STAT-231

Topic : ELEMENTARY CONCEPT OF STATISTICS

Introduction :

The word statistics comes from the Italian word 'Statista' or the German word 'Statistik' each of which means a political state.

In the old days, statistics was regarded as the 'science of statecraft' and was by-product of the administrative activity of the state.

Government record is the earliest foundation of statistics. Government has a traditional function/job to keep records of population, births, deaths, taxes, crop yields and so on. Counting and measuring these types of event may generate many kinds of data. Statistics is said to be a branch of applied mathematics. The present body of the statistical methods, particularly, those concerned with drawing inferences is based on the mathematical theory of probability. Hence, the science of statistics is originated from two main sources, government records and mathematics.

Statistics is used for the collection, analysis and interpretation of data in order to provide basis for making correct decisions. The word statistics was originally applied to data which was collected and required to state/nation for its official purposes. Now, the word statistics is used not only for the material (numerical data) which is analyzed, but also for the methods applied in its analysis.

Hence, the word statistics is used in two senses i) As numerical data and

ii) As statistical methods.

In the first sense, the word statistics is refers to numerical descriptions (counts and measures) of the quantitative aspects of things. In the second sense the word statistics refers to statistical principles and methods used in collection, presentation, analysis and interpretation of data.

Statistics defined : - The term statistics has been defined differently by different authors.

Definition of statistics as per first sense the 'numerical data' given by the author Prof. Horace Secrist.

' By statistics we mean aggregates of facts, affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in systematic manner for a pre-defined purpose and placed in relation to each other.'

From this definition, it is clearly seen that numerical data may be called as statistics if and only if it posses certain characteristics. These characteristics are as follows.

1) Statistics are aggregates of facts : Single and isolated figures are not statistics.
2) Statistics are affected to a marked extent by multiplicity of causes : Generally, facts and figures are affected to a considerable extent by a number of forces operating together. For example, production of rice depends on factor such as

seasonal rainfall, fertilizers applied, quality of seed, soil, methods of cultivation, seasonal weather and so on, hence data collected/related on production may constitute statistics.

3) Statistics are numerically expressed : Statement of facts must contain appropriate numerical data. The qualitative statements do not constitute statistics.
4) Statistics are enumerated or estimated according to reasonable standard of accuracy : Standard of accuracy in counting and measuring any characteristics of the thing is important. The accuracy in measuring distance in centimeter is $1/10^{th}$ of cm, in feet is $1/12^{th}$ of foot, in meter is $1/100^{th}$ of meter, in kilometer is $1/1000^{th}$ Km. In measuring distance around 800 meters, it is not necessary to take into account fractional value (few distance in centimeters may be neglected/ignored).
5) Statistics are collected in a systematic manner : Before collecting statistics (i.e. numerical facts), it is very important to prepare suitable plan of data collection.
6) Statistics are collected for a pre-determined purpose : Before starting to collect data, it is necessary to decide purpose of data collection. The purpose of data collection should be specific and well defined.
7) Statistics should be placed in relation to each other : The numerical facts are called statistics, if and only if they are comparable. Statistics are collected mostly for the purpose of comparison. If related data are placed near about, the comparison will be done easily. Statistical data are often compared period wise or region wise.

Definition of statistics as per second sense 'Statistical methods' or 'Science of statistics' given by Croxton and Cowden.

"Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data."

If one more term 'Organization of data' is added in this definition then the definition becomes. "Statistics may be defined as the science of collection, organization, presentation, analysis and interpretation of numerical data."

From this definition it is seen that there are five stages in a statistical investigation. These are collection of data, organization of data, presentation of data, analysis of data and interpretation of data.

1) Collection : The collection of data is the main foundation of statistical analysis. If collected data are faulty, the conclusion drawn from such data can never be reliable.
2) Organization : Data collected from published sources are generally in organized form. Data which are collected or generated from survey frequently need organization. The data are organized in three different stages namely editing, classification and tabulation.
3) Presentation : The data can be presented through diagrams and graphs.

4) Analysis : The purpose of analyzing data is to dig out information useful for decision making. Some commonly used methods of statistical analysis are measures of central tendency, measures of dispersion, correlation, regression etc.
5) Interpretation : The last stage in statistical investigation is interpretation. The interpretation means drawing numerical conclusions from the data collected and analyzed. Correct interpretation can aid in taking suitable decisions.

Limitations of statistics :

1) Statistics does not deal with individual measurements : Since, the statistics deals with aggregates of facts, the study of individual measurements lies outside the scope of statistics.
2) Statistics deals only with quantitative characteristics : Statistics are numerical statement of facts. Thus, qualitative characteristics like honesty, efficiency, intelligence etc. can not be studied directly. We may study the intelligence of boys on the basis of marks obtained by them in an examination.
3) Statistical results are true only on an average : The conclusions obtained statistically are not universally true. Conclusions are true only under certain conditions.
4) Statistics is only one of the methods of studying a problem : Statistical tool do not provide the best solution under all circumstances. Statistics can not be of much help in studying problems like country's culture, religion and philosophy.
5) Statistics can be misused : The misuse of statistics may arise due to lack of complete information, improper use of statistical methods, improper interpretation, improper understanding of the subject and inexperienced people.

Some definitions in Statistics :

Population : Any well defined set of objects about which a statistical enquiry is being made is called a population or universe. The objects may be people, trees, T.V. sets, bulbs, students, etc. The total number of objects/items/individuals in a population is called the size of the population.

Individual : Each object belonging to a population is called an individual of the population.

Sample : A finite set of objects drawn from the population with an aim to make observations from them, is called the sample.

Sample size : The total number of individuals in a sample is called as the sample size.

Characteristic : The parameter on which information required from an individual during the statistical enquiry (survey) is known as the characteristic of an individual. Observations/information of characteristic of an individual under study may be non-numeric (descriptive) or numeric (qualitative or quantitative). If observation is **non-numeric**, the characteristic under study is called as **qualitative** characteristic (**attribute**). If observation is **numeric**, the characteristic under study is called as **quantitative** characteristic (**variate/variable)**

Attribute : A qualitative characteristic of an individual which can not be expressed

numerically is called as attribute. e.g characteristics like colour of eye, colour of hairs, sex, occupation etc.

Variable : A quantitative characteristic of an individual which can always be expressed numerically is called variate or variable. e.g height, weight, income, marks etc.

Discrete variate : A variate which is not capable of assuming numerical values in a given range is called as a discrete variate. (variate which takes values such as 1, 2, 3, ….. i.e counting numbers or whole numbers)e.g. characters like marks, number of students, number of plants, …..

Continuous variate : A variate which is capable of assuming all the numerical values in a given range is called as a continuous variate. (variate which takes values such as 1, 1.23, 3.546, 8, 9, 3.22 …..) e.g characters like height, weight, income, temperature ……..

Topic : FREQUENCY DISTRIBUTION (CLASSIFICATION)

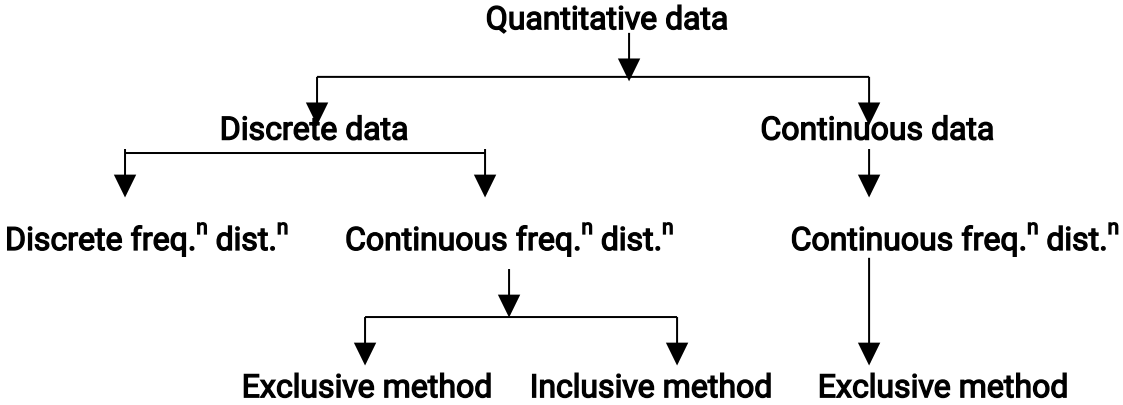
Quantitative Classification (classification according to variable or class interval) :

Quantitative classification refers to the classification of data according to some characteristic that can be measured numerically such as height, weight, income, production etc. In a quantitative classification framing of different classes is important. For example, the students of a college may be classified according to weight as follows

Weight (Kg) (Class)	Number of Students
40-50	250
50-60	50
60-70	29

This type of distribution is known as empirical or simple or univariate frequency distribution. Here we consider only two elements i) the variable i.e. the weight and ii) the frequency i.e. the Number of students.

Classification of numerical data/quantitative data/variable :



Frequency distribution : A frequency distribution or frequency table is simply a table in which the data is grouped in continuous classes (or discrete classes) and the number of cases/items which fall in each class are recorded. The number of cases/items in each class is referred as the **frequency**.

Relative frequency distribution : In a frequency distribution/frequency table if each class frequency is expressed by their proportion (ratio of frequency and total frequency), the table is referred to as relative frequency distribution or simply percentage frequency distribution.

Formation of discrete frequency distribution : To prepare discrete frequency distribution, we have just to count the number of times a particular value is repeated and this repeated number is called **frequency** of that class. In a column under heading class, place all possible values of variables (from given data) from the lowest to the highest value. Prepare another column 'tallies' or 'tally marks' to facilitate counting. For each & every item in the data, put the tally mark or bar (small vertical line) against the particular class (discrete value) to which it relates. To facilitate counting, block of five bars are prepared. Finally, count number of bars and get frequency. If $X_1, X_2, \ldots, X_N$ are N individual values of **discrete** variable X, then the format of discrete frequency distribution is as follows.

