KAMARAJ COLLEGE

SELF FINANCING COURSES

(Reaccreditedwith “A+” Grade by NAAC)

(Affiliated to Manonmaniam Sundaranar University, Tirunelveli.)

STUDY MATERIAL FOR B.Sc., MICROBIOLOGY

BIOSTATISTICS

VI - SEMESTER

ACADEMIC YEAR 2022-23

PREPARED

BY,

DEPARTMENT OF MICROBIOLOGY (SF)

KAMARAJ COLLEGE,

THOOTHUKUDI.

BIOSTATISTICS

Unit – I

Introduction to Biostatistics – Definition, statistical methods, biological measurement, kinds of biological data, functions of statistics and limitation of statistics.

Unit – II

Collection of data, sampling and sampling design, classification and tabulation, types of representations, graphic – bar diagrams, pie diagrams and curves

Unit – III

Measures of central tendency, mean, median, mode, geometric mean

Unit – IV

Measures of dispersion and variability, changes. Deviations –mean deviation, standard deviation, coefficient of variation, Standard error, Skewness – Karl Pearson’s and Bowley’s coefficient of Skewness, Kurtosis.

Unit -V

Correlation Analysis – Scatter diagram – Karl Pearson’s Correlation Coefficient, Regression analysis –Test of significance – ANOVA (one way).

UNIT - I
INTRODUCTION TO BIOSTATISTICS

Definition

Biostatistics is the study of statistics as applied to biological areas. Biological laboratory experiments, medical research (including clinical research), and health services research all use statistical methods. Many other biological disciplines rely on statistical methodology.

Biostatistics is the science of collection, analysis and interpretation of facts and numbers connected with biology. Biostatistics is also called biometrics. Biometrics refers to biological measurements.

STATISTICAL METHODS

The five basic methods are **mean, standard deviation, regression, hypothesis testing, and sample size determination**.

A) Mean

It is one of the simplest and most popular analysis methods easy to apply to data. The mean is the average value of data used in research. In statistics, the term “mean” is commonly used to indicate average. It is calculated by adding the data values and dividing them by the total number of data points. Though it is a common method, it is advised to have other methods supporting it for effective decision-making.

B) Standard Deviation

Standard deviation is a common statistical analysis tool to determine the deviation of a set of values from the mean value. The standard deviation value will be low if the deviation from the mean is small and vice versa.

C) Regression

The regression method helps comprehend the relationship between two or more variables used in the analysis. It shows how one variable is dependent on the other and their inter effect on each other. There is simple linear regression using a single independent variable to interpret the dependent variable and multiple linear regressions using multiple independent variables to interpret the outcome.

D) Hypothesis Testing

The method tests the validity and authenticity of a hypothesis, outcome, or argument. Hypothesis testing is an assumption set at the beginning of the research; after the test is over and a result is obtained based on it, the belief can be either true or false. In addition, it can check whether the null hypothesis or alternative hypothesis is true.

E) Sample Size Determination

The technique derives a sample from the entire population, representing the total population. When there is a large data set, and the analysis gets challenging, a small sample is taken for study and research.

BIOLOGICAL MEASUREMENT

Biostatistics is also called biometrics. Biometrics refers to biological measurements. A simple example is the estimation of oxygen in a few samples. The water samples form a population. The estimation of O2 in each water sample is the collection of data. The amount of oxygen in the water is the data. Arranging values in columns is called tabulation. In one water sample, the amount of oxygen will be higher and in another is lower. It is interpretation.

A) Population:

Population is a group of individuals or study elements or observations. In the estimation of oxygen, all the water samples form a population. The value of each sample is a variable. The population containing limited number of individuals is called a finite population. Eg: Number of students in a class; Number of coconut trees in a grove.

The population containing unlimited number of individuals is called an **infinite population.** Eg: Stars in the Sky; Fishes in the sea.

A) Data:

The values recorded in an experiment or observations are called data. There are two types such as,

1. Primary data- The data collected by investigator. It is the first hand information. The person collecting the data is called investigator.
2. Secondary data- The data collected from another source. **Eg**: Data collected from Newspaper, Journals, etc.,

B) Samples:

A small representative fraction of a population is called a sample. Getting a sample from a population is called sampling. Eg: Only a few rice is examined from a boiling pot to arrive at a conclusion.

KINDS OF BIOLOGICAL DATA

Biostatistics is the science which deals with development and application of the most appropriate methods for the: Collection of data. Presentation of the collected data. Analysis and interpretation of the results. Making decisions on the basis of such analysis.

A) Qualitative variable:

It is a variable or characteristic which cannot be measured in quantitative form but can only be identified by name or categories, for instance place of birth, ethnic group, type of drug, stages of breast cancer (I, II, III, or IV), degree of pain (minimal, moderate, severe or unbearable).

B) Quantitative variable:

A quantitative variable is one that can be measured and expressed numerically and it can be either discrete or continuous. The values of a discrete variable are usually whole numbers, such as the number of episodes of diarrhea or number of children. A continuous variable is a measurement on a continuous scale. Examples include weight, height, blood pressure, age, etc. Although the types of variables could be broadly divided into categorical (qualitative) and numerical (quantitative), it has been a common practice to see four basic types of data (scales of measurement).

- **Nominal data** represent categories or names. There is no implied order to the categories of nominal data. In these types of data, individuals are simply placed in the proper category or group, and the number in each category is counted. Each item must fit into exactly one category.

- **Ordinal Data** have order among the response classifications (categories). The spaces or intervals between the categories are not necessarily equal. For example: strongly agree, agree, no opinion, disagree, strongly disagree. In this situation, we only know that the data are ordered.
- **Interval Data**: In interval data the intervals between values are the same. For example, in the Fahrenheit temperature scale, the difference between 70 degrees and 71 degrees is the same as the difference between 32 and 33 degrees. But the scale is not a RATIO Scale. 40 degrees Fahrenheit is not twice as much as 20 degrees Fahrenheit.
- **Ratio Data:** The data values in ratio data do have meaningful ratios, for example, age is a ratio data, and someone is 40 is twice as old as someone is 20.

Both interval and ratio data involve measurement. Most data analysis techniques that apply to ratio data also apply to interval data. Therefore, in most practical aspects, these types of data (interval and ratio) are grouped under metric data. In some other instances, these type of data are also known as numerical discrete and numerical continuous.

C) Numerical discrete:

Numerical discrete data occur when the observations are integers that correspond with a count of some sort. Some common examples are:

- the number of bacteria colonies on a plate,
- the number of cells within a prescribed area upon microscopic examination
- the number of heart beats within a specified time interval,
- a mother's history of number of births (parity) and pregnancies (gravidity), etc.
- **Numerical continuous:**

The scale with the greatest degree of quantification is a numerical continuous scale. Each observation theoretically falls somewhere along a continuum. One is not restricted, in principle, to particular values such as the integers of the discrete scale. The restricting factor is the degree of accuracy of the measuring instrument. Most clinical measurements, such as blood pressure, serum cholesterol level, height, weight, age etc. are on a numerical continuous scale.

FUNCTIONS OF BIOSTATISTICS

1. To Understand and Assess Medical Literature

Measures of statistics such as variability and central tendency are often used in the medical literature. Errors do occur in data that is published in medical research, sometimes even in the well-respected textbooks. Possessing the necessary knowledge will help you critically evaluate and apply original research data.

