Chapter One

1. Introduction

This chapter introduces the subject matter of statistics, the art of learning from data. It describes the two branches of statistics, descriptive and inferential. The idea of learning about a population by sampling and studying certain of its members is described. It is reasonable to start by thinking about this question, relating it to your own experiences, and talking it over with friends. However, if you want to convince others and obtain a consensus, it is then necessary to gather some objective information. We are using it to indicate the modern approach to learning about a complicated question. Namely, one must collect relevant information, or data, and these data must then be described and analyzed in such a way that valid conclusion is made. This is the subject matter of statistics.

1.1 Definition and Classification of Statistics

The common usage of the word "statistics" has, therefore, two meanings. In one sense "statistics" is the plural form which refers to the numerical facts and figures collected for a certain purposes. **Statistics as a numerical data:** In this sense statistics is defined as aggregates of numerical expressed facts (figures) collected in a systematic manner for a predetermined purpose.(*in plural sense*)

In the other sense;" statistics" refers to a field of study or to a body of knowledge or to a subject that is concerned with systematic collection and interpretation of numerical data to make a decision. In this sense the word statistics is singular. **Statistics as a subject (field of study):** in this sense statistics is defined as the science of collecting, organizing, presenting, analyzing and interpreting numerical data to make decision on the bases of such analysis.(***in singular sense)***

In this course, we shall be mainly concerned with statistics as a subject, that is, as a field of study.

Classification of statistics

Anyone can apply statistical techniques to, virtually, every branch of science and art. These techniques are so diverse that statisticians commonly classify them into the following two broad categories (Descriptive statistics and inferential statistics)

Descriptive Statistics: it is an area of statistics which is mainly concerned with the methods and techniques used in collection, organization, presentation, and analysis of a set of data without making any conclusions or inferences. According to this definition the activities in the area of Descriptive Statistics include:

- Gathering data
- Editing and classifying them
- Presenting data in tables
- drawing diagrams and graphs for them
- Calculating averages and measures of dispersions.

Descriptive statistics doesn't go beyond describing the data themselves.

Examples of activities of descriptive statistics:

- Recording a students' grades throughout the semester and then finding the average of these grades.
- Drawing graphs that show the difference in the scores of males and females.

All the above examples simply summarize and describe a given data. Nothing is inferred or concluded on the basis of the above description.

Inferential Statistics: Inferential statistics is an area of statistics which deals with the method of inferring or drawing conclusion about the characteristics of the population based upon the results of a sample. Statistics is concerned not only with collection , organization , presentation and analysis of data but also with the inferences which can be made after the analysis is completed. In collecting data concerning the characteristics of a set of elements, or the element can even be infinite. Instead of observing the entire set of objects, called the population, one observes a subset of the population called a sample.Inferential statistics utilizes sample data to make decision for entire data set based on sample. Examples of inferential Statistics are:-

1. "There is a definitive relationship between smoking and lung cancer". This statement is the result of continuous research of many samples taken and studied. Therefore, it is an **inference** made from sample results.
2. As a result of recent survey of public opinion, most Americans are in favor of building additional nuclear power plant.

1.2 Definition of Some Basic Statistical Terms

In this section, we will define those terms which will be used frequently**.**

Data: Data as a collection of related facts and figures from which conclusions may be drawn.In other words data is simply a scientific term for facts, figures, information and measurement.

Population (Sampled/target population)*:* A population is a totality of things, objects, peoples, etc about which information is being collected. It is the totality of observations with which the researcher is concerned. The population represents the target of an investigation, and the objective of the investigation is to draw conclusions about the population hence we sometimes call it *target population.* Example: population of trees under specified climatic conditions, population of animals fed a certain type of diet, population of households, etc.

Census: a complete enumeration of the population. But in most real problems it cannot be realized, hence we take sample.

Sample: A sample is a part of a population selected to draw conclusions about the population**.**

Sampling: The process of selecting a sample from the population.

Sample size: The number of elements or observation to be included in the sample.

Statistic: It is a value computed from the sample, used to describe the sample.

Parameter: It is a descriptive measure (value) computed from the population. It is the population measurement used to describe the population. Example: population mean and standard deviation.

Sampling frame:-A list of people, items or units from which the sample is taken.

Variable: A certain characteristic whose value changes from object to object and time to time.

Sample size*:* The number of elements or observation to be included in the sample.

Survey: is an investigation of certain population to access its characteristics when no special control is exerted over any factors influence the population characteristics.

Experiment: is the collection of data when special control is made over one or more factors influencing the population characteristics.

Census survey: It is the process of examining the entire population.

Censes survey (studying the whole population without considering samples) requires a great deal of time, money and energy. Trying to study the entire population is in most cases technically and

economically not feasible. To solve this problem, we take a representative sample out of the population on the basis of which we draw conclusions about the entire population.

Therefore, sampling survey

- Helps to estimate the parameter of a large population.
- Is cheaper, practical, and convenient.
- Save time and energy and easy to handle and analysis.

1.3 Stages in Statistical Investigation

Before we deal with statistical investigation, let us see what statistical data mean. Each and every numerical data can't be considered as statistical data unless it possesses the following criteria.

- The data must be aggregate of facts
- They must be affected to a marked extent by a multiplicity of causes
- They must be estimated according to reasonable standards of accuracy
- The data must be collected in a systematic manner for predefined purpose
- The data should be placed in relation to each other

Statistician/practitioners should be involved at all the different stages of statistical investigation when planning to conduct scientific research. This includes *formulating the problem* and then *collecting*, *organizing* (classifying), *presenting*, *analyzing* and *interpreting* of statistical data.

Problem Formulation: problem of study is identified and objective of the study could be defined.

Data Collection: This is a stage where we gather information for our purpose

- If data are needed and if not readily available, then they have to be collected.
- Data may be collected by the investigator directly using methods like interview, questionnaire, and observation or may be available from published or unpublished sources.
- Data gathering is the basis (foundation) of any statistical work.
- Valid conclusions can only result from properly collected data.

Data Organization: It is a stage where we edit our data .A large mass of figures that are collected from surveys frequently need organization. The collected data involve irrelevant figures, incorrect facts, omission and mistakes. Errors that may have been included during collection will have to be edited. After editing, we may classify (arrange) according to their common characteristics. Classification or arrangement of data in some suitable order makes the information easer for presentation.

Data Presentation: The organized data can now be presented in the form of tables, diagram and graphs. At this stage, large data will be presented in tables in a very summarized and condensed manner. The main purpose of data presentation is to facilitate statistical analysis. Graphs and diagrams may also be used to give the data a bright meaning and make the presentation attractive.

Data Analysis: This is the stage where we critically study the data to draw conclusions about the population parameter. The purpose of data analysis is to dig out information useful for decision making. Analysis usually involves highly complex and sophisticated mathematical techniques. However, in this course only the most commonly used methods of statistical analysis are included in next chapters. Such as the calculations of averages, the computation of majors of dispersion, regression and correlation analysis are covered.

Data Interpretation: This is the stage where one draw valid conclusions from the results obtained through data analysis. Interpretation means drawing conclusions from the data which form the basis for decision making. The interpretation of data is a difficult task and requires a high degree of skill and experience. If data that have been analyzed are not properly interpreted, the whole purpose of the investigation may be defected and fallacious conclusion be drawn. So that great care is needed.

1.4 Applications, Uses and Limitations of Statistics

Application of Statistics

The scope of statistics is indeed very vast; and applicable in almost all fields of human endeavor. Apart from helping elicit an intelligent assessment from a body of figures and facts, statistics is indispensable tool for any scientific enquiry-right from the stage of planning enquiry to the stage of conclusion. It applies almost all sciences: pure and applied, physical, natural, biological, medical, agricultural, engineering, and others. It also finds applications in social and management sciences, in commerce, business and industry, applicable in some process e.g. invention of certain drugs, extent of environmental pollution and industries especially in quality control area etc.

Uses of statistics

Today the field of statistics is recognized as a highly useful tool to making decision process by managers of modern business, industry, frequently changing technology. It has a lot of functions in everyday activities. The following are some of the most important uses of statistics.

- *Statistics condenses and summarizes complex data*. The original set of data (raw data) is normally voluminous and disorganized unless it is summarized and expressed in few numerical values.
- *Statistics facilitates comparison of data*. Measures obtained from different set of data can be compared to draw conclusion about those sets. Statistical values such as averages, percentages, ratios, etc, are the tools that can be used for the purpose of comparing sets of data.
- *Statistics helps in predicting future trends*. Statistics is extremely useful for analyzing the past and present data and predicting some future trends.
- *Statistics influences the policies of government.* Statistical study results in the areas of taxation, on unemployment rate, on the performance of every sort of military equipment, family planning, etc, may convince a government to review its policies and plans with the view to meet national needs.
- Statistical methods are very *helpful in formulating and testing hypothesis* and to develop new theories.