Class (X)	Tally marks	Frequency (f)
x_1	|||| |||	f_1
x_2	|||| |||| ||	f_2
:	:	:
:	:	:
x_n	|||	f_n
		$N = \sum f_i$

Where, $x_1, x_2, \ldots, x_n$ are n all possible values of discrete variable X.

Continuous frequency distribution : Some technical terms -

Class Limits : The class limits are the lowest and highest values that can be included in the class.

Lower Limit of Class : The lower limit of the class is the value below which there can be no item in the class.

Upper Limit of Class : The upper limit of a class is the value above which there can be no item in the class.

Class interval (width of class) : The difference between upper and lower class limit of a class is known as class interval or width of that class. The width of class is depends on difference between the largest and the smallest item, the number of classes to be formed and the details required as per problem.

Class frequency : The number of observations/values/items corresponding to a particular class is known as frequency of that class or class frequency.

Class mid-points or class mark : It is the value lying half-way between the lower and upper limits of the class interval. Mid-point of the class represents that class. It can be calculated with the help of following formula.

Upper limit of a class + Lower limit of

Mid-point of a class = $\frac{\qquad\qquad}{2}$

According to class-intervals of continuous frequency distribution, there are two methods of classification i) Exclusive method and ii) Inclusive method

i) **Exclusive method:** (First explain quantitative classification) When the class intervals are so fixed that the upper limit of one class is the lower limit of the next class, it is known as the exclusive method of classification. **OR** When the class intervals are so fixed that the lower class limit is included in it but the upper class limit is excluded from it, then it is known as exclusive method of classification. This type of classification is useful for discrete as well as continuous data.

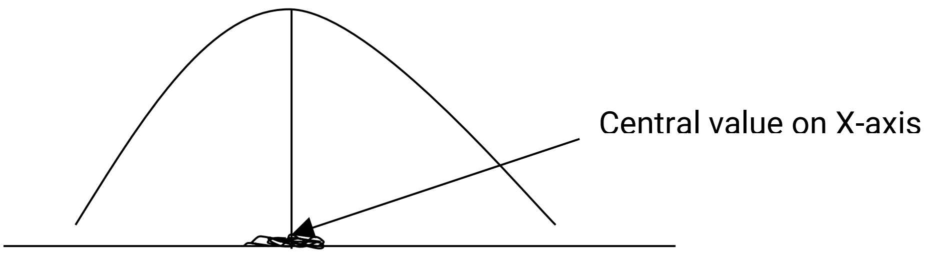
ii) **Inclusive method:** (First explain quantitative classification) When the class intervals are so fixed that the lower limit as well as upper limit of one class is included in that class itself, then it is known as inclusive method of classification. This type of classification is useful only for discrete data.

(Note : The number of classes should be between 4 and 20. Class interval should be 5 or multiple of 5. The lower limit of the class should be either 0 or 5 or multiple of 5. To ensure continuity and to get correct class intervals, we should use 'exclusive' method of classification. It is desirable to use class intervals (width of class) of equal sizes. This enables us to comparison of frequencies among different classes and subsequent statistical analysis. Avoid open end classification.)

Topic 3. MEASURES OF CENTRAL TENDENCY

Introduction:

If we observe the frequency distribution of quantitative data, we find that the frequencies of central classes are high as compared to starting and ending classes. If we draw frequency polygon or frequency curve of such a frequency distribution then it looks like bell shaped as shown in diagram. If individual values are located on X-axis, maximum values lie around central values as shown in diagram.



From this it is seen that maximum number of values/observations/items crowded at central portion or around central value. The property of concentration of the values around a central value is called central tendency of the frequency distribution. The central value around which there is a concentration is called as measure of central tendency **or** measure of location **or** an average **or** expected value.

One of the most important objectives of statistical analysis is to get single value that describes the characteristics of the entire mass of data, such a value is called the central value **or** an average **or** the expected value of the variable.

Average defined (by croxten and cowden) :

Since, average represents the entire data, it value lies somewhere in between

the two extremes. Hence, an average is referred to as a measure of central tendency.
(Objectives of average : To get single value that describes the characteristic of the entire group . To facilitate comparison)

Characteristics (requisites or properties or qualities or requirements) of a ideal (good) average : Average is a single value representing a group of values, hence, it should satisfies following properties.

1. It should be easy to understand : Average should be easy to understand, otherwise, its use becomes very limited.
2. It should be simple to calculate : If it is simple to calculate, it can be used widely. Sometimes, in the interest of greater accuracy, use of more difficult average is desirable.
3. It should be based on all items : The average should be depends upon each and every item of the given data, so that, if any of the items is dropped/changed in magnitude, the average itself is altered.
4. It should not be unduly affected by extreme observations : Each and every items of data should influence the value of the average, none of the items should influence it unduly. If some very small values or very large values (extreme values) present in a given data, such values unduly affect some averages.
5. It should be rigidly defined : An average should be properly defined, so that it has one and only one interpretation. It should be preferably be defined by an algebraic/mathematical formula.
6. It should be capable of further algebraic treatment : If selected average is useful in further algebraic/mathematical treatment/computation then its utility can be enhanced.
7. It should have sampling stability : If there is no significant difference between averages obtained/calculated from two or more samples drawn from a population, then we can say that such an average have sampling stability.

Types of averages : Arithmetic mean, Medium, Mode, Geometric mean and Hermonic mean

Arithmetic mean (A.M.) : The arithmetic mean is most popular and widely used measure of central tendency or average. Arithmetic mean is obtained by adding together all the items and dividing this total by the number of items.

Simple arithmetic mean - Individual observations/series (ungrouped data) :

If $X_1, X_2, X_3, \ldots\ldots\ldots\ldots, X_N$ are N various values of the variable X, then the simple arithmetic mean of all the values can be calculated by the formula.

$$\bar{X} \quad \frac{X_1 + X_2 + \ldots\ldots\ldots\ldots + X_N}{N}$$

$$\text{i.e } \bar{X} = \frac{\sum X_i}{N}$$

Where, $\bar{X}$ = Simple arithmetic mean, $\sum X_i$ = Sum of all the values, N= Number of values.

Arithmetic mean - Discrete series/Discrete frequency distribution/Discrete grouped data

If $X_1, X_2, \ldots\ldots\ldots, X_N$ are N various values of discrete variable X, but given in the form of discrete frequency distribution (1^{st} and 2^{nd} column) then to calculate arithmetic mean it is necessary to prepare following table.

Class (X)	Frequency (f)	Frequency * Class (f * X)
x_1	f_1	$f_1 * x_1$
x_2	f_2	$f_2 * x_2$
:	:	:
:	:	:
:	:	:
x_n	f_n	$f_n * x_n$
	$N = \sum f_i$	$\sum f_i\ x_i$

$$\bar{X} = \frac{\sum f_i\ x_i}{N}$$

Where, $\bar{X}$ = Arithmetic mean of discrete series

$\sum f_i x_i$ = Sum of product of various discrete classes and frequencies.

N = Sum of all the frequencies i.e. total number of values

Arithmetic mean - Continuous series/Continuous frequency distribution/Continuous grouped data :

If $X_1, X_2, \ldots\ldots\ldots, X_N$ are N various values of variable X, but given in the form of continuous frequency distribution (1^{st} and 2^{nd} column) then to calculate arithmetic mean it is necessary to prepare following table.

Class (X)	Frequency (f)	Mid-points (m)	Frequency * mid-points (f * m)
$L_1 - U_1$	f_1	m_1	$f_1 * m_1$
$L_2 - U_2$	f_2	m_2	$f_2 * m_2$
:	:	:	:
:	:	:	:
:	:	:	:
$L_n - U_n$	f_n	m_n	$f_n * m_n$
	$N = \sum f_i$		$\sum f_i\ m_i$

$$\bar{X} = \frac{\sum f_i\ m_i}{N}$$

Where, $\bar{X}$ = Arithmetic mean of continuous series

$\sum f_i\ m_i$ = Sum of product of mid-points of the classes and their respective frequencies.

N = Sum of all the frequencies i.e. total number of values

Merits of arithmetic mean or mean :

1. It is the simplest average to understand and easiest to compute
2. It is affected by each and every item of the series i.e it is based on all items of the series.
3. It is defined by a rigid mathematical formula
4. It is capable of further subsequent algebraic treatment
5. It is relatively reliable in the sense of sampling
6. It is the centre of the gravity, balancing the values on either side of it.
7. It is calculated value and not based on position in the series.

Demerits/Limitation of arithmetic mean or mean :

1 It is unduly affected due to very small or very large items (extreme values).

2 It is not possible to calculate arithmetic mean if given frequency distribution is open ended.

Median :

The median refers to the middle value in a distribution. The place of the median in a series (ascending or descending) is such that an equal number of items lie on either side of it. The median is determined/calculated by its location/position in the series (ascending or descending). Hence, it is also called positional average.

Median - Individual observations/series (ungrouped data)

If $X_1, X_2, X_3, \ldots\ldots\ldots\ldots, X_N$ are N various values of the variable X, then to find the median, first arrange these values in ascending or descending order of magnitude and then apply one of the following formula.

If N is odd, Median = Size of $(N+1)/2^{th}$ item i.e. middle item/observation

If N is even, Median = [Size of $(N/2)^{th}$ items + size of $(N/2)+1^{th}$ item] /2

$$= \frac{\text{Sum of two middle items}}{2}$$ (i.e arithmetic mean of middle two values)

Median - Discrete series/Discrete frequency distribution/Discrete grouped data :

If $X_1, X_2, \ldots\ldots\ldots, X_N$ are N various values of discrete variable X, but given in the form of discrete frequency distribution (1^{st} and 2^{nd} column) then to calculate median prepare following table.