2. To Reap the Professional Benefits

Understanding the most frequently used and the crucial descriptive and inferential bio statistical methods will help appreciate how the application of the theories of measurement, statistical inference, and decision trees contribute to improved clinical decisions and eventually to improved patient care and outcomes.

3. To Effectively Collaborate with Statisticians

Having an adequate understanding of the vocabulary and fundamental concepts of biostatics will help in a fruitful collaboration with biostatisticians. A strong statistical practice is important in many medical research projects. Also, medical thinking is crucial to the formulation and application of statistical strategies.

4. To Discover the Patterns Obscured by the Variability of Responses in Living Systems

Statistics provides the tools to make an appropriate choice by judging the "significance" of the observed differences or changes.

5. To Establish that a Test Therapeutic Product is Safe and Effective

In a pivotal clinical trial, to demonstrate that a therapeutic product is safe and effective, a sample of the population is treated. Then a statistical inference is used to determine the safety and efficacy of the product

6. To Uphold the Integrity of Clinical Trial

Understanding the key role in the drug development process right from trial design to protocol development will help in protecting the integrity of the clinical trial.

7. It Is a Decision-Making Tool

Statistics is a useful decision-making tool in the clinical research arena. When working in a field where a p-value can determine the next steps on the development of a drug or procedure, decision-makers must understand the theory and application of statistics.

8. To Freely Communicate Statistical and Epidemiological Information to Patients and Colleagues

Given that many are vulnerable to the suggestions of published literature and other questionable sources, it is imperative to keep abreast of medical knowledge and communicate with confidence.

9. To Avoid Mis-Interpretation of Statistical Methods

A Harvard report on clinical research suggests that researchers often misinterpret statistical methods due to poor knowledge of statistical concepts. A clinical research professional will do well to understand statistical concepts such as confidence Intervals, Multiplicity, Subgroup Analysis, Parametric vs. Non-parametric statistical methods, Sample Size Calculation, Types of endpoints, Statistical Reporting, Missing Data, Adaptive Trial Design, and Bayesian Model.

10.To Properly Use the Statistical Tools

Much statistical software is now available to professionals.

LIMITATIONS OF BIOSTATISTICS

1. Statistics is not concerned with individual observation.
2. Statistics do not analyse qualitative phenomenon.

3. Statistical generalisations are true only on average.
4. Improper use of statistics.

UNIT 2 -

COLLECTION OF DATA

In Statistics, data collection is a process of gathering information from all the relevant sources to find a solution to the research problem. It helps to evaluate the outcome of the problem. The data collection methods allow a person to conclude an answer to the relevant question. Most of the organizations use data collection methods to make assumptions about future probabilities and trends. Once the data is collected, it is necessary to undergo the data organization process.

The main sources of the data collections methods are “Data”. Data can be classified into two types, namely primary data and secondary data. The primary importance of data collection in any research or business process is that it helps to determine many important things about the company, particularly the performance. So, the data collection process plays an important role in all the streams. Depending on the type of data, the data collection method is divided into two categories namely,

- Primary Data Collection methods
- Secondary Data Collection methods

1. Primary Data Collection Methods:

Primary data or raw data is a type of information that is obtained directly from the first-hand source through experiments, surveys or observations. The primary data collection method is further classified into two types. They are

- Quantitative Data Collection Methods
- Qualitative Data Collection Methods

i) Quantitative Data Collection Methods

It is based on mathematical calculations using various formats like close-ended questions, correlation and regression methods, mean, median or mode measures. This method is cheaper than qualitative data collection methods and it can be applied in a short duration of time.

ii) Qualitative Data Collection Methods

It does not involve any mathematical calculations. This method is closely associated with elements that are not quantifiable. This qualitative data collection method includes interviews, questionnaires, observations, case studies, etc. There are several methods to collect this type of data. They are

iii) Observation Method

Observation method is used when the study relates to behavioural science. This method is planned systematically. It is subject to many controls and checks. The different types of observations are:

- Structured and unstructured observation
- Controlled and uncontrolled observation
- Participant, non-participant and disguised observation

iv) Interview Method

The method of collecting data in terms of verbal responses. It is achieved in two ways, such as

- Personal Interview – In this method, a person known as an interviewer is required to ask questions face to face to the other person. The personal interview can be structured or unstructured, direct investigation, focused conversation, etc.
- Telephonic Interview – In this method, an interviewer obtains information by contacting people on the telephone to ask the questions or views, verbally.

v) Questionnaire Method

In this method, the set of questions are mailed to the respondent. They should read, reply and subsequently return the questionnaire. The questions are printed in the definite order on the form. A good survey should have the following features:

- Short and simple
- Should follow a logical sequence
- Provide adequate space for answers
- Avoid technical terms
- Should have good physical appearance such as colour, quality of the paper to attract the attention of the respondent

2. Secondary data collection method

Secondary data is data collected by someone other than the actual user. It means that the information is already available, and someone analyses it. The secondary data includes magazines, newspapers, books, journals, etc. It may be either published data or unpublished data.

Published data are available in various resources including

- Government publications
- Public records
- Historical and statistical documents
- Business documents
- Technical and trade journals

Unpublished data includes

- Diaries
- Letters
- Unpublished biographies, etc

SAMPLING DESIGN

A sample that is representative of the group as a whole is called a **sampling method**. There are two primary types of sampling methods that you can use in your research:

- **Probability sampling** involves random selection, allowing you to make strong statistical inferences about the whole group.
- **Non-probability sampling** involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

Population vs. sample

- The **population** is the entire group that you want to draw conclusions about.
- The **sample** is the specific group of individuals that you will collect data from.

The population can be defined in terms of geographical location, age, income, or many other characteristics.

It can be very broad or quite narrow: maybe you want to make inferences about the whole adult population of your country; maybe your research focuses on customers of a certain company, patients with a specific health condition, or students in a single school.

It is important to carefully define your target population according to the purpose and practicalities of your project.

If the population is very large, demographically mixed, and geographically dispersed, it might be difficult to gain access to a representative sample. A lack of a representative sample affects the validity of your results, and can lead to several research biases, particularly sampling bias.

Sampling frame

The sampling frame is the actual list of individuals that the sample will be drawn from. Ideally, it should include the entire target population (and nobody who is not part of that population).

Example: Sampling frame you are doing research on working conditions at a social media marketing company. Your population is all 1000 employees of the company. Your sampling frame is the company's HR database, which lists the names and contact details of every employee.

Sample size

The number of individuals you should include in your sample depends on various factors, including the size and variability of the population and your research design. There are different sample size calculators and formulas depending on what you want to achieve with statistical analysis.

A) Probability sampling methods

Probability sampling means that every member of the population has a chance of being selected. It is mainly used in quantitative research. If you want to produce results that are representative of the whole population, probability sampling techniques are the most valid choice.

There are four main types of probability sample.

1. Simple random sampling

In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include the whole population.

To conduct this type of sampling, you can use tools like random number generators or other techniques that are based entirely on chance.

2. Systematic sampling

Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct. Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.

If you use this technique, it is important to make sure that there is no hidden pattern in the list that might skew the sample. For example, if the HR database groups employees by team, and team members are listed in order of seniority, there is a risk that your interval might skip over people in junior roles, resulting in a sample that is skewed towards senior employees.

3. Stratified sampling

Stratified sampling involves dividing the population into subpopulations that may differ in important ways. It allows you draw more precise conclusions by ensuring that every subgroup is properly represented in the sample.

To use this sampling method, you divide the population into subgroups (called strata) based on the relevant characteristic (e.g., gender identity, age range, income bracket, job role).