Limitations of Statistics

Even though, statistics is widely used in various fields of natural and social sciences, which closely related with human inhabitant. It has its own limitations as far as its application is concerned. Some of these limitations are:

- *Statistics doesn't deal with single (individual) values*. Statistics deals only with aggregate values. But in some cases single individual is highly important to consider in some situations. Example, the sun, a deriver of bus, president, etc.
- *Statistics can't deal with qualitative characteristics*. It only deals with data which can be quantified. Example, it does not deal with marital status (married, single, divorced, widowed) but it deal with number of married, number of single, number of divorced.
- *Statistical conclusions are not universally true*. Statistical conclusions are true only under certain condition or true only on average. The conclusions drawn from the analysis of the sample may, perhaps, differ from the conclusions that would be drawn from the entire population. *Statistical interpretations require a high degree of skill and understanding of the subject*. It requires extensive training to read and interpret statistics in its proper context. It may lead to wrong conclusions if inexperienced people try to interpret statistical results.

- *Statistics can be misused*. Sometimes statistical figures can be misleading unless they are carefully interpreted. Example: The report of head of the minister about Ethio-Somalia terrorist attack mission dismissed terrorists 25% at first day, 50% at second day, 75% at third day. However, we doubt about the mechanisms how the mission is measured and quantified. This leads miss use of statistical figures.

1.5 Types of variables and Scales of Measurements

Variables and Attributes: A variable in statistics is any characteristic, which can take on different values for different elements when data are collected. A *quantitative* or *qualitative* characteristic that varies from observation to observation in the same group is called a *variable*. In case of quantitative variables, observations are made using interval scales whereas in case qualitative variables nominal scales are used. Conventionally, the quantitative variables are termed as ***variables*** and qualitative variables are termed as ***attributes.*** Example: attributes like gender, religion, marital status, coding etc. are attributes.

Types of Variables

A. **Continuous Variables:** - are usually obtained by measurement not by counting. These are variables which assume or take any decimal value when collected. The variables like age, time, height, income, price, temperature, length, volume, rate, time, amount of rainfall and etc are all continuous since the data collected from such variables can take decimal values.
B. **Discrete Variables:** - are obtained by counting. A discrete variable takes always whole number values that are counted. Example: Variables such as number of students, number of errors per page, number of accidents on traffic line, number of defective or non-defectives in production, etc.

Scales of Measurements: Normally, when one hears the term *measurement*, they may think in terms of measuring the length of something (i.e. the length of a piece of wood) or measuring a quantity of something (i.e. a cup of flour). This represents a limited use of the term measurement. In statistics, the term measurement is used more broadly and is more appropriately termed *scales of measurement*. A scale of measurement refers to ways in which variables or numbers are defined and categorized and/or is the assignment of numbers to objects or events in a systematic fashion. Each scale of measurement has certain properties which in turn determine the appropriateness for use of certain statistical analyses. The various measurement scales results from the facts that measurement may be carried out under different sets of

rules. Four levels of measurement scales are commonly distinguished: nominal, ordinal, interval, and ratio; and each possessed different properties of measurement systems.

i. ***Nominal Scale:-***Consists of 'naming' observations or classifying them into various mutually exclusive categories. Sometimes the variable under study is classified by some quality it possesses rather than by an amount or quantity. In such cases, the variable is called attribute. *Example* Religion (Christianity, Islam, Hinduism, etc), Sex (Male, Female), Eye color (brown, black, etc.), Blood type (A, B, AB and O), Departments (HWRE, IWRE, WSEE, etc), etc.

ii. ***Ordinal Scale: -***Whenever observations are not only different from category to category, but can be ranked according to some criterion. The variables deal with their relative difference rather than with quantitative differences. Ordinal data are data which can have meaningful inequalities. The inequality signs < or > may assume any meaning like 'stronger, softer, weaker, better than', etc. *Example:* Patients may be characterized as (unimproved, improved & much improved), Individuals may be classified according to socio-economic as (low, medium & high), Letter for grading system (A, B, C, D, F), speed of certain machine, authority, career, etc.

Note: Qualitative variables can be either Nominal or Ordinal scales of measurements.

iii. ***Interval Scale:*** With this scale it is not only possible to order measurements, but also the distance between any two measurements is known but not meaningful quotients. There is no true zero point but arbitrary zero point. Interval data are the types of information in which an increase from one level to the next always reflects the same increase. Possible to add or subtract interval data but they may not be multiplied or divided. *Example:* Temperature of zero degrees does not indicate lack of heat. The two common temperature scales; Celsius (C) and Fahrenheit (F). We can see that the same difference exists between 10^oC (50^oF) and 20^oC (68^OF) as between 25^oc (77^oF) and 35^oc (95^oF) i.e. the measurement scale is composed of equal-sized interval. But we cannot say that a temperature of 20^oc is twice as hot as a temperature of 10^oc because the zero point is arbitrary.

iv. ***Ratio Scale: -*** Characterized by the fact that equality of ratios as well as equality of intervals may be determined. Fundamental to ratio scales is a true zero point. All arithmetic operations are used to calculate values of ratio scale. Most statistical data analysis procedures do not distinguish between the interval and ratio properties of the measurement scales. *Example:* Variables such as age, height, length, volume, rate, time, amount of rainfall, etc. are require ratio scale.

Note: Quantitative variables can be either Interval or Ratio scales of measurements.

Chapter Two

2. Methods of Data Collection and Presentation

2.1. Sources of Data

Before one deal with methods of data collection it is necessary to see sources of data from which data are collected, there are two sources of data that pervades the nature of data. These are Primary sources and Secondary sources. The **primary sources** of data are objects or persons from which we collect the figures used for firsthand information. The data obtained from sources are measurements observed or recorded as a part of an original study or surveys being conducted are called ***primary data, or*** data measured or collect by the investigator or the user directly from the source. The key point here is that the data you collect is unique to you and your research and, until you publish, no one else has access to it. Examples of primary data are weight, feeling, current air pressure, etc.

The **secondary sources** are either published or unpublished materials or records. ***Secondary data*** can be literally defined as second-hand information and data or information that was either gathered by someone else (e.g., researchers, institutions, other NGOs, etc.) or for some other purpose than the one currently being considered, or often a combination of the two. Some of the sources of secondary data are *government document*, *official statistics*, *technical report*, *scholarly journals*, *trade journals*, *review articles*, *reference books*, *research institutes*, *universities*, *hospitals*, *libraries*, *library search engines*, *computerized data base* and *world wide web* (WWW). Examples of secondary data are age, birth date, vital statistics from reports, etc.

Before use of secondary data investigator should examine:

- The type and objective of the situations.
- The purpose for which the data are collected and compatible with the present problem.
- The nature and classification of data is appropriate to our problem.
- There are no biases and misreporting in the published data.
- Reliability, homogeneity, and completeness.

2.2. Methods of Data Collection

Planning to data collection requires: identify source and elements of the data, decide whether to consider sample or census, if sampling is preferred, decide on sample size, selection method, etc, decide measurement procedure, and set up the necessary organizational structure. Then, data can be collected by one of the following methods based on nature of items of information, objectives of the enquiry, operational feasibility and cost. In primary data collection, you collect the data yourself using methods such as interviews, observations, laboratory experiments and questionnaires.

1) ***Observation:*** It involves recording the behavioral patterns of people, objects and events in a systematic manner. The technique ranges from single visual observation to those requiring special skills like direct observation/examination. *Example:* measuring height, weight, temperature, chemical component in water, etc. And it may include laboratory experiment; conducting laboratory experiments on fields of chemical, biological sciences and so on.
2) ***Questionnaire:*** It is a popular means of collecting data, but is difficult to design and often require many rewrites before an acceptable questionnaire is produced. In this method a set of questions are administered (provided) to respondent either physically or through mail (Email, Postal, etc). Schedule through enumerations is the method in which investigator approach to the informant with prepared questionnaire and got replies to the questions.

Advantages:

- Can be used as a method in its own right or as a basis for interviewing or a telephone survey.
- Can be posted, e-mailed or faxed.
- Can cover a large number of people or organizations and wide geographic coverage.
- Relatively cheap and avoids embarrassment on the part of the respondent.
- Respondent can consider responses, and there is no interviewer bias.