Class (X)	Frequency (f)	Cumulative frequency (CF <=)	
x_1	f_1	cf_1	
x_2	f_2	cf_2	
:	:	:	
()	()	()	>= (N+1)/2 (say)
:	:	:	
x_n	f_n	cf_n	

i) Arrange the values in ascending or descending order of magnitude (smallest to largest)

ii) Find out cumulative frequencies.

iii) Find the values of (N + 1)/ 2

iv) Look at the cumulative frequency column and find which cumulative frequency is equal to (N+1)/2 or just higher than (N+1)/2 and detect a discrete class

corresponding to such a cumulative frequency (such discrete class is denoted by () in 1st column). Such a discrete class becomes median value for given discrete series.

Median–Continuous series/ Continuous frequency distribution/ Continuous grouped data :

If $X_1, X_2, \ldots\ldots\ldots, X_N$ are N various values of variable X, but given in the form of continuous frequency distribution (1st and 2nd column) then to calculate median prepare following table.

Class (X)	Frequency (f)	Cumulative frequency (CF <=)	
$L_1 - U_1$	f_1	cf_1	
$L_2 - U_2$	f_2	cf_2	
:	:	:	
() : ()	()	()	>= N/2 (say)
:	:	:	
$L_n - U_n$	f_n	cf_n	

i) Find cumulative frequencies
ii) Calculate N/2
iii) Look at the cumulative frequency column and find which cumulative frequency is either equal to N/2 or higher than N/2 and detect a continuous class corresponding to such a cumulative frequency. Such a continuous class is known median class (such continuous class is denoted by () : () in 1st column).

$$\text{Median} = L_m + \frac{N/2 - c f_{m-1}}{f_m} * i_m$$

Where, L_m = Lower limit of the median class
N = Sum of the frequency column = $\sum f_i$
Cf_{m-1}=Cumulative frequency of the class preceding the median class
f_m = Simple frequency of the median class
i_m = The class interval of median class.

Properties of median (algebraic/mathematical)

The sum of the deviation of the items from median, ignoring signs, is the least.

i.e. $\sum | X_i - M_d | < \sum | X_i - A |$

Where, M_d = Median and A = Any other value other than median

Merits:

i) It is useful in case of open end classes.
ii) Extreme values do not affect the median as strongly as they do the arithmetic mean
i.e it is not unduly affected by extreme values (very small or very large values).
iii) In markedly skewed distribution, the median is useful
iv) It is useful in case of ranked or scored qualitative data
v) It can be determined graphically.

Demerits/Limitations:

i) To find median it is necessary to arrange data in ascending or descending order of magnitude. If number of values are too large then it is difficulty to arrange such data in ascending or descending order of magnitude.

ii) It is not determined/calculated by each and every observation i.e it is not based on all the observations.

iii) It is not capable of further algebraic treatment

iv) If the number of items in a series is even, the median is determined/calculated approximately as the mid-point or arithmetic mean of two middle items.

Mode :

The mode or the modal value is that value in a series of observations, which occurs with the greatest frequency. The mode is the most repeated value in the series of observations. Mode is also called as modal value or most typical value or most representative value or peak in distribution.

Mode of distribution (definition by croxtend and cowden):

The mode of a distribution is the value at the point around which the items tend to be most heavily concentrated. It may be regarded as the most typical value of a series of values.

Mode - Individual observations/series (ungrouped data)

$X_1, X_2, X_3, \ldots\ldots\ldots\ldots, X_N$ are N various values of the variable X, then to find mode, count the number of times the various values repeat and the value occurring maximum number of times is the modal value.

Mode – Discrete series/Discrete frequency distribution/Disscrete grouped data :

If $X_1, X_2, \ldots\ldots\ldots, X_N$ are N various values of discrete variable X, but given in the form of discrete frequency distribution as follows.

Class (x)	:	x_1	x_2	. . . ()		x_n
Frequency (f)	:	f_1	f_2	. . . ()		f_n

To determine mode, find maximum frequency and the discrete class value (such a discrete class is denoted by () in the1st row) corresponding to this frequency is the mode of given discrete series/distribution.

Mode – Continuous series/ Continuous frequency distribution/ Continuous grouped data :

$X_1, X_2, \ldots\ldots\ldots, X_N$ are N various values of variable X, but given in the form of continuous frequency distribution as follows.

Class	:	$L_1 - U_1$	$L_2 - U_2$	. . . ()-()		$L_n - U_n$
Frequency (f)	:	f_1	f_2	. . . ()		f_n

To determine mode, find maximum frequency and the class corresponding to this frequency is the modal class (such a continuous class is denoted by ()-() in the 1st row) of given continuous frequency distribution. Then the value of mode for given continuous series/distribution can be determined by the formula

$$Mo = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} * i$$

Where, Mo = Mode

L = lower limit of modal class
f_o = frequency of the class preceding the modal class
f_1 = Frequency of the modal class
f_2 = Frequency of the class succeeding the modal class.
i = class interval of modal class.

(Note : Class interval of all the classes should be uniform)

Merits :

i) By definition mode is the most typical or representative value of a distribution.
ii) Mode is not unduly affected by extreme values.
iii) It is useful in case of open-end classes.
iv) The value of mode can be determined graphically.

Demerits/Limitations :

i) The value of mode cannot be always be determined (in case of bi-modal or multi-modal series) for all the series of values.
ii) It is not capable of further algebraic treatment.
iii) The value of mode is not based on each and every item of the series.
iv) It is not rigidly defined.

Relationship among Mean, Median and Mode :

Symmetrical distribution : A distribution in which the values of mean, median and mode are coincide (i.e. mean = median = mode) is known as symmetrical distribution.

Skewed or asymmetrical distribution : A distribution in which the values of mean, median and mode are not equal is known as skewed or asymmetrical distribution.

In moderately skewed or asymmetrical distributions the distance between the mean and the median is about one-third the distance between the mean and the mode.

. Mode = 3 * Median – 2 * Mean

Topic : MEASURES OF DISPERSION

Introduction

We have seen the various measures of central value or average (e.g Arithmetic mean, median, mode, geometric mean and harmonic mean) which gives us one single value/figure that represents the entire data or set of observations, but it alone can not adequately describes set of observations. Hence, it is necessary to see the variability or dispersion of the set of observations in addition to its measure of central value or an average. The central value or the average of two or more different distributions/set of observations may be the same but there can be no same dispersion/disparities in the formation of such distributions/ set of observations. Significance of measuring variation/dispersion are to determine the reliability of an average, to serve as basis for the control of the variability, to

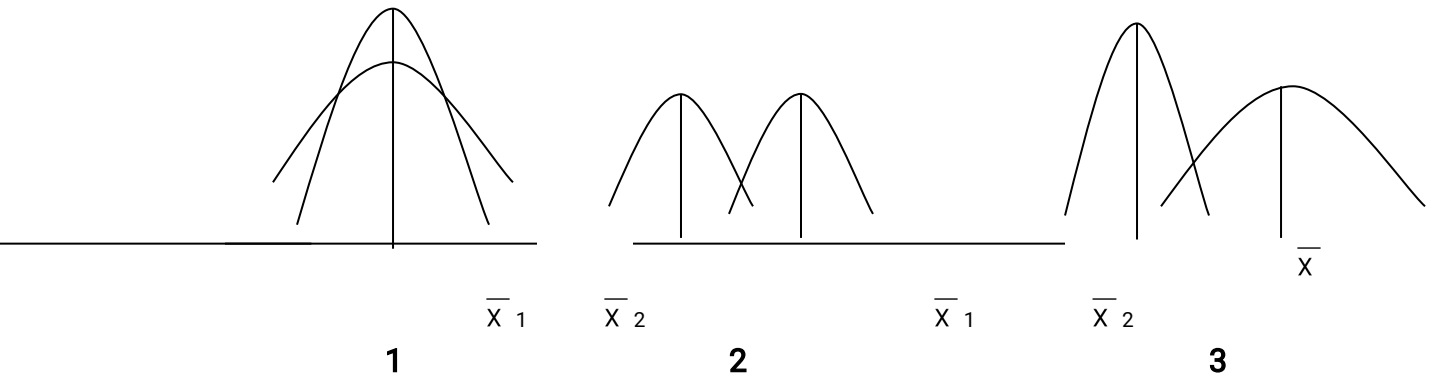
compare two or more series with regard to their variability and it facilitate the use of other statistical measures.

Dispersion defined :

1. Dispersion is the measure of the variation of the items. – A.L. Bowley.
2. The degree to which numerical data tend to spread about an average is called as the variation or dispersion of the data – Spiegel.
3. Dispersion or spread is the degree of the scatter or variation of the variable about a central value – Books and Dick
4. The measurement of the scatterness of the mass of figures in a series about an average is called measure of variation or dispersion – Simpson and Kafka.

By definition, the dispersion (also known as scatter, spread or variation) measures the extent to which the items vary from some central value. Measures of dispersion, also called as averages of the second order. (An average or measures of central value is more meaningful when it is examined in the light of dispersion).

Study the following three different types of figures representing frequency distribution.



In 1st case, there are two distributions and have same central value (Say, $\bar{X}$) but have different dispersions (Flatness of curve indicates dispersion)

In 2nd case, there are two distributions and have different central values (Say, $\bar{X}_1$ and $\bar{X}_2$) but have same dispersion.

In 3rd case, there are two distributions have different central values (Say, $\bar{X}_1$ and $\bar{X}_2$) and different dispersions.

Hence to describe set of observations by measure of central value or an average, it is necessary to support it with other measure like dispersion.

Methods of Studying variation/dispersion

i) The range, ii) The interquartile range or the quartile deviation. iii) The mean deviation or average deviation iv) The standard deviation and v) The Loren curve.

The standard deviation (S. D.)

The standard deviation concept was introduced by Karl Pearson in 1823. It is the most important and widely used measure of studying dispersion. It satisfies most of the properties of a good measure of dispersion. Standard deviation is also known as root mean square deviation. Standard deviation is denoted by the small

Greek letter σ (read as sigma).

Definition : Standard deviation is the square root of the mean of the squared deviations of individual values from the arithmetic mean.