Based on the overall proportions of the population, you calculate how many people should be sampled from each subgroup. Then you use random or systematic sampling to select a sample from each subgroup.

4. Cluster sampling

` Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups.

If it is practically possible, you might include every individual from each sampled cluster. If the clusters themselves are large, you can also sample individuals from within each cluster using one of the techniques above. This is called multistage sampling.

This method is good for dealing with large and dispersed populations, but there is more risk of error in the sample, as there could be substantial differences between clusters. It’s difficult to guarantee that the sampled clusters are really representative of the whole population.

B) Non-probability sampling methods

In a non-probability sample, individuals are selected based on non-random criteria, and not every individual has a chance of being included.

This type of sample is easier and cheaper to access, but it has a higher risk of sampling bias. That means the inferences you can make about the population are weaker than with probability samples, and your conclusions may be more limited. If you use a non-probability sample, you should still aim to make it as representative of the population as possible.

Non-probability sampling techniques are often used in exploratory and qualitative research. In these types of research, the aim is not to test a hypothesis about a broad population, but to develop an initial understanding of a small or under-researched population.

1) **Convenience sampling**

A convenience sample simply includes the individuals who happen to be most accessible to the researcher.

This is an easy and inexpensive way to gather initial data, but there is no way to tell if the sample is representative of the population, so it can't produce generalizable results. Convenience samples are at risk for both sampling bias and selection bias.

2) **Voluntary response sampling**

Similar to a convenience sample, a voluntary response sample is mainly based on ease of access. Instead of the researcher choosing participants and directly contacting them, people volunteer themselves (e.g. by responding to a public online survey).

Voluntary response samples are always at least somewhat biased, as some people will inherently be more likely to volunteer than others, leading to self-selection bias.

3) **Purposive sampling**

This type of sampling, also known as judgement sampling, involves the researcher using their expertise to select a sample that is most useful to the purposes of the research.

It is often used in qualitative research, where the researcher wants to gain detailed knowledge about a specific phenomenon rather than make statistical inferences, or where the population is very small and specific. An effective purposive sample must have clear criteria and rationale for inclusion. Always make sure to describe your inclusion and exclusion criteria and beware of observer bias affecting your arguments.

4) **Snowball sampling**

If the population is hard to access, snowball sampling can be used to recruit participants via other participants. The number of people you have access to "snowballs" as you get in contact with more people. The downside

here is also representativeness, as you have no way of knowing how representative your sample is due to the reliance on participants recruiting others. This can lead to sampling bias.

CLASSIFICATION AND TABULATION

Classification:

The collected data, also known as raw data or ungrouped data are always in a unorganised form and need to be organised and presented in meaningful and readily comprehensible form in order to facilitate further statistical analysis. It is, therefore, essential for an investigator to condense a mass of data into more and more comprehensible and assimilable form. The process of grouping into

different classes or sub classes according to some characteristics is known as classification, tabulation is concerned with the systematic arrangement and presentation of classified data. Thus classification is the first step in tabulation. For Example, letters in the post office are classified according to their destinations viz., Delhi, Madurai, Bangalore, Mumbai etc.,

Objects of Classification:

The following are main objectives of classifying the data:

- It condenses the mass of data in an easily assimilable form.
- It eliminates unnecessary details.
- It facilitates comparison and highlights the significant aspect of data.
- It enables one to get a mental picture of the information and helps in drawing inferences.
- It helps in the statistical treatment of the information collected.

Types of classification:

Statistical data are classified in respect of their characteristics. Broadly there are four basic types of classification namely

a) Chronological classification

b) Geographical classification
c) Qualitative classification
d) Quantitative classification

a) Chronological classification:

In chronological classification the collected data are arranged according to the order of time expressed in years, months, weeks, etc., The data is generally classified in ascending order of time. Example: The estimates of birth rates in India during 1970 – 74 are,

Year	1970	1971	1972	1973	1974
Birth Rate	36.8	36.9	36.6	34.6	34.5

b) Geographical classification:

In this type of classification the data are classified according to geographical region or place. For instance, the production of paddy in different states in Iraq, production of wheat in different countries etc.,

Example:

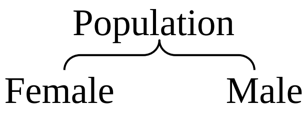
Country	America	China	Denmark	France	Iraq
Yield of wheat in (kg/acre)	1925	893	225	439	862

c) Qualitative classification:

In this type of classification data are classified on the basis of same attributes or quality like sex, literacy, religion, employment etc., Such attributes cannot be measured along with a scale. For example, if the population to be classified in respect to one attribute, say sex, then we can classify them into two namely that of males and females. Similarly, they can also be classified into

'married or 'single' on the basis of another attribute 'marital status'. Thus when the classification is done with respect to one attribute, which is dichotomous in nature, two classes are formed, one possessing the attribute and the other not possessing the attribute. This type of classification is called simple or dichotomous classification.

A simple classification may be shown as under



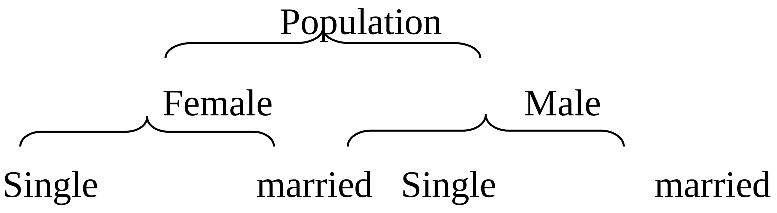
The classification, where two or more attributes are considered and several classes are formed, is called a manifold classification. For example, if we classify population simultaneously with respect to two attributes, e.g sex and marital status, then population are first classified with respect to 'sex' into 'males' and 'females'. Each of these classes may then be further classified into 'married' and single on the basis of attribute 'employment' and as such Population are classified into four classes namely.

I. Male married

II. Male single

III. Female married

IV. Female single Still the classification may be further extended by considering other attributes like marital status etc. This can be explained by the following chart



d) Quantitative classification:

Quantitative classification refers to the classification of data according to some characteristics that can be measured such as height, weight, etc., For example the group of children may be classified according to weight as given below.

In this type of classification there are two elements, namely

I. The variable (i.e) the weight in the above example, and

II. The frequency in the number of children. There are 50 children having weights ranging from 5 to 10 kg, 200 children. having weight ranging between 10 to 15 kg and so on. 3.5

TABULATION

Tabulation is the process of summarizing classified or grouped data in the form of a table so that it is easily understood and an investigator is quickly able to locate the desired information. A table is a systematic arrangement of classified data in columns and rows. Thus, a statistical table makes it possible for the investigator to present a huge mass of data in a detailed and orderly form. It facilitates comparison and often reveals certain patterns in data which are otherwise not obvious. Classification and 'Tabulation', as a matter of fact, are not two distinct processes. Actually they go together. Before tabulation data are classified and then displayed under different columns and rows of table.

Definition:

Tabulation may be defined, as systematic arrangement of data is column and rows. It is designed to simplify presentation of data for the purpose of analysis and statistical inferences.

Major Objectives of Tabulation:

1) To simplify the complex data

2) To facilitate comparison

3) To economise the space

4) To draw valid inference / conclusions

5) To help for further analysis

Differences between Classification and Tabulation :

1) First data are classified and presented in tables; classification is the basis for tabulation.