Disadvantages:

- Historically low response rate (although inducements may help).
- Time delay whilst waiting for responses to be returned
- Several reminders may be required and it assumes no literacy problems.
- No control over who completes, and it is not possible to give assistance if required.
- Respondent can read all questions beforehand and then decide whether to complete or not.

3) ***Interviewing:*** is a technique that is primarily used to gain an understanding of the underlying reasons and motivations for people's attitudes, preferences or behavior. Interviews can be undertaken on a *personal (face to face)or via telephone (indirect method)*. They can be conducted at work, at home, in the street or in a shopping center, or some other agreed location.

Advantages:

- Serious approach by respondent resulting in accurate information and good response rate.
- Characteristics (motives and feelings) of respondent assessed – tone of voice, facial expression, hesitation, etc.
- Interviewer in control and can give help if there is a problem
- If one interviewer used, uniformity of approach.
- Completed and immediate.
- Possible in-depth questions.
- Can use recording equipment.
- Used to pilot than other methods.

Disadvantages:

- Need to set up interviews.
- If many interviewers, training required.
- Time consuming.
- Geographic limitations.
- Can be expensive.
- Normally need a set of questions.
- Respondent bias – tendency to please or impress, create false personal image, or end interview quickly, Embarrassment possible if personal questions.
- Transcription and analysis can present problems (subjectivity).

4) Extract from Records/Documentary Sources: it is method of collecting information (secondary data) from published or unpublished sources. Secondary data also collected from *diaries*. A diary is a way of gathering information about the way individuals spend their time on professional activities.

Advantage of secondary data

- Secondary data may help to clarify or redefine the definition of the problem as part of the exploratory research process.
- Provides a larger database as compared to primary data

- Time saving
- Does not involve collection of data

Disadvantage of secondary data

- Lack of availability
- Inaccurate data
- Lack of relevance
- Insufficient data

5) Focus Group Discussion, Experimental Design, New Product Registration, etc

2.3. Methods of Data Presentation

So far you know how to collect data. So what do we do with the collected data next? Now you have to present the data you have collected so that they can be of use. Thus, the collected data also known as 'raw data' are always in an unorganized form and need to be organized and presented in a meaningful and readily comprehensible form in order to facilitate further statistical analysis. **Raw data:** recorded information in its original collected form, whether it is counts or measurements, is referred to as raw data. Classification is a preliminary and it prepares the ground for proper presentation of data.

The objectives of data classification (organization) are:-

- ☞ To avoid/eliminate unnecessary details and repetitions.
- ☞ To bring clear point of similarity and differences; i.e.to arrange objects or data in groups or classes according to their resemblance and disparities.
- ☞ To enable one to form mental picture of data.
- ☞ To facilitate data presentation.

After collecting and organize data the next important task is effective presentation of bulk volume data. The major objectives of data presentation are:-

- ☞ To presenting data in visual display and more understandable
- ☞ To have great attraction about the data
- ☞ To facilitate quick comparisons using measures of location and dispersion.
- ☞ To enable the reader to determine the shape and nature of distribution to make statistical inference, and to facilitate further statistical analysis.

Methods of data presentation, namely: Tables, Diagrams, and Graphs are commonly used to summarize both qualitative and quantitative data.

2.3.1. Tabular presentation of data

Tables are important to summarize large volume of data in more understandable way. Based on the characteristics they present tables are:

i. Simple (one way table): table which present one characteristics for example age distribution.
ii. Two way table: it presents two characteristics in columns and rows for example age versus sex.
iii. A higher order table: table which presents two or more characteristics in one table.

In statistics usually we use frequency distribution table for different type of data, however sometimes we use the above tables to summarize raw data and to present results of statistical analysis.

Frequency Distribution: is the table that shows data classified in to a number of classes according to some criteria with corresponding frequencies; in other words a frequency distribution is the organization of raw data in table form, using classes and frequencies. Where, Frequency is the number of values in a specific class of the distribution.

There are three basic types of frequency distributions, and there are specific procedures for constructing each type. The three types are categorical, ungrouped and grouped frequency distributions.

A. Categorical Frequency Distribution

The *categorical frequency distribution* is used for data which can be placed in specific categories such as *nominal* or *ordinal level* data. For example, for data such as political affiliation, religious affiliation, blood type, or major field of study categorical frequency distribution is appropriate.

The major components of categorical frequency distribution are *class*, *tally* and *frequency (or proportion).*

Steps of constructing categorical frequency distribution

1. You have to identify that the data is in nominal or ordinal scale of measurement

A	B	C	D
Class	Tally	Frequency	Percent

2. Make a table as show below
3. Put distinct values of a data set in column A
4. Tally the data and place the result in column B
5. Count the tallies and place the results in column C
6. Find the percentage of values in each class by using the formula $\frac{f}{n} x100\%$

   Where, f is frequency, and n is total number of values.

Example 2.1: Twenty-five army inductees were given a blood test to determine their blood type. The data set is given as follows:

A	B	B	AB	O
O	O	B	AB	B
B	B	O	A	O
A	O	O	O	AB
AB	A	O	B	A

Construct a frequency distribution for the above data.

Solution

Our data type is nominal so we use categorical frequency distribution type to present data. After we have followed the above six steps the following frequency distribution was obtained.

A	B	C	D
Class	Tally	Frequency	Percent
A	////	5	20
B	//// //	7	28
O	//// ////	9	36
AB	////	4	16

B. Ungrouped Frequency Distribution

It is a table of all the potential raw values that could possible occurs in the data along with the number of times each actually occurred. In other words ungrouped frequency distribution is the distribution that use individual data values along with their frequencies. It is often constructed for small set of data on discrete variable (when data are numerical), and when the range of the data is small. However, sometimes it is complicate to use ungrouped frequency distribution for large mass of data, as result we use grouped frequency distribution.The major components of this type of frequency distributions are class, tally, frequency, and cumulative frequency.

Cumulative frequencies (CF):- are used to show how many values are accumulated up to and including a specific class. We have less than and more than cumulative frequencies.

*Less than Cumulative Frequency (LCF):-*is the total sum of observations below specified class including that class

More than Cumulative frequency (MCF):- is the total sum of observations above specified class including that class.

Steps of constructing ungrouped frequency distribution

The steps are almost similar with that of categorical frequency distribution

- First find the smallest and largest raw score in the collected data.
- Arrange the data in order of magnitude and count the frequency.
- Put respective frequency, relative frequency, and cumulative frequency along each ordered data.

Example 2.2 A demographer is interested in the number of children a family may have, he/she took sample of 30 families and obtained the following observations.

4	2	4	3	2	8
3	4	4	2	2	8
5	3	4	5	4	5
4	3	5	2	7	3
3	6	7	3	8	4

Construct a frequency distribution for this data.

Solution

These individual observations can be arranged in ascending or descending order of magnitude in which case the series is called array. Array of the number of children in 30 families is:

2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 6, 7, 7, 8, 8, 8

And the variable" number of children in a family" can assume only the values 0, 1, 2, 3, . . . which are discrete; therefore, we construct ungrouped frequency distribution consists of individual data, frequencies, and cumulative frequencies.

Frequency distribution of children in a 30 families is as follow:

No of children (Class)	**No of family (Frequency)**	**LCF**	**MCF**
2	5	5	30
3	7	12	25
4	8	20	18
5	4	24	10
6	1	25	6
7	2	27	5
8	3	30	3

Each individual value is presented separately, that is why it is named ungrouped frequency distribution

C. Grouped Frequency Distribution

It is a frequency distribution when several numbers are grouped in one class; the data must be grouped in which each class has more than one unit in width. We use this type of frequency distribution when the range of the data is large, and for data from continuous variable. But, sometimes for large volume of discrete we use grouped frequency distribution.

Some of basic terms that are most frequently used while we deal with grouped frequency distribution are

- *Lower Class Limits* are the smallest number that can belong to the different class.
- *Upper Class Limits* are the largest number that can belong to the different classes.

- *Class Boundaries (true class limits)* are the number used to separate classes, but without the gaps created by class limits.
- *Class midpoints* are the midpoints of the classes. Each class midpoint can be found by adding the lower class limit/boundary to the upper class limit boundary and dividing the sum by 2.
- *Class width* is the difference between two consecutive lower class limits or two consecutive lower class boundaries.

Steps in constructing grouped frequency distribution

1. Find the highest and the lowest values

2. Find the range; $Range = Maximum - Minimum$ or $R = H - L$

3. Select the number of classes desired. Here, we have two choices to get the desired number of classes:

i. Use Struge's rule. That is, $K = 1 + 3.32 \log n$ where K is the number of class and n is the number of observations. Round the decimal to the upper nearest integer.

ii. Select the number of classes arbitrarily between 5 and 20 conventionally. If you fail to calculate K by Struge's rule, this method is more appropriate.