Standard deviation of individual observations : If X_1, X_2, X_3, ... , X_N are N various values of the variable X, then by definition

$$\text{Standard deviation (S. D.)} = \sqrt{\frac{\sum (X_i - \overline{X})^2}{N}}$$

Where, $\overline{X}$ is arithmetic mean of individual observations

Variance : The square of the standard deviation is known as variance and

$$\text{Variance} = \frac{\sum (X_i - \overline{X})^2}{N}$$ i.e Variance = S.D.^{2}

Where, $\overline{X}$ = is arithmetic mean of individual observations

Standard deviation of discrete series or discrete frequency distribution : If x_1, x_2, ... , x_n are n various values of discrete classes with f_1, f_2, ... , f_n as frequencies then standard deviation is

$$\text{Standard deviation (S. D.)} = \sqrt{\frac{\sum f_i * (x_i - \overline{X})^2}{N}}$$

Where, $N = \sum f_i$ = sum of frequencies

$\overline{X}$ = arithmetic mean of discrete frequency distribution

Standard deviation of continuous series or continuous frequency distribution : If m_1, m_2, ... , m_n are n various mid points of continuous classes with f_1, f_2, ... , f_n as frequencies then standard deviation is

$$\text{Standard deviation (S. D.)} = \sqrt{\frac{\sum f_i * (m_i - \overline{X})^2}{N}}$$

Where, $N = \sum f_i$ = sum of frequencies

$\overline{X}$ = arithmetic mean of continuous frequency distribution

Relative measure of standard deviation is known as coefficient of variation and is given by

$$\text{Coefficient of variation (C.V.\%)} = \frac{\text{S. D.}}{\overline{X}} * 100$$

Topic : STUDY OF CORRELATION ANALYSIS

Introduction :

We have studied measures of central value and measures of dispersion of one variable only. If two variables vary in such a way that movements (increase or decrease) in one variable are accompanied by movements in the other, then these variables are said to be correlated. For example, there exist some relationship between age of husband and wife, price and supply of commodity, rainfall and crop production, height and weight of boys and so on. The degree of relationship

between the variables under consideration is measured through the correlation analysis. The measure of correlation is also known as the correlation coefficient or correlation index or association between variables and it gives direction and degree/magnitude of relationship in single figure/value. It is possible to measure closeness or the relationship between the variables with the help of correlation analysis. Correlation is the appropriate statistical tool to measure relationship. With the help of correlation analysis we can measure in single figure/value the direction and degree of relationship between two series of the values (variables). Once we come to know from correlation analysis that there exists relationship between two variables, we can estimate the value of dependent variable for any value of independent variable. This can be done using regression analysis. The prediction based on correlation analysis is likely to be near to reality.

Definitions :

1) Correlation analysis deals with the association between two or more variables.
2) If two or more quantities vary in sympathy so that movements in one tend to be accompanied by corresponding movements in the other(s) then they are said to be correlated.
3) When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in brief formula is known as correlation.
4) Correlation analysis attempts to determine the degree of relationship between variables.
5) Correlation is an analysis of the co-variation between two or more variables.

Types of correlation : 1. Positive or negative. 2. Simple, multiple and partial..
3. Linear and non-linear.

1) **Positive Correlation :** If both the variables are varying in the same direction, i.e. if as one variable is increasing, the other, on an average is also increasing or, if as one variable is decreasing, the other, on an average is also decreasing, then correlation between such two variables is said to be positive (direct relationship).

Negative correlation: If both the variables are varying in opposite direction i.e. if as one variable is increasing the other, on an average is decreasing or, if as one variable is decreasing the other, on an average is increasing, then correlation between such two variables correlation is said to be negative (inverse relationship).

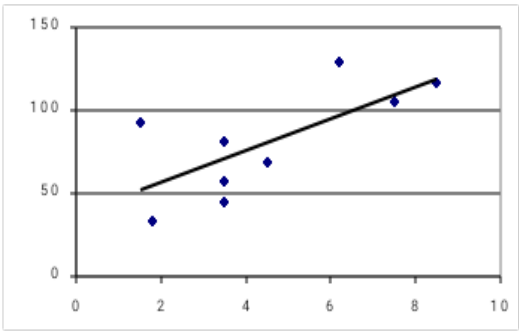
2) **Simple correlation :** The correlation between only two variables is known as simple correlation.

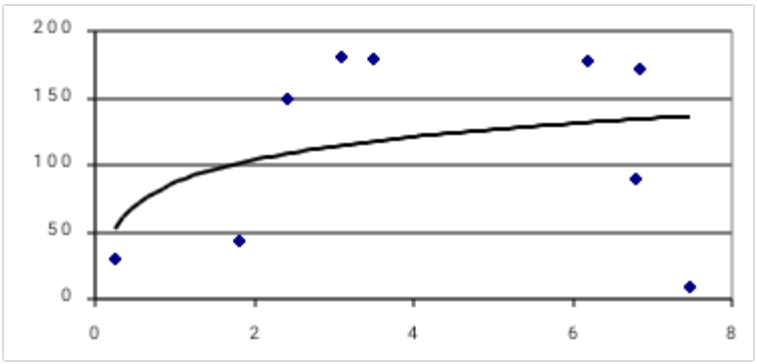
Multiple correlation : In this study three or more than three variables are included, but the correlation between a variable and set of variables is considered.

Partial correlation : In this study three or more than three variables are included but correlation between any two variables is considered by keeping statistical influence of other variables constant.

3) **Linear correlation :** If the amount of change in one variable tends the bear constant ratio to the amount of change in the other variable then the correlation is said to be linear correlation (line of equation in the form $Y= a + b X$).

Non-linear correlation : If the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable then the correlation is said to be non-linear or curvilinear
(equation in the form of $Y = a+bX+cX^2$)



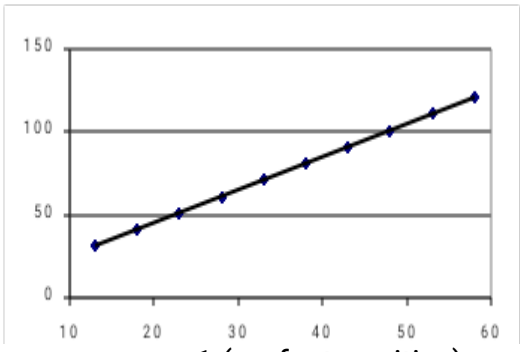


Methods of studying correlation :

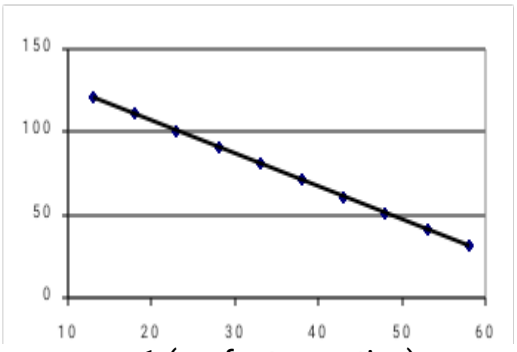
1) Scatter diagram method.
2) Graphic method
3) Karl Pearson's coefficient of correlation
4) Rank method (Spearman's rank correlation coefficient)
5) Concurrent deviation method.
6) Method of least squares.

1) **Scatter diagram method :** The simplest device to see relationship between two variables is a special type of dot chart called scatter diagram. If values of two variables are given then we can draw graph by taking one variable on X-axis and other on Y-axis on a graph paper. For each pair (X,Y) of values if we put a dot then we obtain as many points as the number of pairs of observations. Looking to the scatterness and trend of the various points/dots we come to know some idea about relationship between two variables.

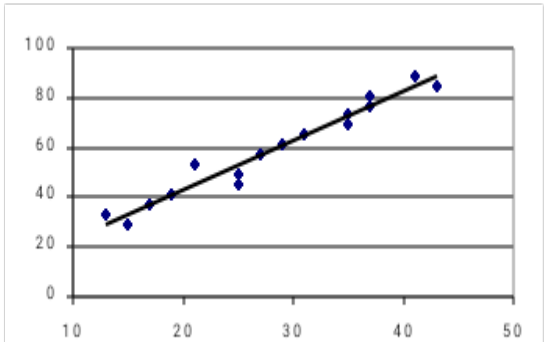
Relationship and direction of correlation.



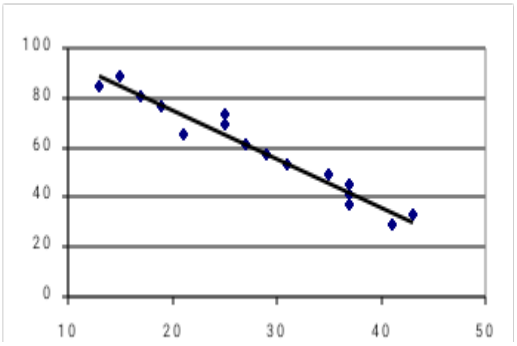
r = +1 (perfect positive)
(Points are lie on line)



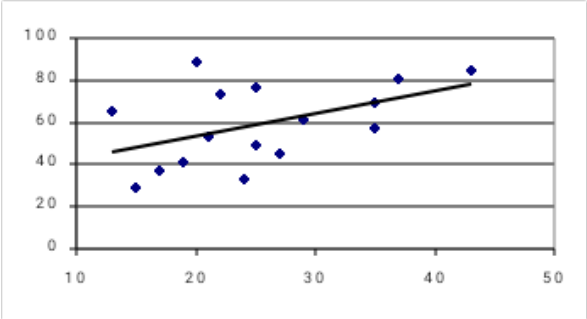
r = -1 (perfect negative)
(Points are lie on line)



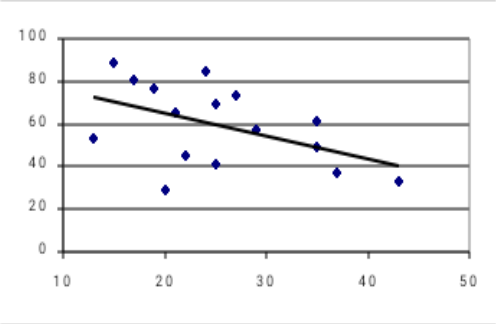
High degree positive
(Points are closely dispersed around line)



High degree negative
(Points are closely dispersed around line)



Low degree positive
Points are widely dispersed around line



Low degree negative
Points are widely dispersed around line

2) **Karl Pearson's coefficient of correlation :** This is the mathematical method of measuring correlation between two variables. The Pearson's coefficient of correlation is denoted by 'r' or r_{xy} and it gives degree/magnitude and direction of relationship between two series or two variables. The mathematical formula is

$$r_{xy} = \frac{Cov\ (X, Y)}{\sigma_X * \sigma_Y}$$

$$r_{xy} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{(Xi - \overline{X})^2}\sqrt{(Yi - \overline{Y})^2}}$$

Where, X_i = i^{th} individual observations of variable X, Y_i = i^{th} individual observations of variable Y, $\overline{X}$ = Arithmetic mean of X-series, $\overline{Y}$ = Arithmetic mean of Y-series, Cov (X, Y) = Covariance between X and y series, σ_X = Standard deviation of X-series, σ_Y = Standard deviation of Y-series

Some remarks :

1) The value of r_{xy} lies between -1 and +1 i.e. $-1 \leq r \leq 1$
2) If r =1, there exists perfect positive correlation
3) If r = -1, there exists perfect negative correlation
4) If r = 0, there does not exist correlation.
5) If r = 0.45, there exist positive correlation with magnitude of correlation is 0.45.
6) If r = - 0.46, there exist negative correlation with magnitude of correlation 0.46.