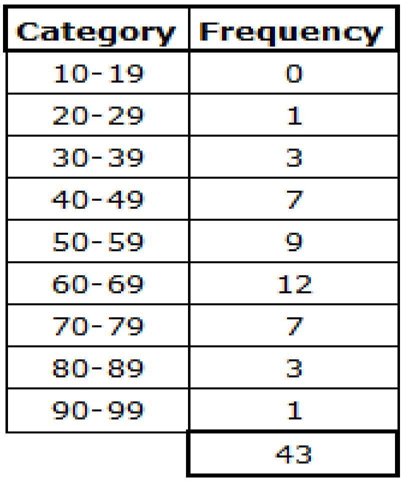
2) Tabulation is a mechanical function of classification because is tabulation classified data are placed in row and columns.

3) Classification is a process of statistical analysis while tabulation is a process of presenting data is suitable structure.

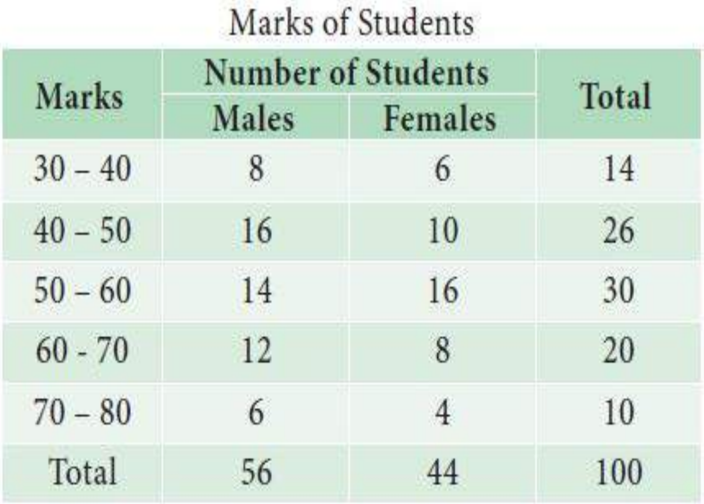
Classification of tables:

Classification is done based on

1. Coverage (Simple and complex table)
2. Objective / purpose (General purpose / Reference table / Special table or summary table)
3. Nature of inquiry (primary and derived table).



Category	Frequency
10-19	0
20-29	1
30-39	3
40-49	7
50-59	9
60-69	12
70-79	7
80-89	3
90-99	1
	43



Marks of Students

Marks	Number of Students		Total
	Males	Females	
30 – 40	8	6	14
40 – 50	16	10	26
50 – 60	14	16	30
60 - 70	12	8	20
70 – 80	6	4	10
Total	56	44	100

PRESENTATION OF DATA

This is the third method of presenting data. This method provides the quickest understanding of the actual situation to be explained by data in comparison to tabular or textual presentations. Diagrammatic presentation of data translates quite effectively the highly abstract ideas contained in numbers into more concrete and easily comprehensible form. Diagrams may be less accurate but are much more effective than tables in presenting the data.

There are various kinds of diagrams in common use. Amongst them the important ones are the following:

(i) Geometric diagram
(ii) Frequency diagram
(iii) Arithmetic line graph

GEOMETRIC DIAGRAM

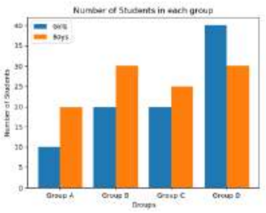
Bar diagram and pie diagram come in the category of geometric diagram. The bar diagrams are of three types: simple, multiple and component bar diagrams.

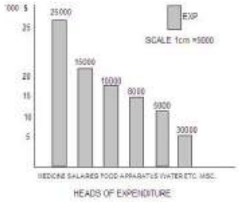
BAR DIAGRAM

A) Simple Bar Diagram

Bar diagram comprises a group of equi spaced and equi width rectangular bars for each class or category of data. Height or length of the bar reads the magnitude of data. The lower end of the bar touches the base line such that the height of a bar starts from the zero unit. Bars of a bar diagram can be visually compared by their relative height and accordingly data are comprehended quickly. Data for this can be of frequency or non-frequency type. In non-frequency type data a particular characteristic, say production, yield, population, etc. at various points of time or of different states are noted and corresponding bars are made of the respective heights according to the values of the characteristic to construct the diagram. The values of the characteristics (measured or counted) retain the identity of each value.

Different types of data may require different modes of diagrammatical representation. Bar diagrams are suitable both for frequency type and non-frequency type variables and attributes. Discrete variables like family size, spots on a dice, grades in an examination, etc. and attributes such as gender, religion, caste, country, etc. can be represented by bar diagrams. Bar diagrams are more convenient for non-frequency data such as income expenditure profile, export/imports over the years, etc.

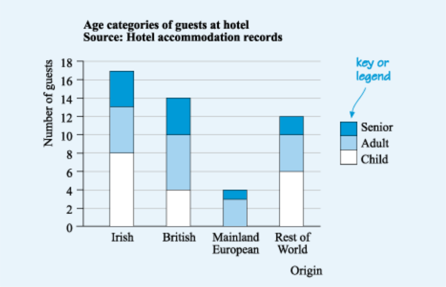




B) Multiple Bar Diagram

Multiple bar diagrams are used for comparing two or more sets of data, for example income and expenditure or import and export for different years, marks obtained in different subjects in different classes, etc.

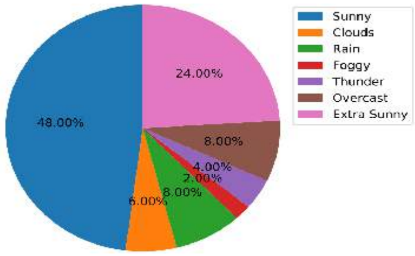
C) **Component Bar Diagram**



Component bar diagrams or charts also called sub-diagrams, are very useful in comparing the sizes of different component parts (the elements or parts which a thing is made up of) and also for throwing light on the relationship among these integral parts. For example, sales proceeds from different products, expenditure pattern in a typical Indian family (components being food, rent, medicine, education, power, etc.), budget outlay for receipts and expenditures, components of labour force, population etc. Component bar diagrams are usually shaded or coloured suitably.

PIE DIAGRAM

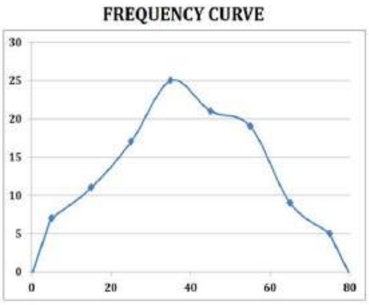
A pie diagram is also a component diagram, but unlike a bar diagram, here it is a circle whose area is proportionally divided among the components it represents. It is also called a pie chart. The circle is divided into as many parts as there are components by drawing straight lines from the centre to the circumference. Pie charts usually are not drawn with absolute values of a category. The values of each category are first expressed as percentage of the total value of all the categories. A circle in a pie chart, irrespective of its value of radius, is thought of having 100 equal parts of 3.6° (360°/100) each. To find out the angle, the component shall subtend at the centre of the circle, each percentage figure of every component is multiplied by 3.6°.



It may be interesting to note that data represented by a component bar diagram can also be represented equally well by a pie chart, the only requirement being that absolute values of the components have to be converted into percentages before they can be used for a pie diagram.

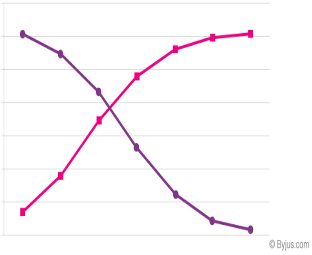
Frequency Curve

The frequency curve is obtained by drawing a smooth freehand curve passing through the points of the frequency polygon as closely as possible. It may not necessarily pass through all the points of the frequency polygon but it passes through them as closely as possible.