When we choose the number of classes, we have to think about the following criteria

- ***The classes must be mutually exclusive***. Mutually exclusive classes have non overlapping class limits so that values can't be placed in to two classes.
- ***The classes must be continuous***. Even if there are no values in a class, the class must be included in the frequency distribution. There should be no gaps in a frequency distribution. The only exception occurs when the class with a zero frequency is the first or last. A class width with a zero frequency at either end can be omitted without affecting the distribution.
- ***The classes must be equal in width***. The reason for having classes with equal width is so that there is not a distorted view of the data. One exception occurs when a distribution is open-ended. i.e., it has no specific beginning or end values.

4. Find the class width (W) by dividing the range by the number of classes

$$W = \frac{R}{K} \text{ or } W = \frac{Range}{Number\ of\ Classes}$$

Note that: Round the value of W up to the nearest whole number if there is a reminder. For instance, $4.7 \approx 5$ and $4.12 \approx 5$.

5. Select the starting point as the lowest class limit. This is usually the lowest score (observation). Add the width to that score to get the lower class limit of the next class. Keep adding until you achieve the number of desired classes (K) calculated in step 3.

6. Find the upper class limit; subtract unit of measurement(U) from the lower class limit of the second class in order to get the upper class limit of the first class. Then add the width to each upper class limit to get all upper class limits. Take care of the last class to include the maximum value of data.

Unit of measurement*:* Is the smallest value of difference between consecutive observations or sometimes it is next value. For instance, 28, 23, 52, and then the unit of measurement of this data set is one. Because take one datum arbitrarily, say 23, then the next value will be 24. Therefore, $U = 24 - 23 = 1$. If the data set is 24.12, 30, 21.2, then give *priority* to the datum with more decimal place. Take 24.12 and guess the next possible value. It is 24.13. Therefore, U = 24.12 - 24.13 = 0.01

Note that: U=1 is the maximum value of unit of measurement and is the value when we don't have a clue about the data.

7. Find the class boundaries. $Lower\ Class\ Boundary = Lower\ Class\ Limit - \frac{U}{2}$ and

$Upper\ Class\ Boundary = Upper\ Class\ Limit - \frac{U}{2}$.

In short, $\text{LCB} = \text{LCL} - \frac{\text{U}}{2}$ and $\text{UCB} = \text{UCL} + \frac{\text{U}}{2}$.

8. Tally the data and write the numerical values for tallies in the frequency column.

9. Find cumulative frequency. We have two type of cumulative frequency namely *less than cumulative frequency* and *more than cumulative frequency*. Less than cumulative frequency is obtained by adding successively the frequencies of all the previous classes including the class against which it is written. The cumulate is started from the lowest to the highest size. More than cumulative frequency is obtained by finding the cumulate total of frequencies starting from the highest to the lowest class.

For example, the following frequency distribution table gives the marks obtained by 40 students:

Class marks	Frequency	Cumulative frequency
0 - 10	4	4
10 - 20	5	9 = 5 + (4)
20 - 30	12	21 = 12 + (4 + 5)
30 - 40	11	32 = 11 + (4 + 5 +12)
40 - 50	8	40 = 8 + (4 + 5 + 12 + 11)

The above table shows how to find less than cumulative frequency and the table shown below shows how to find more than cumulative frequency.

Class marks	Frequency	Cumulative Frequency
0 - 10	4	40 = 4 + (5 + 12 + 11 + 8)
10 - 20	5	36 = 5 + (12 + 11 + 8)
20 - 30	12	31 = 12 + (11 + 8)
30 - 40	11	19 = 11 + (8)
40 - 50	8	8

Example 2.3*:* Consider the following set of data and construct the frequency distribution.

11 29 6 33 14 21 18 17 22 38
31 22 27 19 22 23 26 39 34 27

Solution:

Using steps to construct grouped frequency distribution

1. *Highest value=39, Lowest value=6*
2. $R = 39 - 6 = 33$
3. $K = 1 + 3.32 \log 20 = 5.32 \approx 6$
4. $W = \frac{R}{K} = \frac{33}{6} = 5.5 \approx 6$

5. Select starting point. Take the minimum which is 6 then add width 6 on it to get the next class LCL.

6	12	18	24	30	36

6. Upper class limit. Since unit of measurement is one. $12-1=11$. So 11 is the UCL of the first class. Therefore, $6-11$ is the first class

Class limit	6-11	12-17	18-23	24-29	30-35	36-41

7. Find the class boundaries. Take the formula in step 7. $LCB_i = LCL_i - 0.5$ and $UCB_i = UCL_i + 0.5$

Class Boundary	5.5-11.5	11.5-17.5	17.5-23.5	23.5-29.5	29.5-35.5	35.5-41.5

With respective class marks of

Class marks	8.5	14.5	20.5	26.5	32.5	38.5

8. 9 and 10

Class Limit	Class Boundary	f	Less than CF	More than CF
6-11	5.5-11.5	2	2	20=2+(2+7+4+3+2)
12-17	11.5-17.5	2	2+2=4	18=2+(7+4+3+2)
18-23	17.5-23.5	7	2+2+7=11	16=7+(4+3+2)
24-29	23.5-29.5	4	2+2+7+4=15	9=4+(3+2)
30-35	29.5-35.5	3	2+2+7+4+3=18	5=3+2
36-41	35.5-41.5	2	2+2+7+4+3+2=20	2

Example 2.4 The following data are percentage coverage of forest in countries in Africa. Construct frequency distribution by using sturge's rule.

30, 25, 23, 41, 39, 27, 41, 24, 32, 29, 35, 31, 36, 33, 36, 42, 35, 37, 41, and 29

Solution

1. Given no. of observation(n) = 20,then no. of classes

$$K = 1 + 3.32 \log_{10}^{20} \cong 5 \text{ , where k is number of classes.}$$

2. Class width(W) $= \frac{highestvalue - lowestvalue}{k} = \frac{42 - 23}{5} \cong 4$

Classes	Class boundary	Class mark	Frequency
23 - 26	22.5 –26.5	24.5	3
27 - 30	26.5 – 30.5	28.5	4
31 - 34	30.5 – 34.5	32.5	3
35 - 38	34.5 – 38.5	36.5	5
39 - 42	38.5 – 42.5	40.5	5
Total			20

Relative Frequency Distribution

An important variation of the basic frequency distribution uses relative frequencies, which are easily found by dividing each class frequency by the total of all frequencies. A relative frequency distribution includes the same class limits as a frequency distribution, but relative frequencies are used instead of actual frequencies. The relative frequencies are sometimes expressed as percents.

$$\text{Relative Frequency} = \frac{\text{Class Frequency}}{\text{Sum of all frequencies}}$$

Relative frequency distribution enables us to understand the distribution of the data and to compare different sets of data.

Example 2.5: Relative Frequency and Percentage Table

Class interval	Frequency	Relative Frequency	Percentage
10 but less than 20	3	0.055	5.5
20 but less than 30	5	0.091	9.1
30 but less than 40	9	0.164	16.4
40 but less than 50	18	0.32 7	32.7
50 but less than 60	10	0.182	18.2
60 but less than 70	8	0.145	14.5
70 but less than 80	2	0.036	3.6
Total	**55**	**1.000**	**100**

It may be noted that at times the use of relative frequencies is more appropriate than absolute frequencies. Whatever two or more sets of data contain different number of observation, a comparison with absolute frequencies will be incorrect. In such cases, it is necessary to use the relative frequency

2.3.2. Diagrammatic presentation of the data

We have discussed the techniques of classification and tabulation that help us in organizing the collected data in a meaningful fashion. However, this way of presentation of statistical data does not always prove to be interesting to a layman. Too many figures are often confusing and fail to convey the massage effectively.One of the most effective and interesting alternative way in which a statistical data may be presented is through diagrams and graphs. There are several ways in which statistical data may be displayed pictorially such as different types of graphs and diagrams.

General steps in constructing diagram or graphs

1. Draw and label the x and y axes

2. Choose a suitable scale for the frequencies or cumulative frequencies and label it on the y axis.

3. Represent the class boundaries for the histogram or Ogive or the midpoint for the frequency polygon on the *x* axis.

4. Plot the points, and draw the bars or lines

Most of time we use diagram form of presentation for categorical and numerical data, some of the diagrams are as follow:

A. Pie Chart

Pie chart can used to compare the relation between the whole and its components. Pie chart is a circular diagram and the area of the sector of a circle is used in pie chart. Circles are drawn with radii proportional to the square root of the quantities because the area of a circle is πr^2. To construct a pie chart (sector diagram), we draw a circle with radius (square root of the total). The total angle of the circle is 360^0.