Properties:

1) It lies in between -1 to +1, symbolically $-1 \leq r \leq +1$ or $|r| \leq 1$
2) It is independent of change of scale and origin of the variable X and Y
3) $r_{xy} = r_{yx}$ (i.e. the relationship between the two variables is symmetric or equal)
4) It is geometric mean of two regression coefficients, symbolically $r_{xy} = \sqrt{b_{yx} * b_{xy}}$

Topic : SIMPLE LINEAR REGRESSION

Introduction

The correlation study is concerned with the relationship or closeness or association between variables. If two variables are closely related, then we may interested in estimating (predicting) the value of one variable for given value of another. Regression analysis reveals average relationship between variables and this makes possible to estimate or to predict. We know that there is close relationship between yield of rice and amount of rainfall received, then we may

estimate or predict yield of rice for given amount of rainfall with the help of regression analysis.

Regression analysis provides estimates of value of the dependant variable for the value(s) of independent variable(s). The device used in this estimate procedure is the regression line. The equation of this line is known as the regression equation. We may estimate value of dependent variable for given value(s) of independent variable(s) from this equation.

Definitions

1) Regression is the measure of the average relationship between two or more variables in terms of the original units of the data.
2) The terms regression analysis refers, to the methods by which estimates are made of the values of a variable from knowledge of the values of one or more other variables and to the measurement of the errors involved in this estimation process.
3) Regression analysis attempts to establish the nature of the relationship between variables i.e. to study the functional relationship between the variables and thereby provide a mechanism for prediction, or forecasting.

Regression lines

If we consider two variables say, X and Y, then we shall have two regression lines. i) Regression line of Y on X and ii) Regression line of X on Y.

The regression line of Y on X gives probable value of Y for given value of X and the regression line of X on Y gives probable value of X for given value of Y. If r = 1 or r = - 1 (i.e. perfect relationship) then both the lines will coincide at one place i.e. we will have only one line. The point of intersection of two lines gives value of arithmetic mean of X-series, and Y- series [i.e. $(\bar{X}, \bar{Y})$]. If r = 0, then two lines intersect to each other at right angle i.e. one parallel to X-axis and other parallel to Y-axis. As two intersecting lines comes nearer to each other, degree of magnitude of correlation increases.

Regression equations : Regression equation is the algebraic expression of the regression line. There are two regression lines, hence, there are two regression equations i) Regression equation of Y on X and ii) Regression equation of x on Y.

Method of normal equations :

Regression equation of Y on X : The regression equation of Y on X is expressed as

$$Y = a + b * X \quad \ldots\ldots\ldots (1)$$

Where, a, b = Numerical constant, Y=Dependent variable, X=Independent variable

a is 'intercept of line on Y-axis' and b is 'slope of line with respect to X-axis'.

To determine values of constants a and b algebraically or mathematically, we have to solve following two normal or simultaneous equations (equation 2 and 3).

$\Sigma Y_i = N * a + b * \Sigma X_i$ ..... (2) (equation obtained by taking sum of both the sides of equation 1)

$\Sigma X_i * Y_i = a * \Sigma X_i + b * \Sigma X_i^2$ .....(3) (equation obtained by multiplying equation 1 by X and then taking sum of both the sides)

Regression equation of X on Y : The regression equation of X on Y is expressed as

$$X = a + b * Y \quad \ldots\ldots\ldots (1)$$

Where, a,b = Numerical constant, X=Dependent variable, Y=Independent variable

a is 'intercept of line on X-axis' and b is 'slope of line with respect to Y-axis'.

To determine values of constants a and b algebraically or mathematically, we have to solve following two normal or simultaneous equations (equation 2 and 3).

$\Sigma X_i = N * a + b * \Sigma Y_i$ ..... (2) (equation obtained by taking sum of both the sides of equation 1)

$\Sigma X_i * Y_i = a * \Sigma Y_i + b * \Sigma Y_i^2$ ..... (3) (equation obtained by multiplying equation 1 by Y and then taking sum of both the sides)

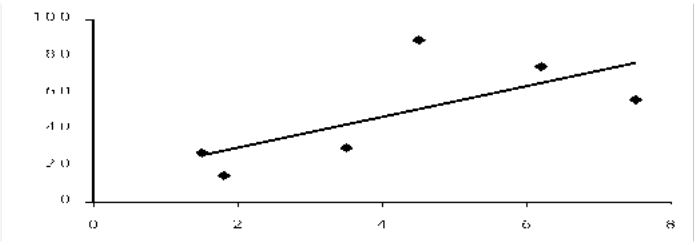
Method of scatter diagram

Regression line of Y on X : In this, scatter diagram of various points (X, Y), line is drawn passing through these scattered points then it is possible to find vertical distance of each point from line. The vertical distance of each point can be determined from line by taking difference between value of Y at a point and corresponding value of Y on the line. Value of Y on a line is also known as calculated value (Yc). According to the method of least squares, the regression line is drawn through the various points in such a manner that the sum of squares of deviations of the actual values of Y from the calculated values (i.e sum of squares of vertical distances of points from the line) is the least. i.e. $\Sigma(Y-Yc)^2$ is the least

Thus, line for which $\Sigma(Y-Yc)^2$ is the least or minimum, fits the points best and such a line is known as best fit line.

Characteristics of best fit line

1) $\Sigma(Y-Yc)^2$ is least or minimum
2) $\Sigma(Y-Yc)=0$
3) The line goes through $(\bar{X}, \bar{Y})$
4) Least square line is a best estimate of the population regression line.

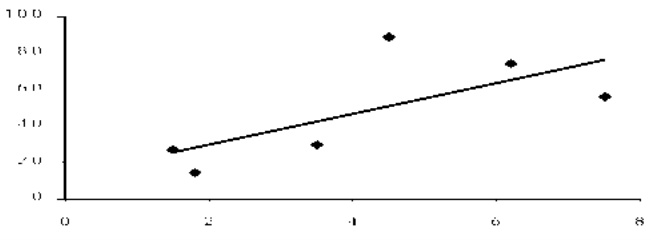


Regression line of X on Y : In this, scatter diagram of various points (X, Y), line is drawn passing through these scattered points then it is possible to find horizontal distance of each point from line. The horizontal distance of each point can be determined from line by taking difference between value of X at a point and corresponding value of X on the line. Value of X on a line is also known as calculated value (Xc). According to the method of least squares, the regression line is drawn through the various points in such a manner that the sum of squares of deviations of the actual values of X from the calculated values (i.e sum of squares of horizontal distances of points from the line) is the least. i.e. $\Sigma(X-Xc)^2$ is the least

Thus, line for which $\Sigma(X-Xc)^2$ is the least or minimum, fits the points best and such a line is known as best fit line.

Characteristics of best fit line

1) $\Sigma(X-Xc)^2$ is least or minimum
2) $\Sigma(X-Xc)=0$
3) The line goes through $(\bar{X}, \bar{Y})$
4) Least square line is a best estimate of the population regression line.



Regression equation of Y on X if deviations taken from mean of X values and Y values.

$$(Y - \overline{Y}) = r_{XY} \frac{\sigma_Y}{\sigma_X}(X - \overline{X})$$

Where, Y = Dependent variable $\overline{Y}$ = Mean of Y-series

X = Independent variable $\overline{X}$ = Mean of X-series

r_{XY} = Correlation coefficient σ_Y =Standard deviation of X-series

σ_X =Standard deviation of Y-series

The term $r_{XY}\frac{\sigma_Y}{\sigma_X}$ is known as regression coefficient of Y on X and is denoted by b_{YX}, hence

$$(Y - \overline{Y}) = b_{YX}(X - \overline{X})$$

Regression equation of X on Y if deviations taken from mean of X values and Y values.

$$(X - \overline{X}) = r_{XY} \frac{\sigma_X}{\sigma_Y}(Y - \overline{Y})$$

Where, X = Dependent variable X = Mean of X-series

Y = Independent variable Y = Mean of Y-series

r_{XY} = Correlation coefficient σ_Y =Standard deviation of X-series

σ_X =Standard deviation of Y-series

The term $r_{XY} \frac{\sigma_X}{\sigma_Y}$ is known as regression coefficient of X on Y and is denoted by b_{XY}, hence

$$(X - \overline{X}) = b_{XY}(Y - \overline{Y})$$

Some remarks

1. Both the regression coefficients have the same sign (+ve or –ve), that is both b_{YX} and by b_{XY} be negative or positive.
2. The magnitude of both the regression coefficients never exceeds 1 at a time, if one of the magnitude is greater than 1 the other must be less than 1.
3. The Coefficient of correlation have the same signs as that of regression coefficients (since $r_{XY} = \sqrt{b_{yx} * b_{xy}}$). i.e if b_{YX} and b_{XY} are negative then r_{XY} is negative and if b_{YX} and b_{XY} are positive then r_{XY} is positive.
4. Regression coefficients are independent of change of origin but not on the scale.