Ogive:

Ogive is also called cumulative frequency curve. As there are two types of cumulative frequencies, for example ‘‘less than’’ type and ‘‘more than’’ type, accordingly there are two ogives for any grouped frequency distribution data. Here in place of simple frequencies as in the case of frequency polygon, cumulative frequencies are plotted along y-axis against class limits of the frequency distribution. For ‘‘less than’’ ogive the cumulative frequencies are plotted against the respective upper limits of the class intervals whereas for more than ogives the cumulative frequencies are plotted against the respective lower limits of the class interval. An interesting feature of the two ogives together is that their intersection point gives the median of the frequency distribution. As the shapes of the two ogives suggest, ‘‘less than’’ ogive is never decreasing and ‘‘more than’’ ogive is never increasing.



UNIT - 3

MEASURES OF CENTRAL TENDENCY

MEAN:

The measures of central tendencies are given by various parameters but the most commonly used ones are mean, median and mode.

What is Mean?

Mean is the most commonly used measure of central tendency. It actually represents the average of the given collection of data. It is applicable for both continuous and discrete data.

It is equal to the sum of all the values in the collection of data divided by the total number of values.

Suppose we have n values in a set of data namely as x1, x2, x3, …, xn, then the mean of data is given by:

x¯=x1+x2+x3+……..+xnn

It can also be denoted as: x¯=∑i=1nxin

For grouped data, we can calculate the mean using three different methods of formula.

Direct method	**Assumed mean method**	**Step deviation method**
Mean x—=∑i=1nfixi∑i=1nfi Here, $\sum f_i$ = Sum of all frequencies	Mean (x—)=a+∑fidi∑fi Here, a = Assumed mean $d_i = x_i – a$ $\sum f_i$ = Sum of all frequencies	Mean (x—)=a+h∑fiui∑fi Here, a = Assumed mean $u_i = (x_i – a)/h$ h = Class size $\sum f_i$ = Sum of all frequencies

MEDIAN:

Generally median represents the mid-value of the given set of data when arranged in a particular order.

Median: Given that the data collection is arranged in ascending or descending order, the following method is applied:

- If number of values or observations in the given data is odd, then the median is given by $[(n+1)/2]^{th}$ observation.
- If in the given data set, the number of values or observations is even, then the median is given by the average of $(n/2)^{th}$ and $[(n/2)+1]^{th}$ observation.

The median for grouped data can be calculated using the formula,

Median=l+(N2−cf)×h

MODE:

The most frequent number occurring in the data set is known as the mode.

Consider the following data set which represents the marks obtained by different students in a subject.

Name	Anmol	Kushagra	Garima	Ashwini	Geetika	Shakshi
Marks Obtained (out of 100)	73	80	73	70	73	65

The maximum frequency observation is 73 (as three students scored 73 marks), so the mode of the given data collection is 73.

We can calculate the mode for grouped data using the below formula:

$$\text{Mode}=l+(f1-f02f1-f0-f2)\times h$$

UNIT - 4

MEASURES OF DISPERSION AND VARIABILITY

DEVIATION: MEAN DEVIATION

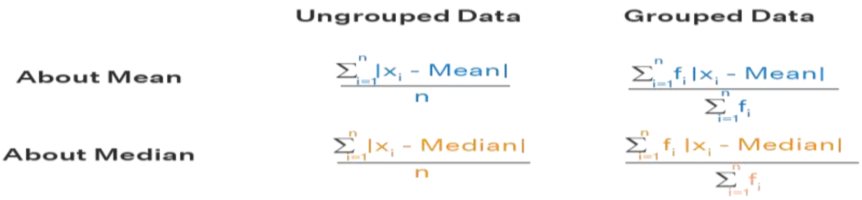
Definition:

The difference between the observed value of a data point and the expected value is known as deviation in statistics. Thus, mean deviation or mean absolute deviation is the average deviation of a data point from the mean, median, or mode of the data set. Mean deviation can be abbreviated as MAD.

Mean Deviation Example

Suppose we have a set of observations given by {2, 7, 5, 10} and we want to calculate the mean deviation about the mean. We find the mean of the data given by 6. Then we subtract the mean from each value, take the absolute value of each result and add them up to get 10. Finally, we divide this value by the total number of observations (4) to get the mean deviation as 2.5.

Mean Deviation Formula



	Ungrouped Data	Grouped Data
About Mean	$\frac{\sum_{i=1}^{n} \lvert x_i - \text{Mean} \rvert}{n}$	$\frac{\sum_{i=1}^{n} f_i \lvert x_i - \text{Mean} \rvert}{\sum_{i=1}^{n} f_i}$
About Median	$\frac{\sum_{i=1}^{n} \lvert x_i - \text{Median} \rvert}{n}$	$\frac{\sum_{i=1}^{n} f_i \lvert x_i - \text{Median} \rvert}{\sum_{i=1}^{n} f_i}$

Depending upon the type of data available as well as the type of the central point, there can be several different formulas to calculate the mean deviation. Given below are the different mean deviation formulas.

Mean Deviation Formula for Ungrouped Data

Data that is not sorted or classified into groups and remains in raw form is known as ungrouped data. To calculate the mean deviation for ungrouped data the formula is as follows:

MAD = ∑n1|xi−¯¯¯x|n∑1n|xi−x¯|n

Here, xixi represents the ith observation, ¯¯¯xx¯ represents the central point (mean, median, or mode), and 'n' is the number of observations in the data set.

Mean Deviation Formula for Grouped Data

When data is organized and classified into groups it is known as grouped data. Grouping of data is done by continuous and discrete frequency distributions. The mean deviation formulas for grouped data are given below:

Mean Deviation for Continuous Frequency Distribution

Such a type of grouped data consists of class intervals. The frequency of repetition of an observation within each interval is given by the continuous frequency distribution. The mean deviation formula is as follows:

MAD = ∑n1fi|xi−¯¯¯x|∑n1fi∑1nfi|xi−x¯|∑1nfi

fifi is the frequency of repetition of xixi. xixi denotes the mid value of the class interval.

Mean Deviation for Discrete Frequency Distribution

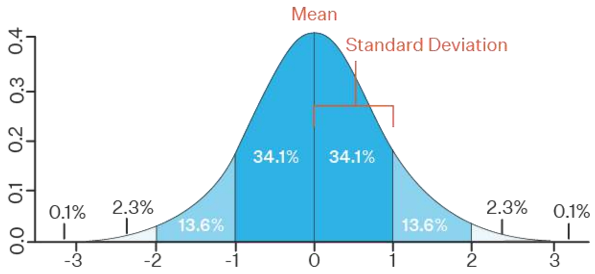
In this type of data, the individual data points are specified and the frequency with which they occur is also mentioned. To calculate the mean deviation for a discrete frequency distribution, the formula is given as follows:

MAD = ∑n1fi|xi−¯¯¯x|∑n1fi∑1nfi|xi−x¯|∑1nfi

xixi denotes the specified individual value and fifi is the frequency of occurrence of that value.

STANDARD DEVIATION

Standard deviation is the degree of dispersion or the scatter of the data points relative to its mean, in descriptive statistics. It tells how the values are spread across the data sample and it is the measure of the variation of the data points from the mean. The standard deviation of a sample, statistical population, random variable, data set, or probability distribution is the square root of its variance.



