given the AM of the two data sets and their sizes, then the combined AM of two data sets can be obtained by

$$\bar{X}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1+n_2}$$

where,

$\bar{x}_{12}$ = Combined mean of two data sets 1 & 2.
$\bar{x}_1$ = Mean of 1st data
$\bar{x}_2$ = Mean of 2nd data
n_1 = Size of the 1st data
n_2 = " " " 2nd data.

If $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_K$ are the means of K series of sizes $n_1, n_2, \ldots, n_K$, respectively then the mean $\bar{x}$ of the composite series is given by

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \ldots + n_K\bar{x}_K}{n_1+n_2+\ldots+n_K}$$

PROPERTIES OF AM :-

(a) If $\bar{x}$ is the mean of $x_1, x_2, \ldots x_n$, then mean of $ax_1+b, ax_2+b, \ldots ax_n+b$, is $a\bar{x}+b$.

(b) Mean is dependent of change of origin but it is independent of change of scale.

(c) Algebraic sum of the deviations of a set of values from their AM is zero.

(d) The sum of the squares of the deviations of a set of values is minimum, when taken about mean.

GEOMETRIC MEAN (GM) :-

The nth root of the product of the values is called geometric mean.

(1) # GEOMETRIC MEAN FOR UNGROUPED DATA :-

If $x_1, x_2, \ldots, x_n$ are n non-zero values of a variate X, then geometric mean is

$$GM = (x_1 \cdot x_2 \cdots x_n)^{1/n}$$

$$\log GM = \frac{1}{n}(\log x_1 + \log x_2 + \cdots + \log x_n)$$

$$\log GM = \frac{1}{n}\sum_{i=1}^{n} \log x_i$$

$$GM = \text{antilog}\left(\frac{1}{n}\sum_{i=1}^{n} \log x_i\right)$$

GEOMETRIC MEAN FOR GROUPED DATA :-

If $x_1, x_2, \ldots, x_n$ are n observations whose corresponding frequencies are $f_1, f_2, \ldots, f_n$, then geometric mean is given by

$$GM = (x_1^{f_1} \cdot x_2^{f_2} \cdots x_n^{f_n})^{1/N}$$

$$\log GM = \frac{1}{N}(f_1 \log x_1 + f_2 \log x_2 + \cdots + f_n \log x_n)$$

$$\log GM = \frac{1}{n}\sum_{i=1}^{n} f_i \log x_i$$

$$GM = \text{antilog}\left(\frac{1}{N}\sum_{i=1}^{n} f_i \log x_i\right)$$

COMBINED GEOMETRIC MEAN :-

If G_1 and G_2 are the geometric means of two series of sizes n_1 and n_2 respectively,

then the GM of the combined series is given by

$$\log GM = \frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2}$$

as median divides a distribution into two equal parts,

$$N = f_1 + f_2 + \cdots + f_n .$$

HARMONIC MEAN :-

It is the reciprocal of the AM of the reciprocals of the observations of any series.

HARMONIC MEAN FOR UNGROUPED DATA :-

If $x_1, x_2, \ldots, x_n$ are n non-zero values of a variate X, then harmonic mean is

$$HM = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \left(\frac{1}{x_i}\right)}$$

HARMONIC MEAN FOR GROUPED DATA :-

If $x_1, x_2, \ldots, x_n$ on n observations, whose corresponding frequencies of each variate is $f_1, f_2, \ldots, f_n$, then

$$HM = \frac{f_1 + f_2 + \cdots + f_n}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \cdots + \frac{f_n}{x_n}} = \frac{N}{\sum_{i=1}^{n} \left(\frac{f_i}{x_i}\right)}$$

$$= \frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n} \frac{f_i}{x_i}}$$

Where, $N = f_1 + f_2 + \cdots + f_n$

NOTE:- If AM is to be find, and if GM and HM of a statistical sequence is known, then AM can be obtained as

$$AM = \frac{(GM)^2}{(HM)}$$

or

$$(GM)^2 = (AM) \times (HM)$$

MEDIAN :-

It is the middle most or the central values of the variate in a set of observations when the observation are arranged either in ascending or in descending order of their magnitudes.