	Correlation	Regression
1	Correlation study is concerned with the relationship between variables.	Regression study is concerned with the average relationship between variables.
2	Correlation indicates the extent of relationship between variables.	Regression analysis helps in estimating or predicting value of dependent variable.
3	Cause and effect relationship may not be identified separately.	The cause and effect relationship is clearly indicated through regression equation.

4	The correlation coefficient are equal or symmetric i.e. $r_{yx}=r_{xy}$	The regression coefficients are different i.e. $b_{yx} \neq b_{xy}$
5	The correlation coefficient is independent of change of scale and origin.	The regression coefficients are independent of change of origin only.

Topic : POBABILITY

Introduction

The word 'probability' or 'chance' or 'possibility' is commonly used in day-to-day conversation. In layman's terminology the word probability means there is uncertainty about the happening of event. We use the word probability in the form of percent on the basis of some judgment or sense or opinion or wishful thinking. But in statistical sense, the word probability is not based on such things; it is based on collected/generated/available numerical facts/data. Hence, in statistics the term probability is established by definition and is not connected with beliefs or any form of wishful thinking.

Some terms

Experiment : The term experiment refers to describe an act which can be repeated under some given conditions.

Random experiment : Random experiment are those experiments whose results depend on chance. For example, tossing of a coin (two sides), throwing of a die (which has six sides) and drawing a card from pack of playing cards are some examples of random experiments.

Outcomes : The results of a random experiment are called outcomes or events.

Random experiment and event : If in an experiment all the possible outcomes are known in advance and none of the outcome can be predicted with certainty, then such an experiment is called a random experiment and the outcome of a random experiment is called as event or chance event. Events are denoted by capital letters A, B, C … etc.

Dependent events : Dependent events are those events in which the occurrence or non-occurrence of an event in any one trial affects the probability of other events in other trials.

Equally likely events : Events are said to be equally likely when one does not occur more often than the others.

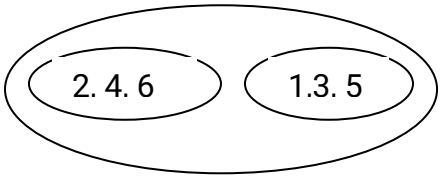
Simple and compound events : In simple events, we consider the probability of the happening or non-happening of single outcome. On the other hand, in case of compound events, we consider the joint occurrence of two or more outcomes.

Sample Space : The set of all possible outcomes of a random experiment is called as sample space of the experiment and it is denoted by S

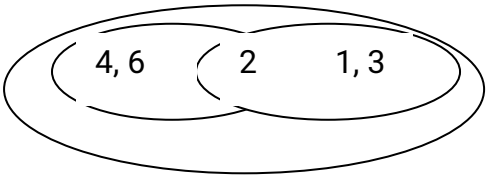
Mutually exclusive events : Two events are said to be mutually exclusive or incompatible when both cannot happen simultaneously in a single trial. For example if single coin is tossed then either head or tail can be up, both can not be up at the same time. Hence, event of getting head (say, event A) and event of getting tail (say, event B) are mutually exclusive events. In tossing a die, the event of getting even numbers (Say, A) and event of getting odd numbers (Say, B) are mutually exclusive events, but event of getting even numbers (say, A) and event of getting numbers less than or equal to 3 (Say, B) are not mutually exclusive. In this, number 2 (outcome) is

common to both the events.

Event A and B are mutually exclusive events



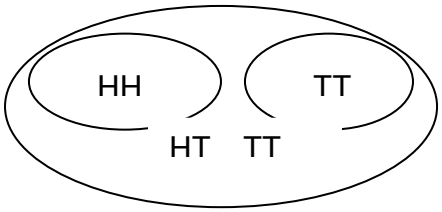
Event A and B are not mutually exclusive events



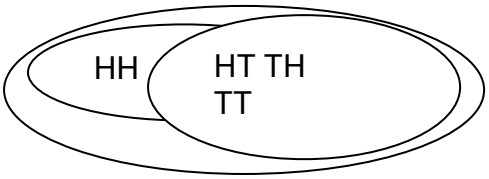
Independent events : Two or more events are said to be independents when the outcome of one does not affect, and is not affected by the other. If coin is tossed as successively two times (twice) the result of 2nd throw is not affected by the result of 1st throw. If we draw a card from a pack of cards (from 52 cards) and not placed back into the pack of cards the probability of getting a particular card at 2nd draw is affected by 1st draw. Hence events of getting particular card from pack of cards are said to be independent events if and only if we put drawn card back into the pack of cards. Otherwise it is said to be dependent events. [Note : getting black card in 1st draw (say, event A) and getting queen in 2nd draw (say, event B)]

Exhaustive events : Events are said to be exhaustive when their totally includes all the possible outcomes of a random experiments.

Sample space i.e **S**



Events A and B are not exhaustive events
(A – Getting both H B – Getting both T)



Events A and B are exhaustive events
(A - Getting at least one H B – Getting at lest one T)

Complementary events : Suppose there are two events A and B. A is called the complementary event of B, if A and B are mutually exclusive and exhaustive events.

Probability (Mathematical approach) :

Probability is defined as the ratio of number of favorable cases to the total number of equally likely cases. If probability of occurrence of event A is denoted by P(A), then by definition.

$$P(A) = \frac{\text{Number of favourable cases to event A}}{\text{Number of equally likely cases}}$$

Suppose there are '**a**' number of favourable cases to event A and there are n number of equally likely case, then

$$p = \frac{a}{n}$$

Where, p = Probability of event A

If probability of non occurrence of event A is denoted by P(not A), then by definition.

$$P(\text{not A}) = \frac{\text{Number of unfavourable cases to event A}}{\text{Number of equally likely cases}}$$

Suppose there are '**n-a'** number of unfavourable cases to event A (since '**a'** number of favourable cases to event A) and there are n number of equally likely case, then

$$q = \frac{n - a}{n}$$

$$q = \frac{n}{n} - \frac{a}{n}$$

$$q = 1 - \frac{a}{n}$$

$$q = 1 - p$$

Hence, $p + q = 1$

Where, q = Probability of not getting event A

Axiomatic theorem of probability.

Additional theorem :

If two events A and B are mutually exclusive, then the probability of the occurrence of either A or B is the sum of the individual probability of A and B symbolically.

- $P(A \text{ or } B) = P(A \cup B) = P(A) + P(B)$

If two events A and B are not mutually exclusive, then the probability of the occurrence of either A or B is by subtracting probability of occurrence of event A and event [i.e. P (A and B)] from the individual probabilities of A and B, symbolically.

- $P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(AB)$,

Proof of $P(A \text{ or } B) = P(A \cup B) = P(A) + P(B)$

P(A or B) = Probability of getting event A or event B

Suppose there are n equally likely cases, and number of favourable cases to event A is say **a** and number of favorable cases to event B is say **b**

$$P(A \text{ or } B) = \frac{\text{Number of favorable cases to A or B}}{\text{Total number of equally likely cases}}$$

$$P(A \text{ or } B) = \frac{a + b}{n}$$

$$P(A \text{ or } B) = \frac{a}{n} \quad \frac{b}{n}$$

$$P(A \text{ or } B) = \frac{\text{Number of favourable cases to event A}}{\text{Number of equally likely cases}} + \frac{\text{Number of favourable cases to event B}}{\text{Number of equally likely cases}}$$

$$P(A \text{ or } B) = P(A) + P(B)$$

Multiplication theorem

If two events A and B are independent events then the probability that they both will

occur is equal to the product of their individual probabilities, symbolically.

P(A and B) = P(A∩B) = P(A) * P(B)

(For proof of P(A and B) = P(A∩B) = P(A) * P(B) refer book)

Theoretical distributions

Binomial Distribution

The binomial distribution is also known as Bernoulli's distribution. Binomial distribution is a probability distribution expressing the probability of one set of dichotomous alternative, which is success or failure (only two possible outcomes).

For an event with probability of occurrence or success say, p and non-occurrence or failure say q, the probability distribution of number of occurrences (say, r) of A or number of successes of A in '**n**' trails follow Binomial distribution and is given by,

P (r success) = $P(r) = {}^{n}C_r\, q^{n-r}\, p^r$

Where, p = Probability of success in single trial

q = l - p = Probability of failure in single trail

n = Number of trails

r = Number of successes in n trails.

It should be noted that the variable r (as number of success) in the binomial distribution is a discrete one and not continuous.

The term obtained ${}^{n}C_r\, q^{n-r}\, p^r$ for r=0,1,2,3,..... r, r+1, ... is the term from binomial expansion of $(q+p)^n$.

$(q+p)^n = {}^{n}C_0\, q^n\, p^0 + {}^{n}C_1\, q^{n-1}\, p^1 + {}^{n}C_2\, q^{n-2}\, p^2 + {}^{n}C_3\, q^{n-3}\, p^3 + \ldots\ldots + {}^{n}C_r\, q^{n-r}\, p^r + \ldots\ldots\ldots + {}^{n}C_n\, q^{n-n}\, p^n$

[Note :

Properties of Binomial distribution

1) It is a discrete distribution
2) Mean of binomial distribution is product of number of trials and probability of success i.e Mean = n p
3) Standard deviation of binomial distribution is square root of product of number of trials and probability of success and probability of failure. i.e S.D. = $\sqrt{n\, p\, q}$

   Therefore Variance = n p q

Poisson distribution

Poisson distribution is used when the probability of success of individual event is very small i.e. p is very small. This distribution is used to describe the behaviour of rare events. For example, in the cases like number accidents on the road, number of printing mistakes and so on, in such a evidence chances of occurrence are very small, hence the probability p is also small.