The angles of each component are calculated by the formula

$Angle\ of\ Sector = \frac{Component\ Part}{Total} x360^0$ These angles are made in the circle by mean of a protractor to show different components. The arrangement of the sectors is usually anti-clock wise.

Example2.6: The following table gives the details of monthly budget of a family. Represent these figures by a suitable diagram.

Item of Expenditure	Family Budget
Food	$ 600
Clothing	$ 100
House Rent	$ 400
Fuel and Lighting	$ 100
Miscellaneous	$ 300
Total	$ 1500

Solution: The necessary computations are given below:

Items	Family Budget		
	Expenditure $	Angle of Sectors	Percent
Food	600	144°	40
Clothing	100	24°	6.67
House Rent	400	96°	26.67
Fuel and Lighting	100	24°	6.67
Miscellaneous	300	72°	20
Total	1500	360°	100

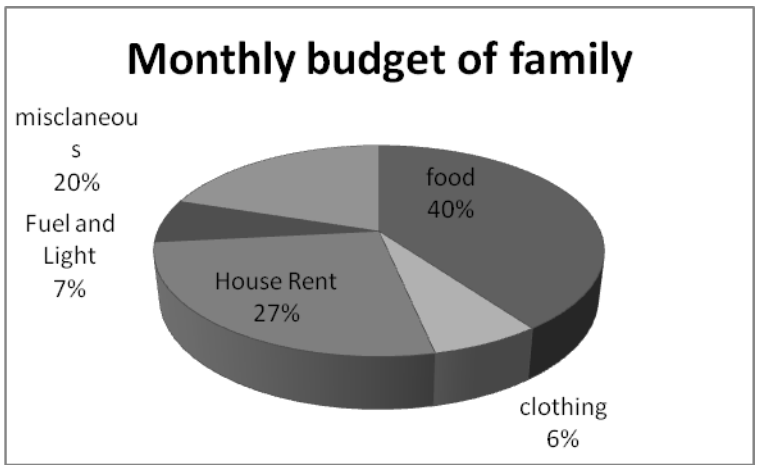


Figure 2.1 pie chart of monthly budget of a family

B. Bar Charts

The bar charts (simple bar chart, multiple bar charts, and stratified or stacked bar chart) use vertical or horizontal bins to represent the frequencies of a distribution. While we draw bar chart, we have to consider the following two points. These are

- Make the bars the same width
- Make the units on the axis that are used for the frequency equal in size

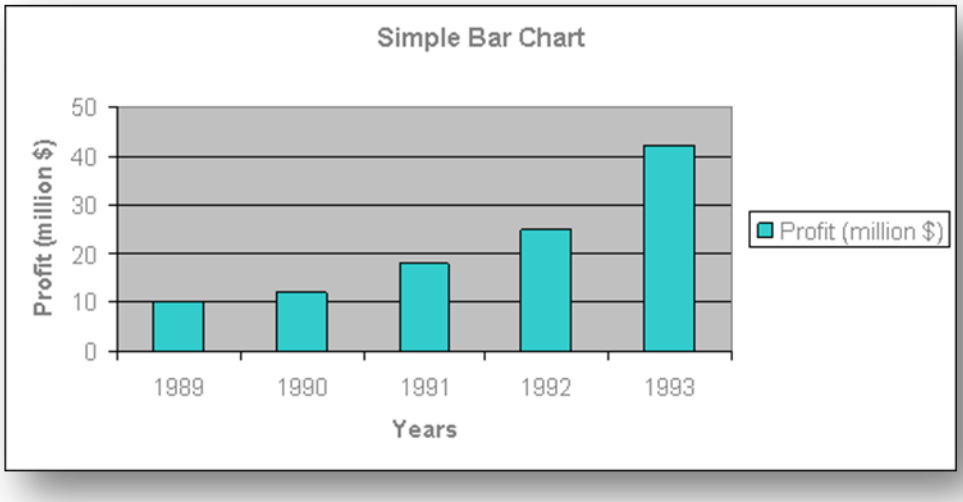
i. **Simple bar chart** is used to represents data involving only one variable classified on spatial, quantitative or temporal basis.In simple bar chart, we make bars of equal width but variable length, i.e. the magnitude of a quantity is represented by the height or length of the bars.The following steps are undertaken in drawing a simple bar diagram:

- Draw two perpendicular lines one horizontally and the other vertically at an appropriate place of the paper.

- Take the basis of classification along horizontal line (X-axis) and the observed variable along vertical line (Y-axis) or vice versa.
- Marks signs of equal width for each class and leave equal or not less than half width in between two classes.
- Finally, marks the values of the given variable to prepare required bars.

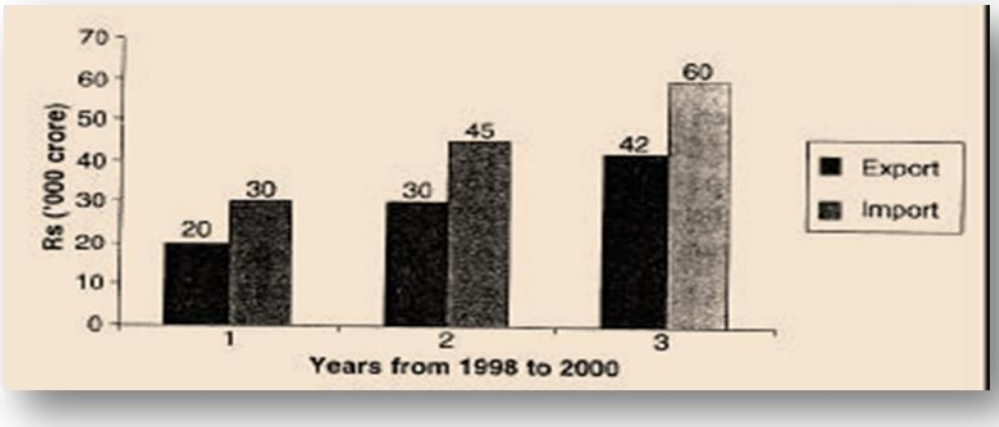
Example 2.7: Draw simple bar diagram to represent the profits of a bank for 5 years.

Years	1989	1990	1991	1992	1993
Profit (million $)	10	12	18	25	42



ii. Multiple Bars

When two or more interrelated series of data are depicted by a bar diagram, then such a diagram is known as a multiple-bar diagram. Suppose we have export and import figures for a few years. We can display by two bars close to each other, one representing exports while the other representing imports figure shows such a diagram based on hypothetical data.



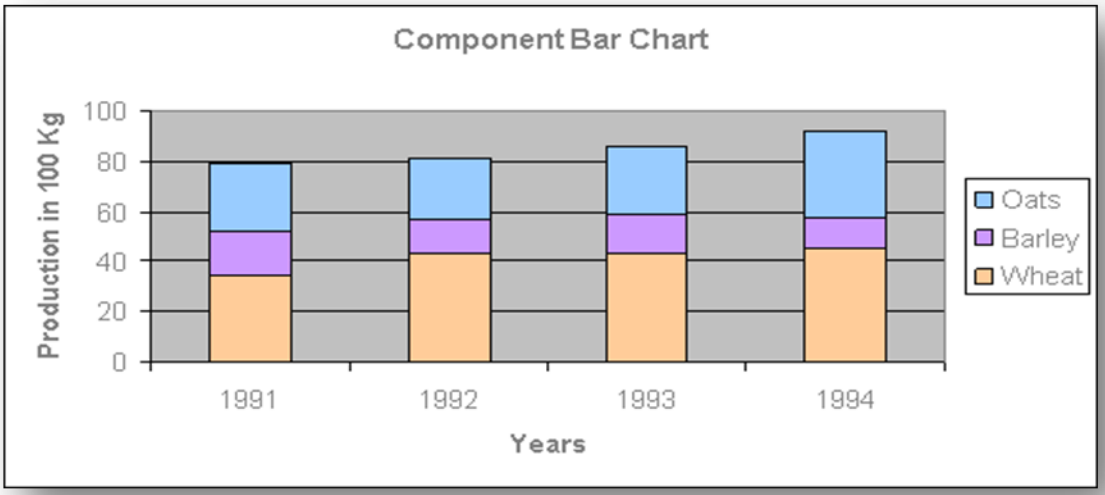
Multiple bar chart should be noted that multiple bar diagrams are particularly suitable where some comparison is involved.

iii. ***Stratified (Stacked) Bar Chart****is used to represent data in which the total magnitude is divided into different or components.* In this diagram, first we make simple bars for each class taking total magnitude in that class and then divide these simple bars into parts in the ratio of various components. This type of diagram shows the variation in different components within each class as well as between different classes. Stratified bar diagram is also known as component bar chart.