When we have n number of observations and the observations are x1,x2,.....xnx1,x2,.....xn, then the mean deviation of the value from the mean is determined as ∑ni=1(xi−¯x)2∑i=1n(xi−x¯)2. However, the sum of squares of deviations from the mean doesn't seem to be a proper measure of dispersion. If the average of the squared differences from the mean is small, it indicates that the observations xixi are close to the mean ¯xx¯. This is a lower degree of dispersion. If this sum is large, it indicates that there is a higher degree of dispersion of the observations from the mean ¯xx¯. Thus we conclude that ∑ni=1(xi−¯x)2∑i=1n(xi−x¯)2 is a reasonable indicator of the degree of dispersion or scatter.

We take 1n∑ni=1(xi−¯x)21n∑i=1n(xi−x¯)2 as a proper measure of dispersion and this is called the variance(σ2). The square root of the variance is the standard deviation.

Steps to Calculate Standard Deviation

a) Find the mean, which is the arithmetic mean of the observations.

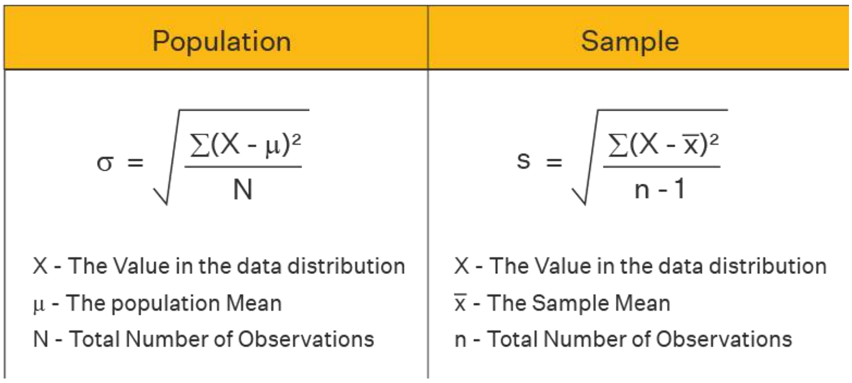
b) Find the squared differences from the mean. (The data value - mean)^2

c) Find the average of the squared differences. (Variance = The sum of squared differences ÷ the number of observations)

d) Find the square root of variance. (Standard deviation = √Variance)

Standard Deviation

Formula The spread of statistical data is measured by the standard deviation. The degree of dispersion is computed by the method of estimating the deviation of data points. You can read about dispersion in summary statistics. As discussed, the variance of the data set is the average square distance between the mean value and each data value. And standard deviation defines the spread of data values around the mean. Here are two standard deviation formulas that are used to find the standard deviation of sample data and the standard deviation of the given population



Population	Sample
$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}}$	$s = \sqrt{\frac{\sum(X - \bar{x})^2}{n - 1}}$
X - The Value in the data distribution μ - The population Mean N - Total Number of Observations	X - The Value in the data distribution $\bar{x}$ - The Sample Mean n - Total Number of Observations

.

Formula for Calculating Standard Deviation: The population standard deviation formula is given as: σ=√1N∑Ni=1(Xi−μ)2σ=1N∑i=1N(Xi−μ)2
Here,
σ = Population standard deviation
μ = Assumed mean
Similarly, the sample standard deviation formula is:

s=√1n−1∑ni=1(xi−¯x)2s=1n−1∑i=1n(xi−x¯)2

Here,
s = Sample standard deviation
¯xx¯ = Arithmetic mean of the observations

Standard Deviation of Ungrouped Data

The calculations for standard deviation differ for different data. Distribution measures the deviation of data from its mean or average position. There are two methods to find the standard deviation.

- actual mean method
- assumed mean method

A) Actual Mean Method

σ = √(∑x−¯x)x−x¯)^2 /n)

Consider the data observations 3, 2, 5, 6. Here the mean of these data points is 16/4 = 4.
The squared differences from mean = (4-3)2+(2-4)2 +(5-4)2 +(6-4)2= 10

Variance = Squared differences from mean/ number of data points =10/4 =2.5
Standard deviation = √2.5 = 1.58

B) Assumed Mean Method

When the x values are large, an arbitrary value (A) is chosen as the mean. The deviation from this assumed mean is calculated as d = x - A.

σ = √[(∑(d)2 /n) - (∑d/n)2]

Standard Deviation of Grouped Data

When the data points are grouped, we first construct a frequency distribution.

Standard Deviation of Grouped Discrete Frequency Distribution

For n number of observations, x1,x2,.....xnx1,x2,.....xn, and the frequency, f1,f2,f3,...fnf1,f2,f3,...fn the standard deviation is:

σ=√1N∑Ni=1fi(Xi−¯x)2σ=1N∑i=1Nfi(Xi−x¯)2. Here N = ∑Ni=1fi

Difference between Mean deviation and Standard deviation

Mean Deviation	Standard Deviation
We use the central points (mean, median, or mode) to find the mean deviation.	We only use the mean to find the standard deviation.
We take the absolute value of the deviations to find the mean deviation.	To find the standard deviation, we use the square of the deviations.
It is less frequently used.	It is the most common measure of variability and is more frequently

	used.
If the data has a greater number of outliers, mean absolute deviation is used.	If there are a lesser number of outliers in the data, then standard deviation is used.

COEFFICIENT OF VARIATIONS

Coefficient of variation formula (CV), also known as relative standard deviation (RSD), is a standardized measure of the dispersion of a probability distribution or frequency distribution. When the value of the coefficient of variation is lower, it means the data has less variability and high stability.

The formula for coefficient of variation is given below:

coefficient of variation=Standard Deviation/Mean×100%

As per sample and population data type, the formula for standard deviation may vary.

Sample Standard Deviation=∑i=1n(Xi−X—)2n−1

Population Standard Deviation=∑i=1n(Xi−X—)2n

Where,
x_i = Terms given in the data

x—=Mean
n = Total number of terms.

STANDARD ERROR

Meaning:

The standard error is one of the mathematical tools used in statistics to estimate the variability. It is abbreviated as SE. The standard error of a statistic or an estimate of a parameter is the standard deviation of its sampling distribution. We can define it as an estimate of that standard deviation. Formula: The accuracy of a sample that describes a population is identified through the SE formula. The sample mean which deviates from the given population and that deviation is given as;

$$SE_{\overline{x}} = \frac{S}{\sqrt{n}}$$

Where S is the standard deviation and n is the number of observations. Standard Error of the Mean (SEM)

The standard error of the mean also called the standard deviation of mean, is represented as the standard deviation of the measure of the sample mean of the population. It is abbreviated as SEM. For example, normally, the estimator of the population mean is the sample mean. But, if we draw another sample from the same population, it may provide a distinct value.

Thus, there would be a population of the sampled means having its distinct variance and mean. It may be defined as the standard deviation of such sample means of all the possible samples taken from the same given population. SEM defines an estimate of standard deviation which has been computed from the sample. It is calculated as the ratio of the standard deviation to the root of sample size, such as:

$$SEM = \frac{s}{\sqrt{n}}$$

Where 's' is the standard deviation and n is the number of observations. The standard error of the mean shows us how the mean changes with different tests, estimating the same quantity. Thus if the outcome of random variations is notable, then the standard error of the mean will have a higher value. But, if there is no change observed in the data points after repeated experiments, then the value of the standard error of the mean will be zero.