The Poisson distribution is defined as

$$P(X = r) = \frac{e^{-m}\, m^r}{r!}$$

Where, r = 0, 1, 2, 3, 4 … = number of occurrences

e = 2.7183, the base of natural logarithms

m= the mean of Poisson distribution

= n p = Average number of occurrences

n = Total number of cases

p = Probability of occurrence of an event.

Properties

1) It is discrete distribution
2) Mean = n p
3) Standard deviation = √m =√n p
   Therefore variance = n p and hence, Mean = Variance

Normal distribution

The normal distribution is also called as normal probability distribution. This is a most useful theoretical distribution for continuous variable. The normal distribution is an approximation to the binomial distribution.

If **μ** is mean and **σ** is standard distribution of variable X, then the normal distribution of variable X, the normal distribution is defined as,

$$P(X = x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

The constant μ and σ are called parameters of normal distribution. Normal is denoted by $N(\mu, \sigma^2)$

If **N** is the total frequency then equation to normal curve corresponding to normal distribution with mean **μ** and standard deviation **σ** is given by

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

The quantity $\frac{N}{\sigma\sqrt{2\pi}}$ is equal to the maximum ordinate of the normal curve

If μ = 0 and σ = 1, then the normal distribution is called standard normal distribution and is given by

$$P(X = x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

Some remarks

1) The shape of normal distribution is different for different values of μ and σ, but there is unique shape of normal distribution for any given values of μ and

σ.

2) Normal distribution is limiting case of Binomial distribution (if n→ ∞ and neither p or q is very small i.e. p and q → 0.5)
3) Normal distribution is limiting case of Poisson distribution (if m is large i.e. n and p are large i.e n→ ∞ and p → 0.5)
4) The mean of normally distributed population lies at the centre of its normal curve.
5) The two tails of normal distribution never touches the x-axis
6) Poisson distribution is limiting case of Binomial distribution if n is large and p tends to o.

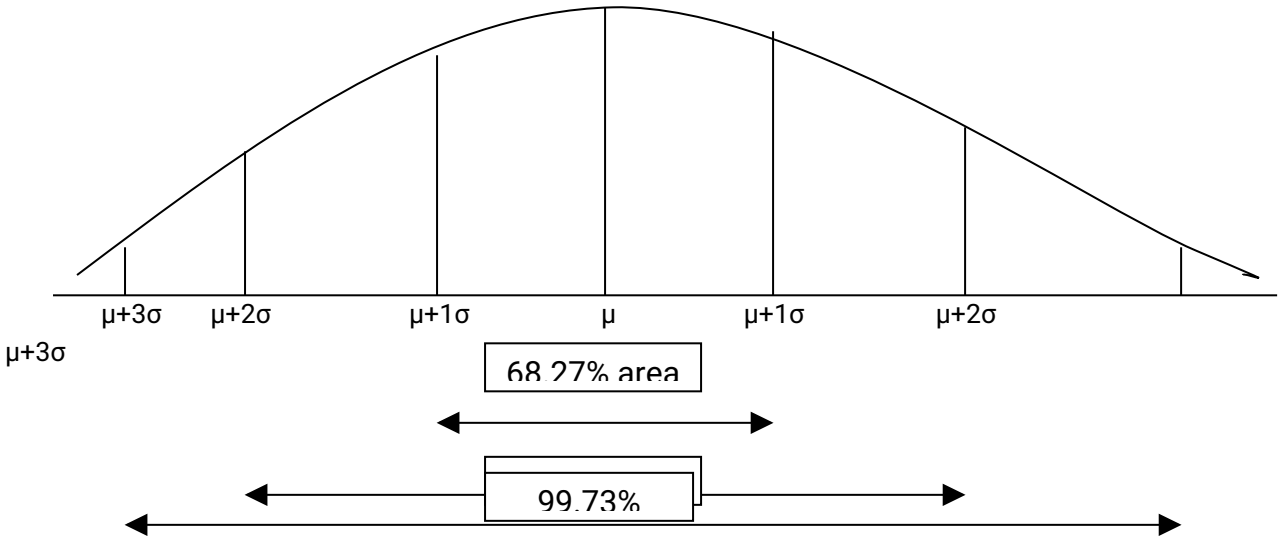
Importance

1) We can calculate maximum and minimum limits with in which population values lies (i.e distribution of total number of frequencies or total number of values N).
   e.g. Mean − 1σ to Mean + 1σ → 68.27% values
   Mean − 2σ to Mean + 2σ → 95.45% values
   Mean − 3σ to Mean + 3σ → 99.73% values
2) If population has normal distribution with mean μ and standard deviation σ and if we draw sample of size n from this population, then for large n, mean of this sample $\bar{X}$ is distributed as normal distribution with mean μ and standard deviation σ/√n
3) As sample size n becomes large, many discrete distributions follow normal distribution.

Properties of normal distribution

1) The normal curve is bell-shaped and symmetric.
2) The height of the normal curve is maximum at the mean, hence, Mean = Mode = Median
3) There exist only one maximum point of the normal curve, and which occurs at the mean. The height of normal curve declines as we go at either direction form the mean. The curve approaches nearer and nearer to base unit (i.e X-axis) but never touches it.
4) Normal curve has only one maximum point, hence normal curve has only one mode i.e. curve of normal distribution is unimodal.
5) The point of inflection i.e. point where the change in curvature occurs are x- σ and x + σ (i.e. x ± σ)
6) The variable distributed according to the normal curve is a continuous one.
7) The first and third quartiles are equidistance from the median i.e. 2nd quartile.
8) The area under the normal curve is as follows
   Mean − 1σ to Mean + 1σ (i.e. mean ±1σ)→ 68.27% area
   Mean − 2σ to Mean + 2σ (i.e. mean ±2σ)→ 95.45% area
   Mean − 3σ to Mean + 3σ (i.e. mean ±3σ)→ 99.73% area
9) For standard normal distribution mean = 0 and S.D. = 1 Hence above property for standard normal distribution becomes.
   -1 to +1 → 68.27% area
   -1 to +1 →95.45% area
   -1 to +1 →99.73% area

Area under the curve



Topic : Statistical Inference - Test of Hypothesis

Some terms

Population : A population (or Universe) is totality of items or things under consideration. It is a collection of all measurements of a particular type of interest to the decision-makers. Population may be finite or infinite.

Sample : A sample is any group of measurements selected from a population for analysis or study. It is also known as subset of population.

Elementary units : The individual items in a population are called elementary units.

Parameter : Population constant like measures of central tendency of population (e.g. arithmetic mean), measures of dispersion of population are known as parameters (e.g. standard deviation) .

Statistic : Sample constant like measures of central tendency of sample (e.g. arithmetic mean), measures of dispersion of sample are known as statistic (e.g. standard deviation).

Estimation : To use 'statistic' obtained from the sample as estimate of the unknown parameters of the population.

Hypothesis : Hypothesis is an assumption that is made about a population parameter.

Hypothesis testing : To test hypothesis about parent population from which the sample is drawn.

Introduction

Statistical inference is concerned with solving problems related with uncertainty in decision making by using probability concept. Statistical inference refers to process of selecting and using 'sample statistic' to draw conclusion about a population parameters.

Procedure of testing hypothesis

1) **Set up a hypothesis :** The first thing in hypothesis testing is to set up a hypothesis and this hypothesis can be tested on the basis of information generated from the sample.
   There are two types of hypothesis
   a) **Null hypothesis :** It is an assumption that is made about the population

parameters in the form of statement of equality.

For example, H_0: $\mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$ or $\mu_1 = \mu_2 = \mu$

i.e. H_0 : There is no difference between means of two populations or difference between two means is equal to zero or both means are equal to μ.

b) **Alternative Hypothesis** : Any hypothesis which differs from null hypothesis is called alternative hypothesis

e.g H_1: $\mu_1 \neq \mu_2$, H_1: $\sigma_1^2 \neq \sigma_2^2$, H_1: $\mu_1 > \mu_2$, H_1: $\mu_1 < \mu_2$ and so on

2) **Set up a suitable significance level :** The significance level is always expressed as a percentage like 10%, 5%, 2%, 1%. The significance level is nothing but probability of rejecting null hypothesis if it is true.

3) **Setting a test criteria :** Setting a test criteria means selecting any one appropriate statistical test procedure such as t-test, F-test or χ^2.

4) **Doing computations.**

5) **Making decisions :** A statistical conclusion or statistical decision is a decision either to reject or to accept the null hypothesis.

Two types of Errors

There are two types of errors in testing hypothesis. A statistical hypothesis is tested by applying certain test criteria to take a decision regarding rejection or acceptance of the null hypothesis. There are four possibilities while taking such decisions.

a) The hypothesis is true but the test rejects it, is known as type I error.
b) The hypothesis is false but the test accepts it, is known as type II error.
c) The hypothesis is true and test accepts it, is the correct decision.
d) The hypothesis is false and the test rejects it, is also correct decision.

These situations & decisions are summarized as follows

		Decision	
		H_0 is accepted	H_0 is rejected
Situation	H_0 is true	Correct decision	**Type I error**
	H_1 is false	**Type II error**	Correct decision.