Example 2.8: The table below shows the quantity in hundred Kgs of Wheat, Barley and Oats produced on a certain farm during the years 1991 to 1994. Draw stratified bar chart.

Years	Wheat	Barley	Oats	Total
1991	34	18	27	79
1992	43	14	24	81
1993	43	16	27	86
1994	45	13	34	92

Solution: to make the component bar chart, first of all we have to take year wise total production.The required diagram is given below:



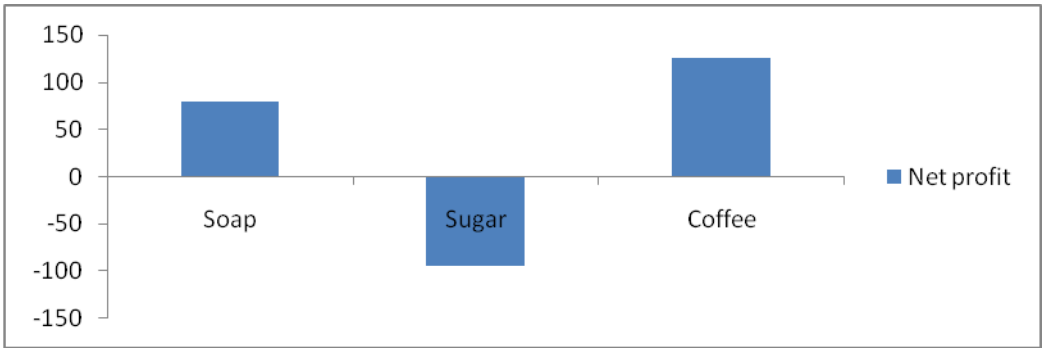
iv. Deviation Bar Diagram:

When the data contains both positive and negative values such as data on net profit, net expense, percent change etc.

***Example 2.9*:**Suppose we have the following data relating to net profit (percent) of commodity.

Commodity	Net profit
Soap	80
Sugar	-95
Coffee	125

Solution:



2.3.3. Graphical Presentation of data

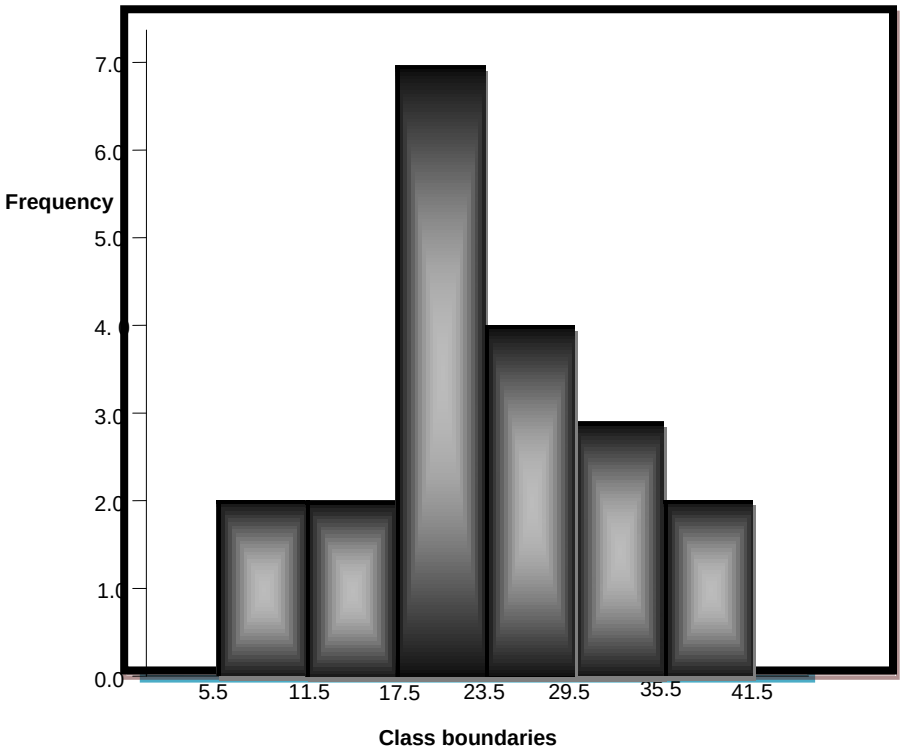
Often we use graphical presentation form for continuous data type; results from the grouped frequency distribution and continuous variables distributed over time.

A. Histogram

Histogram is a special type of bar graph in which the horizontal scale represents classes of data values and the vertical scale represents frequencies. The height of the bars correspond to the frequency values, and the drawn adjacent to each other (without gaps).

We can construct a histogram after we have first completed a frequency distribution table for a data set. TheX axis is reserved for the class boundaries.

Example2.9: The histogram for the data in *example 2.4* is

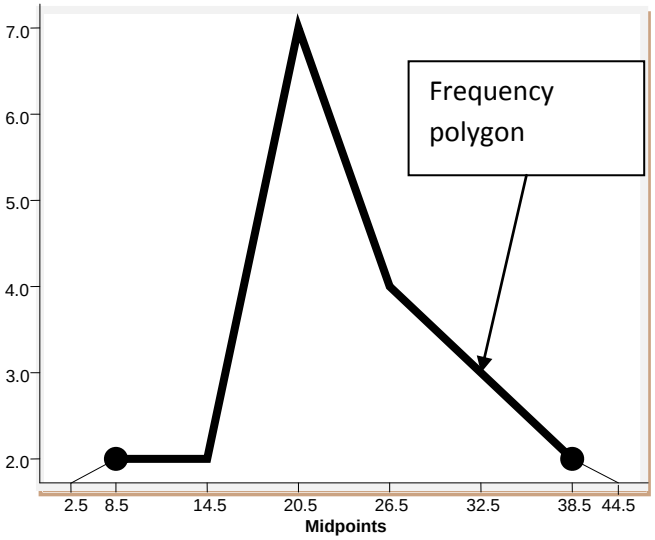


Relative frequency histogram has the same shape and horizontal (x *axis*) scale as a histogram, but the vertical (y *axis*) scale is marked with *relative frequencies* instead of actual frequencies.

B. Frequency Polygon

A *frequency polygon* uses line segment connected to points located directly above class midpoint values. The heights of the points correspond to the class frequencies, and the line segments are extended to the left and right so that the graph begins and ends on the horizontal axis with the same distance that the previous and next midpoint would be located.

Example 2.10:The frequency polygon for the data in *example 2.4* is given as follow.



C. Ogive Graph

An *Ogive* (pronounced as "oh-jive") is a line that depicts *cumulative* frequencies, just as the cumulative frequency distribution lists cumulative frequencies. Note that the Ogive uses class boundaries along the horizontal scale, and graph begins with the lower boundary of the first class and ends with the upper boundary of the last class. Ogive is useful for determining the number of values below some particular value. There are two type of Ogive namely *less than Ogive* and *more than Ogive*. The difference is that less than Ogive uses less than cumulative frequency and more than Ogive uses more than cumulative frequency on y axis.

Example 2.10: Take the data in *example 2.4* and draw less than and more than Ogive curve.

Chapter Three

Measures of Central Tendency (MCT)

3.1. Measure of Central Tendency

Suppose the students from two or more classes appeared in the examination and we wish to compare the performance of the classes in the examination or wish to compare the performance of the same class after some coaching over a period of time. When making such comparisons, it is not practicable to compare the full frequency distributions of marks. For such statistical analysis, we need a single representative value that describes the entire mass of data given in the frequency distribution. This single representative value is called the central value, measure of location or an average around which individual values of a series cluster. This central value or an average enables us to get a gist of the entire mass of data, and its value lies somewhere in the middle of the two extremes of the given observations.

It should be clear to you that the concept of a measure of central tendency is concerned only with quantitative variables and is undefined for qualitative variables as these are immeasurable on a scale.