Standard Error of Estimate (SEE)

The **standard error** of the **estimate** is the estimation of the accuracy of any predictions. It is denoted as SEE. The regression line depreciates the sum of squared deviations of prediction. It is also known as the sum of squares **error. SEE** is the square root of the average squared **deviation**. The deviation of some estimates from intended values is given by standard error of estimate formula.

$$SEE = \sqrt{\frac{\sum(x_i - \bar{x})}{n-2}}$$

Where xi stands for data values, x bar is the mean value and n is the sample size.

SKEWNESS

Measures of skewness help us to know to what degree and in which direction (positive or negative) the frequency distribution has a departure from symmetry. Although positive or negative skewness can be detected graphically depending on whether the right tail or the left tail is longer but, we don’t get idea of the magnitude. Besides, borderline cases between symmetry and asymmetry may be difficult to detect graphically.

Hence some statistical measures are required to find the magnitude of lack of symmetry. A good measure of skewness should possess three criteria:

1) It should be a unit free number so that the shapes of different distributions, so far as symmetry is concerned, can be compared even if the unit of the underlying variables are different;
2) If the distribution is symmetric, the value of the measure should be zero. Similarly, the measure should give positive or negative values according as the distribution has positive or negative skewness respectively; and
3) As we move from extreme negative skewness to extreme positive skewness, the value of the measure should vary accordingly. Measures of skewness can be both absolute as well as relative.

Since in a symmetrical distribution mean, median and mode are identical more the mean moves away from the mode, the larger the asymmetry or skewness. An absolute measure of skewness cannot be used for purposes of comparison because of the same amount of skewness has different meanings in distribution with small variation and in distribution with large variation.

Absolute Measures of Skewness:

Following are the absolute measures of skewness:

1. Skewness (Sk) = Mean – Median
2. Skewness (Sk) = Mean – Mode
3. Skewness (Sk) = (Q3 - Q2) - (Q2 - Q1)

For comparing to series, we do not calculate these absolute mearues we calculate the relative measures which are called coefficient of skewness. Coefficient of skewness are pure numbers independent of units of measurements.

Relative Measures of Skewness:

In order to make valid comparison between the skewness of two or more distributions we have to eliminate the distributing influence of variation. Such elimination can be done by dividing the absolute skewness by standard deviation. The following are the important methods of measuring relative skewness:

A) **Karl Pearson's Coefficient of Skewness:**

This method is most frequently used for measuring skewness. The formula for measuring coefficient of skewness is given by

Sk = Mean – Mode/σ

The value of this coefficient would be zero in a symmetrical distribution. If mean is greater than mode, coefficient of skewness would be positive otherwise negative. The value of the Karl Pearson's coefficient of skewness usually lies between ±1 for moderately skewed distubution. If mode is not well defined, we use the formula

Sk = 3 (Mean – Median) / σ

By using the relationship Mode = (3 Median – 2 Mean)

Here, 3 S 3. – ≤ k ≤ In practice it is rarely obtained.

B) **Bowleys's Coefficient of Skewness**

This method is based on quartiles. The formula for calculating coefficient of skewness is given by

SKB=Q3+Q1−2Q2Q3−Q1

The value of Sk would be zero if it is a symmetrical distribution. If the value is greater than zero, it is positively skewed and if the value is less than zero it is negatively skewed distribution. It will take value between +1 and -1

KURTOSIS

Kurtosis is a statistical measure used to describe a characteristic of a dataset. When normally distributed data is plotted on a graph, it generally takes the form of an upsidedown bell. This is called the bell curve. The plotted data that are furthest from the mean of the data usually form the tails on each side of the curve. Kurtosis indicates how much data resides in the tails.

Distributions with a large kurtosis have more tail data than normally distributed data, which appears to bring the tails in toward the mean. Distributions with low kurtosis have fewer tail data, which appears to push the tails of the bell curve away from the mean.

Types of Kurtosis

There are three categories of kurtosis that a set of data can display; mesokurtic, leptokurtic, and platykurtic. All measures of kurtosis are compared against a normal distribution curve.

Mesokurtic (kurtosis = 3.0)

The first category of kurtosis is mesokurtic distribution. This distribution has a kurtosis similar to that of the normal distribution, meaning the extreme value characteristic of the distribution is similar to that of a normal distribution. Therefore, a stock with a mesokurtic distribution generally depicts a moderate level of risk.

Leptokurtic (kurtosis > 3.0)

The second category is leptokurtic distribution. Any distribution that is leptokurtic displays greater kurtosis than a mesokurtic distribution. This distribution appears as a curve one with long tails (outliers.) The "skinniness" of a leptokurtic distribution is a consequence of the outliers, which stretch the horizontal axis of the histogram graph, making the bulk of the data appear in a narrow ("skinny") vertical range.

A stock with a leptokurtic distribution generally depicts a high level of risk but the possibility of higher returns because the stock has typically demonstrated large price movements.

Platykurtic (kurtosis < 3.0)

The final type of distribution is platykurtic distribution. These types of distributions have short tails (fewer outliers.). Platykurtic distributions have demonstrated more stability than other curves because extreme price movements rarely occurred in the past. This translates into a less-than-moderate level of risk.

Using Kurtosis

Kurtosis is used in financial analysis to measure an investment's risk of price volatility. Kurtosis risk differs from more commonly used measurements such as alpha, beta, r-squared, or the Sharpe ratio. Alpha measures excess return relative to a benchmark index, and beta measures the volatility a stock has compared to the broader market.

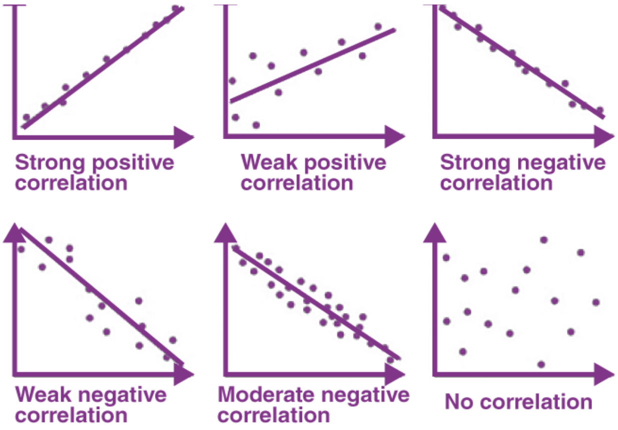
R-squared measures the percent of movement a portfolio or fund has that can be explained by a benchmark, and the Sharpe ratio compares return to risk. Kurtosis measures the amount of volatility an investment's price has experienced regularly.

UNIT - 5

CORRELATION ANALYSIS

SCATTER DIAGRAM

A scatter diagram is used to examine the relationship between both the axes (X and Y) with one variable. In the graph, if the variables are correlated, then the point drops along a curve or line. A scatter diagram or scatter plot gives an idea of the nature of relationship.



In a scatter correlation diagram, if all the points stretch in one line, then the correlation is perfect and is in unity. However, if the scatter points are widely scattered throughout the line, then the correlation is said to be low. If the scatter points rest near a line or on a line, then the correlation is said to be linear.

KARL PEARSON’S CORRELATION COEFFICIENT

Karl Pearson's coefficient of correlation is defined as a linear correlation coefficient that falls in the value range of -1 to +1**. Value of -1 signifies strong negative correlation while +1 indicates strong positive correlation.** Coefficient of Correlation.

A coefficient of correlation is generally applied in statistics to calculate a relationship between two variables. The correlation shows a specific value of the degree of a linear relationship between the X and Y variables, say X and Y.