It divides the arranged series in two equal parts.

MEDIAN OF AN UNGROUPED INDIVIDUAL OBSERVATIONS :-

Let $x_1, x_2, x_3, \ldots, x_n$ be set of observations.

(1) arrange the observations in ascending or descending order.

If n is odd :-

Median (M) : Value of $\left(\frac{n+1}{2}\right)^{th}$ Observation

If n is even :-

Median (M) : $\frac{\left(\frac{n}{2}\right)^{th} \text{ observation} + \left(\frac{n}{2}+1\right)^{th} \text{ observation}}{2}$

MEDIAN OF THE DISCRETE FREQUENCY DISTRIBUTION :-

In case of a discrete frequency distribution,

x_i, f_i, $i = 1, 2, \ldots, n$,

(a) Arrange the data in ascending or descending order and then find the cumulative frequency (cf).

(b) find $\left(\frac{N}{2}\right)$, where $N = \sum_{i=1}^{n} f_i$

⑥ See the (cf) just greater than $\frac{N}{2}$.

The corresponding value of x is median.

ANOTHER:-

Find the cumulative frequency (CF)

If N is odd :-

Median (M) : Value of $\left(\frac{N+1}{2}\right)$ th observation

If N is even :-

Median (M) : ~~Value of~~

$$\frac{\left(\frac{N}{2}\right)\text{th observation} + \left(\frac{N}{2}+1\right)\text{th observation}}{2}$$

where,

$$N = \sum_{i=1}^{i} f_i$$

MEDIAN OF CONTINUOUS FREQUENCY DISTRIBUTION :-

Let the Number of observation be N.

(a) prepare the cumulative frequency (cf) column and obtain $N = \sum f_i$

and find $\frac{N}{2}$.

(b) Find the median class i.e., the class in which the observation whose cf. is equal to or just greater than $\frac{N}{2}$ lies. This class is known as the median class.

(c) Use the formula,

(M) Median =

$$M = l + \left[\frac{\frac{N}{2} - C}{f} \right] \times h$$

where, l : lower limit of the median class
f : frequency of the median class
h : width (size) of the " "
C : Cf of the class preceeding the median class.

MODE :-

It is that value in a series which occurs most frequently. It has the maximum frequency.

It is the observation with maximum frequency, whenever the other observations have for less frequencies.

MODE OF INDIVIDUAL OBSERVATIONS :-

The value which is repeated more no. of times is called mode of the series.

MODE OF A DISCRETE SERIES :-

It is the value of variable consisting highest frequency.

MODE OF A CONTINUOUS SERIES :-

① First find the modal class i.e., which has the maximum frequency.

⑤ Use the formula,

$$\text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) \times h$$

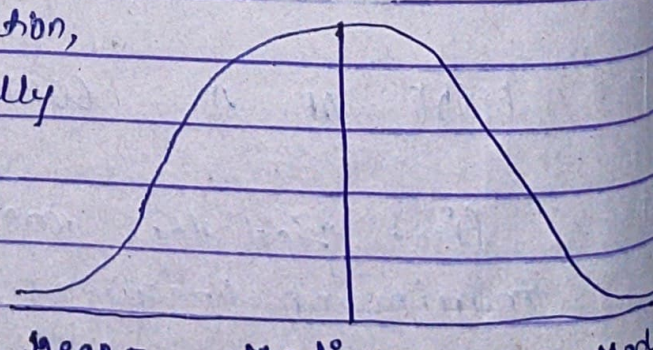
where,

l = lower limit of modal class
h = width (size) of the modal class
f_1 = frequency of " " "
f_0 = " " " class preceeding modal class
f_2 = frequency of the class succeeding modal class.

Symmetric distribution :-

A distribution is a symmetric distribution, if the values of mean, median, mode coincide.

In a symmetric distribution, frequencies are symmetrically distributed on both sides of the centre point of the frequency curve.



Relationship b/w Mean, Median and Mode :-

Mean - Mode = 3 (Mean - Median)

$\Rightarrow$ **Mode = 3 Median - 2 Mean**

Measures of Depression

Dispersion is the measure of the variations.