The three major objectives of measures of central tendency are

- To summarize a set of data by single value
- To facilitate comparison among different data sets
- To use for further statistical analysis or manipulation

3.2. *Summation Notation*

1. $\sum (sigma)$ is used to facilitate the writing of sum

2. $\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots + x_n$

3. $\sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n$

4. $\sum_{i=1}^{n} (x_i + y_i) = (x_1 + y_1) + (x_2 + y_2) + \ldots + (x_n + y_n) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$

5. $\sum_{i=1}^{n} CXi = CX_1 + CX_2 + CX_3 + \ldots + CX_n = C\ (X1 + X_2 + X_3 + \ldots + X_n) = C\sum_{i=1}^{n} Xi$

6. $\sum_{i=1}^{n} C = \underbrace{C + C + C \ldots . + C}_{} = nc$

7. $\sum_{i=1}^{n} (x_i + c) = (x_1 + c) + (x_2 + c) + \ldots (x_n + c)$

$= x_1 + x_2 + \ldots + x_n + c + \ldots + c$

$= \sum_{i=1}^{n} xi + \sum_{i=1}^{n} C = \sum_{i=1}^{n} x_i + nc$

N.B $\sum_{i=1}^{n} x_i^{\,2} \neq \left(\sum_{i=1}^{n} x_i\right)^2$ and $\left(\sum x_i\ y_i\right) \neq \left(\sum x_i\right)\left(\sum y_i\right)$

3.3. Desirable Properties of Good Measures of Central Tendency

1. It should be easy to calculate and understand.
2. It should be based on all the observations during computation.
3. It should be rigidly defined. The definition should be clear and unambiguous so that it leads to one and only one interpretation by different persons.
4. It should be representative of the data, if it's from sample. Then the sample should be random enough to be accurate representative of the population**.**
5. It should have sampling stability. It shouldn't be affected by sampling fluctuations. This means that if we pick (take) two independent random samples of the same size from a given population and compute the average for each of these samples then the value obtained from different samples should not vary much from one another.
6. It shouldn't be affected by the extreme value. If a few very small and very large items are presented in the data, they will influence the value of the average by shifting it to one side or of other side and hence the average chosen should be such that is not influenced by the extreme values.

3.4. Types of Measures of Central Tendency

In statistics, we have various types of measures of central tendencies. The most commonly used types of MCT includes:-

-Mean - Mode

- Median -Quantiles (Quartiles, Decilies, Percentiles)

3.4.1. Mean

Mean is obtained for sample and population data. There are four type of mean which is suitable for a particular type of data.

i. Arithmetic mean (A.M)
ii. Geometric mean (G.M)
iii. Harmonic mean (H.M)
iv. Weighted mean (W.M)

i. Arithmetic Mean ($\bar{X}$)

In classification and presentation of data, we observed that the values of the variable or observations could be put in the form of raw data (individual data), simple (ungrouped) frequency distribution, and grouped frequency distribution.

a) Arithmetic Mean from raw(individual) data

Let X be a variable which takes valuesx_1 ,x_2 ,x_3 ,……,x_n, in a sample size of n from a population of size N for n < N then A.M of a set of observations is the sum of all values in a series divided by the number of items in the series.

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{1}{n}\sum X_i \text{ for sample data}$$

Example:3.1Suppose he scores of a student on six examinations were 10 ,15,30,7,42 , 79 and 83 ,find the arithmetic mean of scores of students.

Solution: there are seven observations (n=7). Symbolically, the arithmetic mean is:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = (10 + 15 + 30 + 7 + 42 + 79 + 83) / 7 = 266 / 7 = 38$$

Example 3.2: The following data are on age of 20 household heads which are obtained during water distribution survey in certain town. Find the average age of household heads.

30, 25, 23, 41, 39, 27, 41, 24, 32, 29, 35, 31, 36, 33, 36, 42, 35, 37, 41, and 29

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{30+25+\ldots+29}{20} = \frac{666}{20} = 33.3$$

Interpretation: the average age of 20 household heads is 33.3 years

NB*: If mean is mentioned, it implies arithmetic mean, but other means are identified by their name.*

It may be noted that the Greek letter μ is used to denote the mean of the population and N to denote the total number of observations in a population. Thus, the population mean

$$\mu = \frac{X_1 + X_2 + \ldots + X_N}{N} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

b) Arithmetic mean from discrete (ungrouped) frequency distribution

In discrete frequency distribution we multiply the values of the variable (X) by their respective frequencies (f) and get the sum of the products ($\sum fXi$). The sum of the products is then divided by the total of the frequencies ($\sum fi = n$).

Values (Xi)	***Frequency(fi)***	***Xi*fi***
X_1	f_1	$x_1 f_1$
X_2	f_2	$x_2 f_2$
.	.	
X_i	f_i	
.	.	
X_k	f_k	$x_k f_k$
	$\sum f_i$	$\sum x_i f_i$

Thus, according to this method, the formula for calculating arithmetic mean for simple /discrete/ frequency distributions becomes: $\overline{X} = \frac{\sum fXi}{\sum fi} = \frac{\sum fXi}{n}$ **where**, $n = \sum fi$

Example 3.3:Following table gives the wages paid to 125 workers in a factory. Calculate the arithmetic mean of the wages.

Wages (in birr):	200	210	220	230	240	250	260
No. of workers:	5	15	32	42	15	12	4

Wages(x)	**No. of Workers(f)**	**fx**
200	5	1000
210	15	3150
220	32	7040
230	42	9660
240	15	3600
250	12	3000
260	4	1040
Total	$N = \sum f = 125$	$\sum fX = 28490$

Solution:using above results in the table we substitute into formula and obtain

$$\overline{X} = \frac{\sum fx}{N} = \frac{28490}{125} = 227.92 birr$$

Interpretation: the average wage paid to 125 workers in the factory is 227.92 birr

c) Arithmetic mean for continuous (grouped) frequency distribution

Arithmetic mean for continuous frequency distribution is given by:

$$\overline{X} = \frac{1}{\sum_{i=1}^{k} f_i} \left(\sum_{i=1}^{n} f_i \; m_i \right)$$, where m_i = midpoint of each ith class interval, k is no. of class

Example 3.4:The following table gives the marks of 58 students in probability and Statistics.

Calculate the average marks of these students.

Marks	**Number of students**
0-10	4
10-20	8
20-30	11
30-40	15
40-50	12
50-60	6
60-70	2

Solution

Marks	**Mid-point (mi)**	**No. of Students (fi)**	$\mathbf{f_i * m_i}$
0-10	5	4	20
10-20	15	8	120
20-30	25	11	275
30-40	35	15	525
40-50	45	12	540
50-60	55	6	330
60-70	65	2	130

$$\sum fi = 58 \sum f_i m_i = 1940$$

So, Arithmetic mean will be $\bar{X} = \frac{1}{\sum_{i=1}^{n} f_i} \left(\sum_{i=1}^{n} f_i \; m_i \right)$ = 1940/58 = 33.45.

Properties of arithmetic mean

1. It is easy to calculate and understand.
2. All observation involved in its calculation.
3. It cannot be computed for open end classes
4. It may not be the values which the variable actually takes and termed as a fictitious (unreal) average. E.g. The figure like on average 2.21 children per family, 3.4 accidents per day.
5. It is affected by extreme values.
6. It is Unique: - a set of data has only one mean.
7. If a constant k is added or subtracted from each value of a distribution, then the new mean for the new distribution will be the original mean plus or minus k, respectively.
8. The sum of the deviation of various values from their mean is zero i.e. $\sum (x_i - \bar{x}) = 0$
9. The sum of the squares of deviation of the given set of observations is minimum when taken from the arithmetic mean i.e. $\sum (x_i - A)^2 \rightarrow$ is minimum when taken from mean than any arbitrary value A from a set of observation.
10. It can be used for further statistical treatment, comparison of means, test of means.
11. When a set of observations is divided into *k* groups and $\bar{x}_1$ is the mean of n_1 observations of group 1, $\bar{x}_2$ is the mean of n_2 observations of group2, …, $\bar{x}_k$ is the mean of n_k observations of group *k* , then the combined mean ,denoted by $\bar{x}_c$, of all observations taken together is given by

$$\bar{X}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_k\bar{x}_k}{n_1 + n_2 + \cdots + n_k} = \frac{\sum_{i=1}^{k} n_i\bar{x}_i}{\sum_{i=1}^{k} n_i}$$

12. If a wrong figure has been used in calculating the mean, we can correct if we know the correct figure that should have been used. Let

- X_{wr} denote the wrong figure used in calculating the mean
- X_c be the correct figure that should have been used
- $\bar{X}_{wr}$ be the wrong mean calculated using X_{wr} , then the correct mean,then $\bar{X}_{correct}$, is given by

$$\bar{X}_{correct} = \frac{n\bar{X}_{wr} + X_c - X_{wr}}{n}$$

Example 3.5*:* Last year there were three sections taking Stat 1044 course in Arba Minch University. At the end of the semester, the three sections got average marks of 80, 83 and 76. If there were 28, 32 and 35 students in each section respectively. Find the mean mark for the entire students.