There are various types of correlation coefficients. However, Pearson’s correlation (also known as Pearson’s R) is the correlation coefficient that is frequently used in linear regression.

Pearson’s Coefficient Correlation

Karl Pearson’s coefficient of correlation is an extensively used mathematical method in which the numerical representation is applied to measure the level of relation between linearly related variables. The coefficient of correlation is expressed by “r”.

Karl Pearson Correlation Coefficient Formula

$$r = \frac{\Sigma \times Y}{\sqrt{X^2}\sqrt{Y^2}}$$

$$r = \frac{\Sigma XY}{N\sigma X \times \sigma Y}$$

$$\sigma X = \text{S.D. of X series}$$

$$\sigma Y = \text{S.D. of Y series}$$

Alternative Formula (covariance formula)

$$\mathrm{Cov}(X,Y) = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{N} = \frac{\sum xy}{N}$$

Pearson correlation example

1) When a correlation coefficient is (1), that means for every increase in one variable, there is a positive increase in the other fixed proportion. For example, shoe sizes change according to the length of the feet and are perfect (almost) correlations.

2) When a correlation coefficient is (-1), that means for every positive increase in one variable, there is a negative decrease in the other fixed proportion. For example, the decrease in the quantity of gas in a gas tank shows a perfect (almost) inverse correlation with speed.

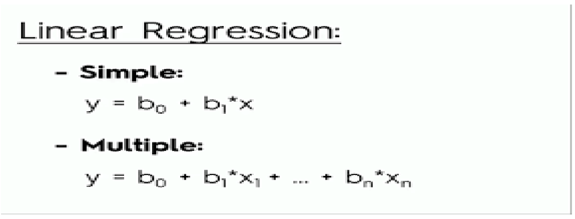
3) When a correlation coefficient is (0) for every increase, that means there is no positive or negative increase, and the two variables are not related.

REGRESSION ANALYSIS

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

Simple regression or ordinary least squares (OLS), linear regression is the most common form of this technique. Linear regression establishes the linear relationship between two variables based on a line of best fit. Linear regression is thus graphically depicted using a straight line with the slope defining how the change in one variable impacts a change in the other. The y-intercept of a linear regression relationship represents the value of one variable when the value of the other is zero. Non-linear regression models also exist, but are far more complex.

Regression analysis is a powerful tool for uncovering the associations between variables observed in data, but cannot easily indicate causation.



Linear Regression:

- **Simple:**
  $y = b_0 + b_1{*}x$
- **Multiple:**
  $y = b_0 + b_1{*}x_1 + \ldots + b_n{*}x_n$

TEST OF SIGNIFICANCE

TEST OF SIGNIFICANCE

A test of significance may be a formal procedure for comparing observed data with a claim (also called a hypothesis), the reality of which is being assessed. It may be a statement with a few of the parameters, like the population proportion p or the population mean µ.

Once the sample data has been collected through an observational study or an experiment, statistical inference will allow the analysts to assess the evidence in favor or some claim about the population from which the sample has been taken from.

Null Hypothesis

Every test for significance starts with a null hypothesis H0. H0 represents a theory that has been suggested, either because it's believed to be true or because it's to be used as a basis for argument, but has not been proved. For example, during a clinical test of a replacement drug, the null hypothesis could be that the new drug is not any better, on average than the present drug. We would write H0: there's no difference between the 2 drugs on average.

Alternative Hypothesis

The alternative hypothesis, Ha, maybe a statement of what a statistical hypothesis test is about up to determine. For example, during a clinical test of a replacement drug, the choice hypothesis could be that the new drug features a different effect, on the average, compared to the current drug. We would write Ha: the 2 drugs have different effects, on the average. The alternative hypothesis may additionally be that the new drug is better, on the average than the present drug. In this case, we might write Ha: the new drug is better than the present drug, on the average.

The final conclusion once the test has been administered is usually given in terms of the null hypothesis. Either we "reject the H0 in favor of Ha" or "we do not reject the H0"; we never conclude "reject Ha", or maybe "accept Ha".

Test of Significance in Statistics

Technically speaking, in the test of significance the statistical significance refers to the probability of the results of some statistical tests or research occurring accidentally. The main purpose of performing statistical research is essential to seek out reality. In this process, the researcher has to confirm the standard of the sample, accuracy, and good measures that require a variety of steps to be done. The researcher determines whether the findings of experiments have occurred thanks to an honest study or simply by fluke.

The significance may be a number that represents probability indicating the results of some study has occurred purely accidentally. The statistical significance may be weak or strong. It does not necessarily indicate practical significance. Sometimes, when a researcher doesn't carefully make use of language within the report of their experiment, the importance could also be misinterpreted.

The psychologists and statisticians search for a 5% probability or less which suggests 5% of results occur thanks to chance. This also indicates that there's a 95% chance of results occurring NOT accidentally. Whenever it's found that the results of our experiment are statistically significant, it refers that we should always be 95% sure the results aren't due to chance.

Process of Significance Testing in Test of Significance

So in this process of testing for statistical significance, the following are the steps:

1. Stating a Hypothesis for Research
2. Stating a Null Hypothesis
3. Selecting a Probability of Error Level
4. Selecting and Computing a Statistical Significance Test
5. Interpreting the results.

ANOVA (One Way)

A one-way ANOVA is a type of statistical test that compares the variance in the group means within a sample whilst considering only one independent variable or factor. It is a hypothesis-based test, meaning that it aims to evaluate multiple mutually exclusive theories about our data. ANOVA is a technique that allows us to compare two or more populations of interval data.

The purpose of a one-way **ANOVA** test is to determine the existence of a statistically significant difference among several group means. The test actually uses variances to help determine if the means are equal or not. In order to perform a one-way ANOVA test, there are five basic assumptions to be fulfilled:

1. Each population from which a sample is taken is assumed to be normal.
2. All samples are randomly selected and independent.
3. The populations are assumed to have **equal standard deviations (or variances)**.
4. The factor is a categorical variable.
5. The response is a numerical variable.

Formula

$$\mathbf{F = MSB/MSW}$$

In this formula, F = coefficient of ANOVA. MSB = Mean sum of squares between the groups. MSW = Mean sum of squares within groups.

The Null and Alternative Hypotheses

The null hypothesis is simply that all the group population means are the same. The alternative hypothesis is that at least one pair of means is different. For example, if there are k groups:

$H0$: $\mu1 = \mu2 = \mu3 = \ldots = \mu k$

Ha: At least two of the group means $\mu1$, $\mu2$, $\mu3$, …, μk are not equal.

The graphs, a set of box plots representing the distribution of values with the group means indicated by a horizontal line through the box, help in the understanding of the hypothesis test. In the first graph (red box plots), $H0$: $\mu 1 = \mu 2 = \mu 3$ and the three populations have the same distribution if the null hypothesis is true. The variance of the combined data is approximately the same as the variance of each of the populations.

Analysis of variance extends the comparison of two groups to several, each a level of a categorical variable (factor). Samples from each group are independent, and must be randomly selected from normal populations with equal variances. We test the null hypothesis of equal means of the response in every group versus the alternative hypothesis of one or more group means being different from the others. A one-way ANOVA hypothesis test determines if several population means are equal. The distribution for the test is the F distribution with two different degrees of freedom.

Assumptions:

1. Each population from which a sample is taken is assumed to be normal.
2. All samples are randomly selected and independent.
3. The populations are assumed to have equal standard deviations (or variances).Top of Form