STATISTICS

Statistics is the science of collection, organization, presentation, analysis and interpretation of the numerical data.

Useful Terms:

1. **Limit of the Class:**
   The starting and end values of each class are called Lower and Upper limit.
2. **Class Interval:**
   The difference between upper and lower boundary of a class is called class interval or size of the class.
3. **Primary and Secondary Data:**
   The data collected by the investigator himself is known as the primary data, while the data collected by a person, other than the investigator is known as the secondary data.
4. **Variable or Variate:**
   A characteristics that varies in magnitude from observation to observation. E.g., weight, height, age, etc., are variables.
5. **Frequency:**
   The number of time an observation occurs in the given data, is called the frequency of the observation.
6. **Discrete Frequency Distribution:**
   A frequency distribution is called a discrete frequency distribution, if the data are presented in such a way that exact measurements of the units are clearly shown.
7. **Continuous Frequency Distribution:**
   A frequency in which data are arranged in classes groups which are not exactly measureable.

Cumulative Frequency Distribution:

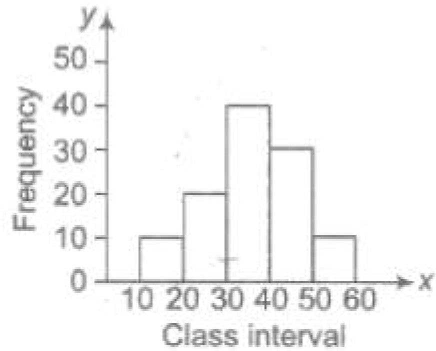
Suppose the frequencies are grouped frequencies or class frequencies. If however, the frequency of the first class is added to that of the second and this sum is added to that of the third and so on, then the frequencies, so obtained are known as cumulative frequencies (cf).

Graphical Representation of Frequency Distributions

(i) Histogram:

To draw histogram of a given continuous frequency distribution, we first mark off all the class intervals along x-axis on a suitable scale. On each of these class intervals on the horizontal axis, we erect (vertical) a rectangle whose height is proportional to the frequency of that particular class, so that the area of the rectangle is proportional to the frequency of the class.

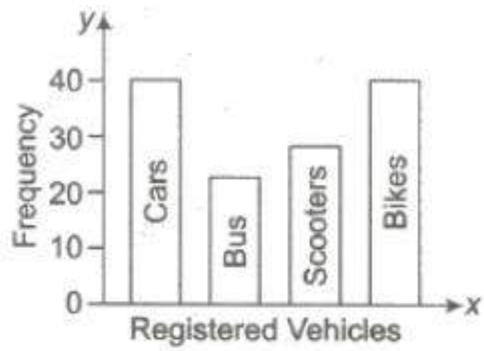
If however the classes are of unequal width, then the height of the rectangles will be proportional to the ratio of the frequencies to the width of the classes.



(ii) **Bar Diagrams:**

In bar diagrams, only the length of the bars are taken into consideration. To draw a bar diagram, we first mark equal lengths for the different classes on the axis, i.e., x-axis.

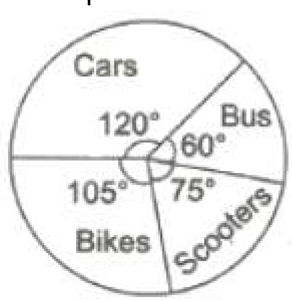
On each of these lengths on the horizontal axis, we erect (vertical) a rectangle whose heights is proportional to the frequency of the class.



(iii) **Pie Diagrams:**

Pie diagrams are used to represent a relative frequency distribution. A pie diagram consists of a circle divided into as many sectors as there are classes in a frequency distribution.

The area of each sector is proportional to the relative frequency of the class. Now, we make angles at the centre proportional to the relative frequencies.