The degree to which numerical data tend to spread about an average value is dispersion of the data.

The measures of dispersion commonly used are :-

(a) Range
(b) Mean Deviation
(c) Standard Deviation

RANGE :-

diff. b/w max. and min. values of variate

i.e., $\boxed{\text{Range} = L - S}$ where,

L : largest value

S : smallest value

NOTE:

$$\text{Coefficient of Range} = \frac{L-S}{L+S}$$

MEAN DEVIATION :-

Mean deviation about a central value is defined as the AM of the absolute deviations of all the values taken about that central value.

M.D. OF INDIVIDUAL OBSERVATIONS :-

If $x_1, x_2, \ldots, x_n$ are n values of a variable x, then the mean deviation from an average A. is given by -

$$M.D.(A) = \frac{1}{n}\sum_{i=1}^{n} |x_i - A| = \frac{1}{n}\sum |d_i|$$

where, $d_i = x_i - A$

M.D. OF DISCRETE FREQUENCY DISTRIBUTION :-

If $x_1, x_2, \ldots, x_n$ are n observation with frequencies $f_1, f_2, f_3, \ldots, f_n$, then mean deviation from an average A is given by

$$\text{M.D. } (A) = \frac{1}{N} \sum f_i |x_i - A|$$

where,

$$N = \sum_{i=1}^{n} f_i$$

M.D. OF CONTINUOUS FREQUENCY DISTRIBUTION :-

The process is same as for a discrete frequency distribution. The only diff. is that here we have to obtain the midpoints of the various classes and take the deviations of these mid-points from the given average A.

NOTE:- Coefficient of M.D. $= \frac{\text{M.D}}{\text{AM}} = \frac{\text{M.D.}}{\bar{x}}$

NOTE :- Mean deviation of a given set of observations is least when taken about their median.

VARIANCE AND STANDARD DEVIATION :-

The variance of a variate x is the AM of the squares of all deviations of x from the AM of the observations and is denoted by

$$Var(x) \text{ or } \sigma^2.$$

The positive sq. root of the variance of a variate x is known as standard deviation.

i.e.,

$$\text{Standard Deviation (S.D.)} = \sqrt{Var(x)} = \sqrt{\sigma^2} = \sigma$$

VARIANCE OF INDIVIDUAL OBSERVATIONS :-

If $x_1, x_2, \ldots, x_n$ are n values of a variable X,

then

$$Var(x) = \frac{1}{n}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2$$

$$S.D. = \sigma = \sqrt{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]}$$

or

$$\sigma = \sqrt{\left[\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2\right]}$$

If the values of variable x are large, calculation of variance from the above formulae is quite tedious and time consuming.

In that case, we take deviation from an arbitrary point A (say), then

$$Var(x) = \frac{1}{n}\left[\sum_{i=1}^{n}(d_i - \bar{d})^2\right] = \frac{1}{n}\sum_{i=1}^{n} d_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} d_i\right)^2$$

where, $d_i = x_i - A$

and → let A be any Assumed Mean, and $d_i = x_i - A$

$$S.D. = \sigma = \sqrt{\left[\frac{\sum_{i=1}^{n} d_i^2}{n} - \left(\frac{\sum_{i=1}^{n} d_i}{n}\right)^2\right]}$$

Also,

$$\text{Var}(x) = \frac{h^2}{n}\left[\sum_{i=1}^{n}(u_i-\bar{u})^2\right]$$

$$\text{Var}(x) = h^2\left[\frac{1}{n}\sum_{i=1}^{n}u_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}u_i\right)^2\right]$$

where,

$$u_i = \frac{x_i - A}{h}$$

VARIANCE OF DISCRETE FREQUENCY DISTRIBUTION :-

If $x_1, x_2, \ldots, x_n$ are n observation with frequencies $f_1, f_2, \ldots, f_n$, then

$$\text{Var}(x) = \frac{1}{N}\left\{\sum_{i=1}^{n} f_i(x_i-\bar{x})^2\right\} = \left(\frac{1}{N}\sum_{i=1}^{n} f_i x_i^2\right) - \bar{x}^2$$

Where, $N = \sum_{i=1}^{n} f_i$

If the value of x or f are large, we take the deviations of the values of variable x from an arbitrary point A (say).