α = Prob (Type I error) = Prob (Reject Ho if it is true)

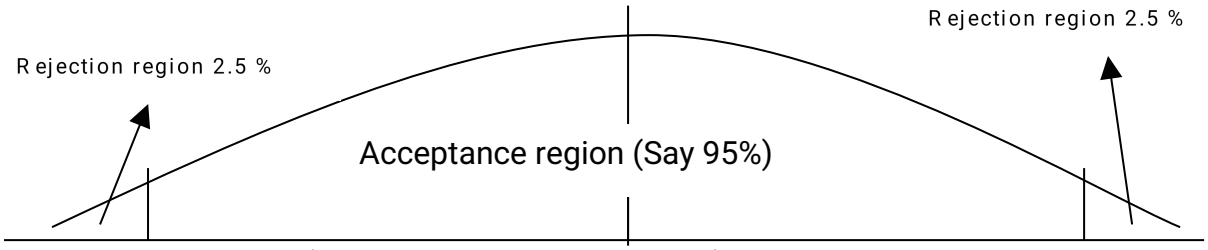
β = Prob (Type II error) = Prob (Accept Ho if it is false)

(Note : While testing hypothesis our aim is to reduce both the type of errors, but it is not possible because as probability of making one type of error can be reduced if we are willing to increase the probability of making the other type of error (as α increases β must be decreases since α + β =1)

To accept false hypothesis is equivalent to accept lot of bad items. Hence, it is more dangerous to accept the false hypothesis (Type II error) than to reject a true hypothesis (Type I error). Hence we keep Type I error at a certain level. Level of significance is the probability of committing Type I error. (The level of significance is also known as size of rejection region or size of the critical region or size of the test).

a) 5% level of significance means (α = 0.05) means out of 100 decisions made, only 5 wrong decision are tolerable and we have 95% confidence that the decision made is correct.

b) 1% level of significance means (α = 0.01)

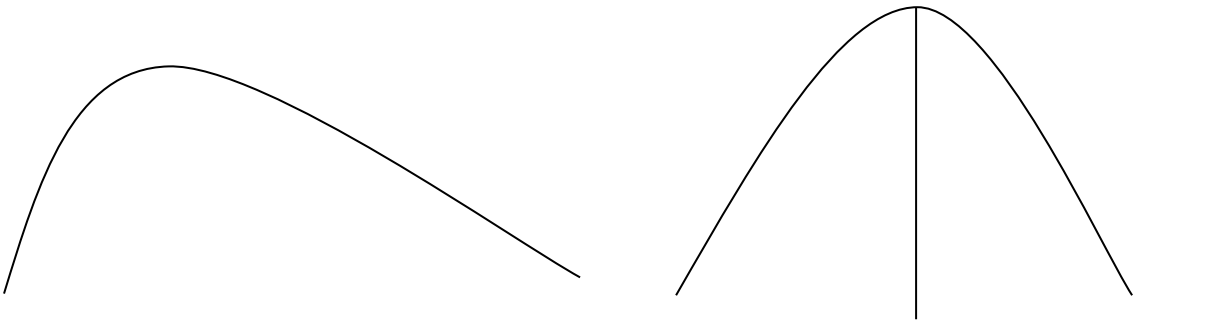


Note : $\propto$ = 0.05 (i.e significance level =5%)

Sampling distribution

It is not possible to study the complete population because some populations are infinite, some populations are finite but large, some populations are finite but it is not possible to study each and every unit of it for various reasons (like, time limit, economy, experienced manpower, destruction of units at the time of study and so on). So, to study ragarding population, we have to select a random sample from the population. The value of statistic obtained form sample is considered as estimate of the population parameter. **These two values, sample statistic and population parameter may not be equal, but difference between these two occurs due to sampling error.**

Consider population having mean μ (population parameter) and standard deviation σ (population parameter). Suppose we have taken 100 random samples each of size say n viz. S_1, S_2, S_3 … S_{100} from population. Form these 100 samples we can calculate 100 sample means (statistic). It will be observed that even though all the samples are of equal size and drawn from the same population, the values of sample means vary from sample to sample. If we draw all possible sample of size n instead of 100 samples from population, then variation of mean of different samples can studied by constructing a frequency distribution table of all means. (means of all the possible samples of size n). Such a frequency distribution of the statistic (sample mean) is called as “sampling distribution” of the statistic.



Sample mean distribution

Mean = μ Standard deviation = σ/√n Mean = μ Standard deviation = σ/√n The arithmetic mean of population and sample distribution remains same.

1) The standard deviation of sampling distribution is equal to ratio of **population standard deviation** and **the square root of the sample size**. (Standard deviation of sample mean distribution (i.e distribution of $\bar{X}$) is also called standard error of sample mean distribution (i.e distribution of $\bar{X}$)).
2) Even if population is not distributed normally, the sampling distribution is

distributed normally.

Students t-distribution : This test is used when sample size is lees than 30. If sample size is less than 30 then such sample is known as small sample. In this, we are not interested to **estimate** population parameter with the help of such a small sample drawn from a population, only our interest to test the hypothesis i.e to test **difference** arises between **observed value** (calculated from sample) and **population value** is due to **sampling fluctuation**. While using this test we assume that the parent population is normally distributed.

Theoretical work on t-distribution was done by W. S. Gosset. He was known by his pen name 'Student', and his work was published by his pen name 'student'. Hence t-distribution is commonly called as student's t-distribution or Students distribution. The t-statistic (since t is derived from sample) is defined as

$$t = \frac{\overline{X} - \mu}{S} * \sqrt{n}$$

$$S = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n-1}}$$

And its distribution is

$$f(t) = c\left(1 + \frac{t}{\gamma}\right)^{-\frac{(\gamma+1)}{2}}$$

$$f(t) = c\left(1 + \frac{t}{\gamma}\right)^{-(\gamma+1)/2}$$

Where,

$$t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

$$t = \frac{\bar{X} - \mu}{S}\sqrt{n}$$

c = a constant required to make area under the curve equal to one

γ = n – 1 (degrees of freedom)

Properties : i) t-value ranges from $-\infty$ to $+\infty$

ii) Value of c depends on γ

iii) Distribution of t is symmetric and mean equal to 0.

iv) Value of standard deviation is always greater than 1 and it tends to 1 as γ tends to 30 or as γ tends to infinity.

i.e. as γ tends to ∞ (or large)

Mean = 0 and Standard deviation = 1

Hence as γ tends to ∞ (or large) t-distribution becomes standard normal distribution.

Application of t-distribution :

I) To test the significance of the mean of a random sample (Single sample)

In deciding, weather the mean of a sample drawn from a normal population deviates significantly or significantly different from population mean, when variance of the population is unknown, we calculate the **statistic**

$$t = \frac{\overline{X} - \mu}{S} * \sqrt{n}$$

$$S = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n-1}}$$

Where, X = Mean

n = Sample size

S = Standard deviation of sample

Ho : $\mu_o = \mu$ i.e Ho : Population mean = μ

If calculated | t | > table value t, then we can say difference between mean of sample and mean of population is significant i.e. sample is not drawn from population having mean μ . If calculated | t | < table value t, then we can say difference between mean of sample and mean of population is not significant i.e. sample is drawn from population having mean μ .

II) To test the significance of the means of a two random samples (Independent, Single sample)

In deciding, weather two independent samples having means μ_1 and μ_1 and standard deviation S_1 and S_2 respectively come (or drawn) from a same normal population, we calculate the **statistic**

$$t = \frac{\overline{X} - \overline{Y}}{S.E.difference} * \sqrt{\frac{n * m}{n+m}}$$

$$S.E.difference = \sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}$$

Where, $\overline{X}$ = Mean of 1st sample

$\overline{Y}$ = Mean of 2nd sample

n = Size of 1st sample

m = Size of 2nd sample

Sx = Standard deviation of 1st sample

Sy = Standard deviation of 2nd sample

Ho : $\mu_x = \mu_y$ i.e Ho : Mean of 1st population = Mean of 2nd population

If calculated | t | > table value t, then we can say difference between mean of sample and mean of population is significant i.e. sample is not drawn from population having mean μ . If calculated | t | < table value t, then we can say difference between mean of sample and mean of population is not significant i.e. sample is drawn from population having mean μ .

III) To test the significance of the means of a two random samples (Dependent, Single sample, matched pair observations or related samples)

In previous test it was assumed that the two samples were independent. Two samples are said to be dependent when the measurements of one sample are related to those in other in any significant or meaningfull manner. In this, samples may consist of pairs of observations of two characters of same object or individual or selected population elements.

In deciding, weather two samples are related to each other (i.e both the samples are drawn from same population), we calculate the **'statistic'**

$$t = \frac{\bar{d}}{S}\sqrt{n}$$

Where, $\bar{d}$ = the mean of difference of pairs of values

S = the standard deviation of difference values

$$\bar{d} = \frac{\sum d_i}{n} \qquad S = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}}$$

Ho : $\mu_d = 0$ i.e Ho : Population mean of difference values = 0

If calculated | t | > table value t, then we can say two samples are not related or samples is not drawn from two populations. If calculated | t | < table value t, then we can say two samples are related or samples are drawn from same population

IV) To test the significance an observed correlation coefficient :

Suppose correlation coefficient between two variables/characters on the basis of samples is say r_{xy} (ie. calculated or observed correlation form two sets of values). With the help of t-test it is possible to test the hypothesis that the correlation coefficient (ρ) between those two variable/characters in population is equal to zero. i.e these two variables/characters are uncorrelated. The test is

$$t = \frac{r_{xy}}{1 - r_{xy}}\sqrt{n-1}$$

Ho : $\rho = 0$ i.e Ho : Population correlation = 0

If calculated | t | > table value t, then we can say correlation coefficient of two variables is different from **zero** i.e they are correlated variables. If calculated | t | < table value t, then we can say correlation coefficient of two variables is equal to i.e they are uncorrelated variables.

F-test or **variance ratio test :**

The object of F-test is to find out weather the two independent estimates of populaton variance differ significantly or weather the two samples may be regarded as drawn from the normal population having the same variance. To test two variances the F ratio is defined as

$$F = \frac{S_x^2}{S_y^2}$$

Where, $S_x^2 = \frac{\sum (X_i - \bar{X})^2}{N_1 - 1}$

$$S_x^2 = \frac{\sum (Y_i - \bar{Y})^2}{N_2 - 1}$$

(Note : $\nu_1 = n_1 - 1 \quad \nu_2 = n_2 - 1$)

if $S_x^2 > S_y^2$ and table value is taken against ν_1 , ν_2 d.f.

$$F = \frac{S_y^2}{S_x^2}$$

if $S_y^2 > S_x^2$ and table value is taken against ν_2 , ν_1 d.f.

Since F- test is based on the ratio of two variance, it is also known as the variance ratio test