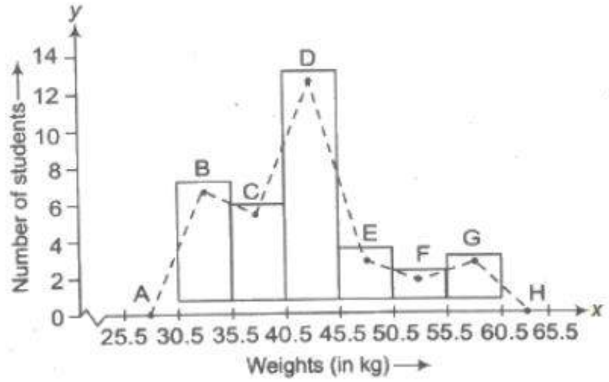


And in order to get the angles of the desired sectors, we divide $360°$ in the proportion of the various relative frequencies. That is,

$$\text{Central angle} = \frac{\text{Frequency} \times 360°}{\text{Total Frequency}}$$

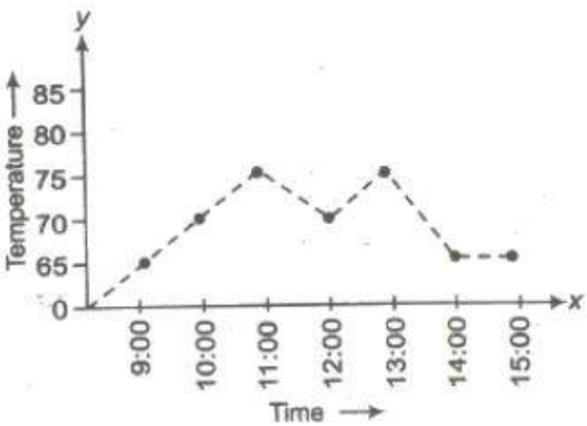
(iv) **Frequency Polygon:**

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 straight lines to obtain the frequency polygon.



(v) Cumulative frequency curve (Ogive)

Ogive is the graphical representation of the cumulative frequency distribution. There are two methods of constructing an Ogive, (i) the 'less than' method; (ii) the 'more than method.



Measure of Central Tendency

Generally, average value of a distribution in the middle part of the distribution, such type of values are known as measure of central tendency.

The following are the five measures of central tendency

Arithmetic Mean:

The arithmetic mean is the amount secured by dividing the sum of values of the items in a series by the number.

1. **Arithmetic Mean for Unclassified Data**

If n numbers, $x_1, x_2, x_3, \cdots, x_n$, then their arithmetic mean

A or $\bar{x} = \frac{x_1+x_2+x_3+\cdots+x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$

2. **Arithmetic Mean for Frequency Distribution**

Let $f_1, f_2, \cdots f_n$ be corresponding frequencies of $x_1, x_2, \cdots, x_n$. Then,

$$A = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_n}{f_1 + f_2 + \cdots f_n} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}$$

3. **Arithmetic Mean for Classified Data**

Class mark of the class interval $a - b$, then $x = \frac{a+b}{2}$

For a classified data, we take the class marks $x_1, x_2, \cdots, x_n$ of the classes as variables, then arithmetic mean

$$A = \frac{\sum xf}{\sum f} = \frac{\sum_{i=1}^{n} \frac{1}{2}(a_i + b_i) \times f_i}{\sum_{i=1}^{n} f_i}$$

Step Deviation Method

$$A = A_1 + \left(\frac{\sum_{i=1}^{n} f_i u_i}{\sum_{i=1}^{n} f_i}\right) h$$

Where, A_1= Assumed mean

$$u_i = \frac{x_i - A_1}{h}$$

f_i= frequency

h = width of interval

4. Combined Mean

If $x_1, x_2, \cdots, x_r$ be r groups of observations, then arithmetic mean of the combined group x is called the combined mean of the observation

$$A = \frac{n_1 A_1 + n_2 A_2 + \cdots n_r A_r}{n_1 + n_2 + \cdots n_r}$$

$\mathrm{A_r} = \text{AM of collection } \mathrm{x_r}$

$\mathrm{n_r} = \text{total frequency of the collection } \mathrm{x_r}$

5. **Weighted Arithmetic Mean**

If w be the weight of the variable x, then the weighted AM

$$A_w = \frac{\sum wx}{\sum w}$$

Shortcut Method:

$$\mathrm{A_w} = \mathrm{A'_w} + \frac{\sum \mathrm{wd}}{\sum \mathrm{w}}$$

$\mathrm{A'_w} = \text{assumed mean}$

$\sum \mathrm{wd}$ = sum of products of the deviations and weight

Properties of Arithmetic Mean

(i) Mean is independent of change of origin and change of scale.
(ii) Algebraic sum of the deviations of a set of values from their arithmetic mean is zero.
(iii) The sum of squares of the deviations of a set of values is minimum when taken about mean.