$\therefore$ $d_i = x_i - A$; $i = 1, 2, 3, \ldots, N$

$$S.D.(\sigma) = \sqrt{\left[\frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \bar{x})^2\right]}$$

Variance $= \sigma^2$

$\therefore$ $$Var(x) = \frac{1}{N}\left[\sum_{i=1}^{n} f_i (d_i - \bar{d})^2\right]$$

$$Var(x) = \frac{1}{N}\left(\sum_{i=1}^{n} f_i d_i^2\right) - \left(\frac{1}{N}\sum_{i=1}^{n} f_i d_i\right)^2$$

where, $N = \sum_{i=1}^{n} f_i$

$$S.D.(\sigma) = \sqrt{\left[\frac{\sum_{i=1}^{n} f_i d_i^2}{N} - \left(\frac{\sum_{i=1}^{n} f_i d_i}{N}\right)^2\right]}$$

where,

$$N = \sum_{i=1}^{n} f_i \quad , \quad N = \sum f_i$$

Sometimes, $d_i = x_i - A$ are divisible by a common number (h) say,

then

$$u_i = \frac{x_i - A}{h} \quad , \quad i = 1, 2, 3, \ldots, N$$

then

$$Var(x) = \frac{h^2}{N}\left[\sum_{i=1}^{n} f_i (u_i - \bar{u})^2\right]$$

$$\Rightarrow \quad Var(x) = h^2\left[\frac{1}{N}\sum_{i=1}^{n} f_i u_i^2 - \left(\frac{1}{N}\sum_{i=1}^{n} f_i u_i\right)^2\right]$$

where, $N = \sum_{i=1}^{n} f_i$

VARIANCE OF A GROUPED OR CONTINUOUS FREQUENCY DISTRIBUTION :-

Some formulae as used in discrete frequency distribution can be used.

$$\text{Var}(x) = \frac{\sum_{i=1}^{n} f_i d_i^2}{N} - \left(\frac{\sum_{i=1}^{n} f_i d_i}{N} \right)^2$$

where, $N = \sum_{i=1}^{n} f_i$, where, $N = \sum f_i$

$$\text{S.D.} = \sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i d_i^2}{N} - \left(\frac{\sum_{i=1}^{n} f_i d_i}{N} \right)^2}$$

NOTE:- Variance is independent of change of origin but dependent on change of scale. Adding or substracting a positive number from each observation of a

group does not affect the variance. If each observation is multiplied by a constant h, then variance of the resulting group becomes h^2 times the original variance.

- While calculating S.D., the deviations are to be taken about arithmetic mean only.

- $$\text{Coefficient of dispersion (S.D.)} = \frac{\text{S.D}}{\text{Mean}} = \frac{\sigma}{\bar{x}}$$

COEFFICIENT OF VARIATION :-

The M.D. and S.D. have the same units in which the data is given.

- $$\text{C.V} = \frac{\text{S.D.}}{\text{Mean}} \times 100 = \frac{\sigma}{\bar{x}} \times 100$$

The series having greater C.V. is said to be more variable and less consistent than the other.

STANDARD DEVIATION OF COMBINED GROUP OR SERIES :-

let $\bar{x}_1, \bar{x}_2$ be the respective means (AM) and σ_1, σ_2 be the respective S.D of the two given series having no. of observations as n_1 and n_2 respectively, then

Combined standard deviation σ_{12} of all the observations taken together is given by

$$\text{Combined SD } \sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2 + D_1^2) + n_2(\sigma_2^2 + D_2^2)}{n_1 + n_2}}$$

$$\text{Variance } (\sigma^2) = \frac{1}{n_1 + n_2}\left[n_1(\sigma_1^2 + D_1^2) + n_2(\sigma_2^2 + D_2^2)\right]$$

where,

$$d_1 = \bar{x}_1 - \bar{x}$$

$$d_2 = \bar{x}_2 - \bar{x}$$

and

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

CORRELATION :-

The tendency of simultaneous variation b/w two variable is called correlation or covariation.

It denotes the degree of inter-dependence b/w variables.