Geometric Mean:

If $x_1, x_2, \cdots, x_n$ be n values of the variable, then

$$G = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \cdots x_n} \ or\ G = antilog\left[\frac{\log x_1 + \log x_2 + \cdots + \log x_n}{n}\right]$$

For frequency distribution

$G = (x_1 f_1 \cdot x_2 f_2 \cdot \cdots \cdot x_n f_n)^{\frac{1}{N}}$, where $N = \sum_{i=1}^{n} f_i$

Or, $G = antilog\left[\frac{f_1\ \log x_1 + f_2 \log x_2 + \cdots + f_n \log x_n}{N}\right]$

Harmonic Mean(HM)

The harmonic mean of n items $x_1, x_2, \cdots, x_n$ is defined as

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

If their corresponding frequencies $f_1, f_2, \cdots f_n$ respectively, then

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

Median

The median of a distribution is the value of the middle variable when the variables are arranged in ascending or descending order.

Median (M_d) is an average of position of the numbers.

1. **Median for Simple Distribution**

Firstly, arrange the terms in ascending or descending order and then find the number of terms n.

(a) If n is odd, then $\left(\frac{n+1}{2}\right) th$ term is the median.

(b) If n is even, then there are two middle terms namely $\frac{n}{2}th\ and\ \left(\frac{n}{2}+1\right)th$ terms. Hence, Median=mean of $\frac{n}{2}th\ and\ \left(\frac{n}{2}+1\right)th$ terms.

2. **Median for Unclassified Frequency Distribution**

(i) First find N/2, where $N = \sum f_i$.

(ii) Find the cumulative frequency of each value of the variable and take value of the variable which is equal to or just greater than N/2.

(iii) This value of the variable is the median.

3. **Median for Classified Data (**Median Class**)**

If in a continuous distribution, the total frequency be N, then the class whose cumulative frequency is either equal to N/2 or is just greater than N/2 is called median class.

For a continuous distribution, median

$$M_d = l + \left(\frac{\frac{N}{2} - C}{f}\right) \times h$$

Where, $\mathrm{l} =$ lower limit of the median class

$\mathrm{f} =$ frequency of the median class

$\mathrm{N} =$ total frequency $= \sum f$

$\mathrm{C} =$ cumulative frequency of the class just before the median class

$\mathrm{h} =$ length of the median class

Quartiles:

The median divides the distribution in two equal parts. The distribution can similarly be divided in more equal parts (four, five, six etc.). Quartiles for a continuous distribution is given by

$$Q_1 = l + \left(\frac{\frac{N}{4} - C}{f}\right) \times h$$

Where, $\mathrm{l} =$ lower limit of the quartile class

$\mathrm{f} =$ frequency of the quartile class

$\mathrm{N} =$ total frequency $= \sum f$

C=cumulative frequency corresponding to the class just before the first quartile class

$\mathrm{h} =$ the length of the first quartile class

Similarly,

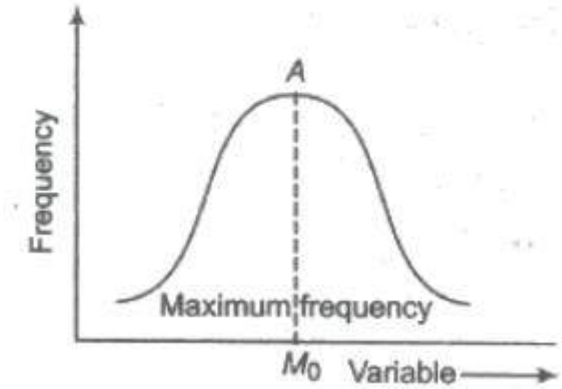
$$Q_3 = l + \left(\frac{\frac{3N}{4} - C}{f}\right) \times h$$

Where symbols have the same meaning as above only taking third quartile in the place of first quartile.

Mode

The mode (M_O) of a distribution is the value at the point about which the items tend to be most heavily concentrated. It is generally the value of the variable which appears to occur most frequently in the distribution.

1. **Mode for a Raw Data**
   Mode from the following numbers of a variable 70,80,90,96,70,96,96,90 is 96 as 96 occurs maximum number of times.



2. **For classified distribution**
   The class having the maximum frequency is called the modal class and the middle point of the modal class is called the crude mode.
   The class just before the modal class is called pre-modal class and the class after the modal class is called the post-modal class.

Mode for Classified Data (Continuous Distribution**)**

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

Where, l=lower limit of the modal class

f_0=frequency of the modal class

f_1=frequency of the pre-modal class

f_2=frequency of the post-modal class

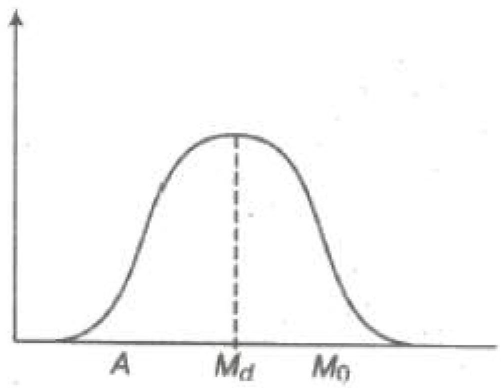
h = length of the class interval

Relation between Mean, Median and Mode

(i) $Mean - Mode = 3(Mean - Median)$
(ii) $Mode = 3Median - 2Mean$

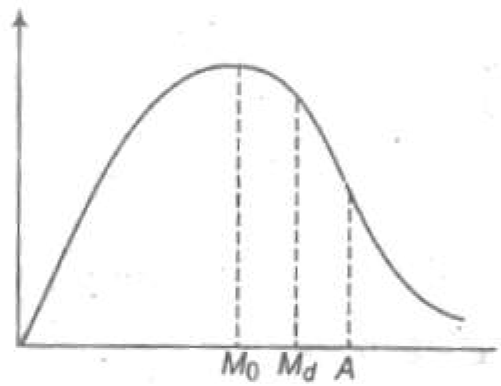
Symmetrical and Skew Distribution

A distribution is symmetric , if the same number of frequencies is found to be distributed at the same linear tense on either side of the mode. The frequency curve is bell shaped and $A = M_d = M_O$

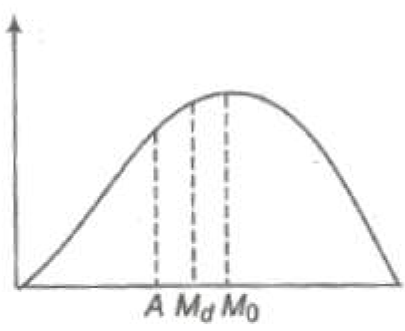


In anti-symmetric or skew distribution, the variation does not have symmetry.

(i) If the frequencies increases sharply at beginning and decreases slowly after modal value, then it is called positive skewness and $A > M_d > M_O$.



(ii) If the frequencies increases slowly and decreases sharply after modal value, the skewness is said to be negative and $A < M_d < M_O$



Measure of Dispersion

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

1. **Range:**
   The difference between the highest and the lowest element of a data called its range.
   $range = x_{max} - x_{min}$

$\therefore$ the coefficient of range $= \dfrac{\mathrm{x_{max}} - \mathrm{x_{min}}}{\mathrm{x_{max}} + \mathrm{x_{min}}}$

It is widely used in statistical series relating to quality control in production.