CORRELATION COEFFICIENT :-

The no. showing the degree or extent to which x are related to each other is called correlation coefficient.

It is denoted by $P(x,y)$ or

r_{xy} or simply r.

METHODS OF CALCULATING CORRELATION COEFFICIENT

(1) KARL PEARSON'S COEFFICIENT OF CORRELATION :-

Covariance (x,y) = $cov(x,y)$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \rightarrow \frac{1}{n}\sum_{i=1}^{n} x_iy_i - \bar{x}\bar{y}$$

let σ_x and σ_y be the SD of variables x and y, respectively.

Then, coefficient of correlation

$$r(x,y) = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

$$= \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{\{n\sum x_i^2 - (\sum x_i)^2\}\{n\sum y_i^2 - (\sum y_i)^2\}}}$$

(2) RANK CORRELATION (SPEARMAN'S) :-

Let d be the difference b/w paired ranks and n be the no. of items ranked.

Then, ρ the coefficient of rank correlation is given by

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d^2}{n(n^2-1)}$$

NOTE:
The rank correlation lies b/w -1 and 1.

PROPERTIES OF CORRELATION :-

(i) $-1 \leq r(x,y) \leq 1$

(ii) If $r = 1$, then the coefficient of correlation is perfectly positive.

(iii) If $r = -1$, the correlation is perfectly negative.

(iv) The correlation coefficient is a pure no. independent of the unit of measurement.

(v) The coefficient of correlation is independent of the change in origin and scale.

(vi) If $-1 < r < 1$, it indicates the degree of linear relationship b/w x and y, where as its sign tells about the direction of relationship.

(vii) If x and y are two independent variables, $r = 0$

(viii) If $r = 0$, x and y are said to be uncorrelated. If

does not imply that the two variates are independent.

$$r(x,y) = 0$$

(iv) If x and y are random variables and a, b, c and d are any nos. such that $a \neq 0$, $c \neq 0$, then

$$r(ax+b, cy+d) = \frac{|ac|}{ac} r(x,y).$$

LINES OF REGRESSION :-

It is the line which gives the best estimate to the value of one variable for any specific value of the other variable.

Therefore, the line of regression is the line of best fit and is obtained by the principle of least squares.

REGRESSION ANALYSIS :-

(1) The line of regression of y on x or regression line of y on x is given by

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

(ii) The line of regression of x on y or regression line of x on y is given by

$$x - \bar{x} = r \frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

(iii) Regression coefficient of y on x, is denoted by b_{yx},

$$b_{yx} = r \frac{\sigma_y}{\sigma_x} = \frac{cov(x,y)}{\sigma^2_x}$$

(iv) Regression coefficient of x on y, is denoted by b_{xy},

$$b_{xy} = r \frac{\sigma_x}{\sigma_y} = \frac{cov(x,y)}{\sigma^2_y}$$

(v) If θ is the angle between the two regression lines, then

$$\tan\theta = \frac{(1-r^2)}{|r|} \cdot \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma^2_y}$$

where, $\tan\theta = \frac{M_2 - M_1}{1 + M_1 M_2}$

(4) If $r = 0$, $\theta = \frac{\pi}{2}$, then the two regression lines are perpendicular to each other.

(5) If $r = 1$ or -1, $\theta = 0, \pi$, then the regression lines coincide.

PROPERTIES OF THE REGRESSION COEFFICIENTS :-

(i) Both regression coefficients and correlation coefficient r have the same sign.

(ii) Coefficient of correlation is the geometric mean between the regression coefficients.

(iii) If one of the regression coefficient is greater than unity, the other must be less than unity;

$$0 < |b_{xy} \cdot b_{yx}| \leq 1, \text{ if } r \neq 0$$

i.e., if $|b_{xy}| > 1$,

$|b_{yx}| < 1$.

(iv) Regression coefficients are independent of the change of origin but not of scale.

(v) Arithmetic Mean of the regression coefficient is greater than the correlation coefficient.

(vi) The two lines of regression cut each other at the points $(\bar{x}, \bar{y})$.

Thus, on solving the two lines of regression, we get the values of means of the variables in the bivariate distribution.