(i) Inter quartile range $= Q_3 - Q_1$

(ii) Semi-inter quartile range (quartile deviation) = $QD = \frac{Q_3 - Q_1}{2}$

(iii) Coefficient of quartile deviation $= \dfrac{Q_3 - Q_1}{Q_3 + Q_1}$

(iv) $QD = \frac{2}{3}SD$

2. **Mean Deviation:**

The arithmetic mean of absolute deviations of the values of the variable from a measure of their Average(mean, median, mode) is called Mean Deviation (MD). It is denoted by δ.

(i) For simple (discrete) distribution

$$\delta = \sum \frac{|x - z|}{n}$$

Where, n= number of terms, $z = A \text{ or } M_d \text{ or } M_O$

(ii) For unclassified frequency distribution

$$\delta = \frac{\sum f|x - z|}{\sum f}$$

(iii) For classified distribution

$$\delta = \frac{\sum f|x - z|}{\sum f}$$

Here, x is for class mark of the interval.

(iv) $MD = \frac{4}{5}SD$

(v) Average (mean or Median or Mode) $= \dfrac{\text{mean deviation from the average}}{\text{Average}}$

Note: the mean deviation is the least when measured from the median.

Coefficient of Mean Deviation

it is the ratio of MD and the mean from which the deviation is measured. Thus, the coefficient of $MD = \frac{\sum |x - \bar{x}|}{n}$

3. **Standard deviation** $(\boldsymbol{\sigma})$

Standard deviation is the square root of the arithmetic mean of the squares of deviations of the terms from their AM and it is denoted by σ.

The square of the standard deviation is called the **variance** and it is denoted by the symbol σ^2.

(i) For simple distribution

$$\sigma = \sqrt{\frac{\sum(x - \bar{x})^2}{n}} = \sqrt{\frac{\sum d^2}{n}}$$

(ii) For frequency distribution

$$\sigma = \sqrt{\frac{\sum f(x-\bar{x})^2}{\sum f}} = \sqrt{\frac{\sum fd^2}{\sum f}}$$

(iii) For classified data

$$\sigma = \sqrt{\frac{\sum f(x-\bar{x})^2}{\sum f}} = \sqrt{\frac{\sum fd^2}{\sum f}}$$

Here, x is class mark of the interval

Shortcut Method for SD $\boldsymbol{\sigma} = \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2}$

Where, $d = x - A'$ and A'= Assumed mean

Standard deviation of the Combined Series

If n_1, n_2 are the sizes, $\overline{X_1}, \overline{X_2}$ are the means and σ_1, σ_2 are the standard deviation of the series, then the standard deviation of the combined series is

$$\sigma = \sqrt{\frac{n_1(\sigma_1^2 + \sigma_2^2) + n_2(\sigma_1^2 + \sigma_2^2)}{n_1 + n_2}}$$

$$where, d_1 = \overline{X_1} - \bar{X} \; and \; d_2 = \overline{X_2} - \bar{X}$$

Effects of Average and Dispersion on Change of Origin and Scale

	Change of origin	Change of scale
Mean	Dependent	Dependent
Median	Not dependent	Dependent
Mode	Not dependent	Dependent
Standard Deviation	Not dependent	Dependent
Variance	Not dependent	Dependent

Important points to be remembered:

(i) The ratio of SD(σ) and the AM($\bar{x}$) is called the coefficient of standard deviation $\left(\frac{\sigma}{x}\right)$.

(ii) The percentage form of coefficient of SD i.e., $\left(\frac{\sigma}{x}\right) \times 100$ is called variation.

(iii) The distribution for which the coefficient of variation is less is called more consistent.

(iv) Standard deviation of first n natural numbers is $\sqrt{\frac{n^2-1}{12}}$.

(v) Standard deviation is independent of change of origin, but it is depend on change of scale.

4. **Root Mean Square Deviation(RMS):**

The square root of AM of squares of the deviations from an assumed mean is called the root mean square deviation. Thus,

(i) For simple(discrete) distribution

$S = \sqrt{\frac{\sum(x-A')^2}{n}}$, A' = Assumed mean

(ii) For frequency distribution

$$S = \sqrt{\frac{\sum f(x-A')^2}{\sum f}}$$

If A'=A then $S = \sigma$

Important Points to be Remembered:

(i) The RMS deviation is the least when measured from AM.

(ii) The sum of the squares of the deviation of the values of the variables is the least when measured from AM.

(iii) $\sigma^2 + A^2 = \frac{\sum f x^2}{\sum f}$

(iv) For discrete distribution f=1, thus $\sigma^2 + A^2 = \frac{\sum x^2}{n}$.

(v) The mean deviation about the mean is less than or equal to the SD. i.e., $MD \le \sigma$

Correlation:

The tendency of simultaneous variation between two variables is called correlation or covariance. It denotes the degree of inter-dependence between variables.

1. **Perfect Correlation:**
   If the two variables vary in such a manner that their ratio is always constant, then the correlation is said to be perfect.
2. **Positive or Direct Correlation:**
   If an increase or decrease in one variable corresponds to an increase or decrease in the other, then the correlation is said to be positive.
3. **Negative or Indirect Correlation:**
   If an increase or decrease in one variable corresponds to decrease or increase in the other, then the correlation is said to be negative.

Covariance:

Let $(x_i, y_i), i = 1,2,3,\cdots,n$ be a bivariate distribution where $x_1, x_2, \cdots, x_n$are the values of variable x and $y_1, y_2, \cdots, y_n$those as y, then the cov(x,y) is given by

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

Where, $\bar{x}\ and\ \bar{y}$ are mean of variables x and y.

(ii) $cov(x,y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)$

Karl Pearson's Coefficient of Correlation:

The correlation coefficient r(x,y) between the variable x and y is given by

$$r(x,y) = \frac{cov(x,y)}{\sqrt{var(x)var(y)}} \; or \; \frac{cov(x,y)}{\sigma_x \sigma_y}$$

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

If $(x_i, y_i), i = 1,2,\cdots,n$ is the bivariate distribution, then

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), -1 \le r(x,y) \le 1$$

Properties of Correlation:

(i) $-1 \le r \le 1$

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

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

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

(v) If $-1 \le r \le 1$, it indicates that the degree of linear relationship between x and y, where as its sign tells about the direction of relationship.

(vi) If x and y are two independent variables, r=0.

(vii) If r=0, x and y are said to be uncorrelated. It does not imply that the two variates are independent.

(viii) If x and y are random variables and a,b,c and d are any numbers such that $a \neq 0, c \neq 0$, then $r(ax+b, cy+d) = \frac{|ac|}{ac} r(x,y)$

(ix) **Rank Correlation** (Spearman's)

Let d be the difference between paired ranks and n be the number of items ranked. The coefficient of rank correlation is given by

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

(a) The rank coefficient correlation lies between -1 and 1.

(b) If two variables are correlated, then points in the scatter diagram generally cluster around a curve which we call the curve of regression.

(x) **Probable error and Standard Error:**

If r is the correlation coefficient in a sample of n pairs of observations, then its standard error is given by $\frac{1-r^2}{\sqrt{n}}$

And the probable error of correlation coefficient is given by $(0.6745)\frac{1-r^2}{\sqrt{n}}$

Regression:

The term regression means stepping back towards the average.

Line of Regression:

The line of regression 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

(i) Line of regression of y on x,

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

(ii) Line of regression of x on y,

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

(iii) Regression coefficient of y on x and x on y is denoted by

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

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

(iv) Angle between two regression line is given by

$$\theta = \tan^{-1}\left\{\left(\frac{1-r^2}{r}\right)\left(\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}\right)\right\}$$

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

(b) If r=1, or -1, $\theta = 0$, so the regression line coincide.

Properties of the Regression Coefficients

(i) Both regression coefficients and r have the same sign.
(ii) Coefficient of correlation is the geometric mean between the regression coefficients.
(iii) $0 < |b_{xy}b_{yx}| < 1$, if $r \neq 0$
(iv) Regression coefficients are independent of the change of origin but not of scale.
(v) If two regression coefficient have different sign, then r=0.
(vi) Arithmetic mean of the regression coefficients is greater than the correlation coefficient.