Solution: here we need to calculate combined mean

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3}{n_1 + n_2 + n_3} = \frac{28(80) + 32(83) + 35(76)}{28 + 32 + 35} = \frac{7556}{95} = \underline{79.54}$$

Example 3.6: An average weight of 10 students was calculated to be 65 kg, but latter, it was discovered that one measurement was misread as 40 kg instead of 80 kg. Calculate the corrected average weight.

Solution: $\bar{X}_{correct} = \frac{n\bar{X}_{WI} + X_C - X_{WI}}{n} = \frac{10(65)+80-40}{10} = 69$

Exercise*:* The average score on the mid-term examination of 25 students was 75.8 out of 100. After the mid-term exam, however, a student whose score was 41 out of 100 dropped the course. What is the average/mean score among the 24 students?

ii. Geometric Mean (G.M)

The geometric mean is usually used in average rates of change, ratios, percentage distribution, and logarithmical distribution.

The geometric mean is the n^{th} root of the product of n positive values.

a) Geometric mean for raw (individual) data series

If $X1, X_{2,\ldots,}X_n$ are n positive values, then their geometric mean is $G.M = (X_1X_{2\ldots}X_n)^{1/n}$.

b) Geometric mean for simple (discrete) frequency distribution

When the observed values $x_1, x_2, \ldots\ldots\ldots x_n$ have the corresponding frequencies $f_1.f_2\ldots\ldots\ldots f_n$ respectively then geometric mean is obtained by

$$G.M = \sqrt[n]{x_1^{f_1}, x^{f_2}{}_2 \ldots . x^{f_n}{}_n}$$

$$= \frac{1}{n}\sum_{i=1}^{n} f_i \log x_i \qquad \text{where,} \quad n = \sum_{i=1}^{n} f_i$$

c) Geometric mean for continuous (grouped) frequency distribution

Whenever the frequency distributions are grouped (continuous), class marks of the class interval are considered as $X_{i,}$ and the above formula can be used as:

$$G.M = \sqrt[n]{m_1^{f_1}, m_2^{f_2} \ldots . m_n^{f_n}}$$

$$= \frac{1}{n}\sum_{i=1}^{n} f_i \log m_i \quad \text{where} \quad n=\sum_{i=1}^{n} f_i \quad \text{and} \quad m_i \text{ is class mark of } i^{th}\text{class.}$$

Note: in case of number of observation is more than two it may be tedious taking out from square root, in that case calculation can be simplified by taking natural logarithm with base 10.

$$G.M = \sqrt[n]{x_1, x_2 \ldots . x_n} \qquad G.M = (x_1 \ldots . x_n)^{\frac{1}{n}} \quad \text{take log in both sides.}$$

$$\log(G.M) = \frac{1}{n}\log(x_1, \ldots x_n) = \frac{1}{n}(\log x_1 + \log x_2 + \ldots + \log x_n) = \frac{1}{n}\sum_{i=1}^{n} \log x_i$$

$$G.M = \text{Antilog}\left[\frac{1}{n}\sum_{i=1}^{n} \log x_i\right]$$

This shows that the logarithm of G. M is the mean of the logarithms of individual observations. In similar manner we compute for ungrouped and grouped frequency distributions.

Example3.7: The ratios of prices in 1999 to those in 2000 for 4 commodities were 0.9, 1.25, 1.75 and 0.85. Find the average price ratio.

Solution $G.M = \text{antilog}\dfrac{\sum \log X_i}{n} = \text{antilog}\dfrac{(\log 0.92 + \log 1.25 + \log 1.75 + \log 0.85)}{4}$

$$= \text{antilog}\frac{(0.963 - 1 + 0.0969 + 0.2430 + 0.9294 - 1)}{4} = \text{antilog}0.5829 = 1.14$$

What is the arithmetic mean of the above values? $\overline{X} = \dfrac{0.92 + 1.25 + 1.75 + 0.85}{4} = 1.19$

This is quite different from the geometric mean in case of ratio data.

Properties of geometric mean

a. Its calculations are not as such easy.
b. It involves all observations during computation
c. It may not be defined even it a single observation is negative.
d. If the value of one observation is zero its values becomes zero.
e. It is affected by sample fluctuations.

iii. Harmonic mean (H.M)

The harmonic mean is used to average rates rather than simple values. It is usually appropriate in averaging kilometers per hour, speed, and rates at time. The Harmonic mean is the reciprocal of the arithmetic mean of the reciprocal of the single values

a) Harmonic mean for raw data

If $X_1, X_2, X_3, \ldots, X_n$ are n values, then their harmonic mean is

$$H.M = \frac{n}{\frac{1}{X_1}+\frac{1}{X_2}+\ldots+\frac{1}{X_n}} = \frac{n}{\sum\frac{1}{X_i}}$$

b) Harmonic mean for ungrouped frequency distribution

$$H.M = \frac{n}{\frac{f_1}{X_1}+\frac{f_2}{X_2}+\ldots+\frac{f_k}{X_k}} = \frac{n}{\sum\frac{f_i}{X_i}} \text{ where, } n = \sum_{i=1}^{k} f_i$$

c) Harmonic mean for grouped frequency distribution

In calculating harmonic mean for grouped frequency distribution data m_i is replace in place of Xi in above equation under b

$$H.M = \frac{n}{\frac{f_1}{m_1}+\frac{f_2}{m_2}+\ldots+\frac{f_k}{m_k}} = \frac{n}{\sum\frac{f_i}{m_i}} \text{ where, } n = \sum_{i=1}^{k} f_i$$

***Example 3.8*:** Find the harmonic mean of the values 2, 3 &6.

$$\text{H.M} = \frac{3}{1/2+1/3+1/6} = \frac{3}{\frac{3+2+1}{6}} = \frac{3\times 6}{6} = 3$$

***Example 3*.9:** A driver covers the 300km distance at an average speed of 60 km/hr makes the return trip at an average speed of 50km/hr. What is his average speed for total distance?

Solution

Trip	Distance	Average speed	Time taken
1st	300km	60km/hr	5hrs
2nd	300km	50km/hr	6hrs
Total	600km	---------	11hrs

Average speed for the whole distance= $\frac{Total\ dis\tan ce}{Total\ time\ taken}$ =600km/11hrs=54.55km/hr.

Using harmonic formula it is H.M= $\frac{2}{1/60+1/50}$ =600/110=54.55km/hr.

Note that A.M= $\frac{60+50}{2}$ =55km/hr G.M= $\sqrt{60\times 50}$ =54.7km/hr.

In general: A.M ≥G.M≥H.M

Properties of harmonic mean

i. It is based on all observation in a distribution.
ii. Used when a situations where small weight is given for larger observation and larger weight for smaller observation
iii. Difficult to calculate and understand
iv. It is not affected by extreme values and sample fluctuation.
v. Appropriate measure of central tendency in situations where data is in ratio, speed or rate.

iv. Weighted Arithmetic Mean

In finding arithmetic mean and others type of mean for series of data, all items were assumed to be of equally importance (each value in the data set has equal weight). But, when the observations have different weight, we use *weighted average*. Weights are assigned to each item in proportion to its relative importance.

If $x_1, x_2, ..., x_k$ represent values of the items and $w_1, w_2, ..., w_k$ are the corresponding weights, then the weighted arithmetic mean, $(\bar{x}_w)$ is given by

$$\bar{X}_w = \frac{w_1x_1 + w_2x_2 + \cdots + w_kx_k}{w_1 + w_2 + \cdots + w_k} = \frac{\sum_{i=1}^{k} w_i x_i}{\sum_{i=1}^{k} w_i}$$

Example 3.10*:* A student's final mark in Mathematics, Physics, Chemistry and Biology are respectively 82, 80, 90 and 70.If the respective credits received for these courses are 3, 5, 3 and 1, determine the approximate average mark the student has got for one course.

Solution: We use a weighted arithmetic mean, weight associated with each course being taken as the number of credits received for the corresponding course.

x_i	82	80	90	70
w_i	3	5	3	1

Therefore, $\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} = \frac{(3\times 82) + (5\times 80) + (3\times 90) + (1\times 70)}{3+5+3+1} = 82.17$

3.4.2 Median ($\tilde{X}$)

Median is defined as the value of the middle item (or the mean of the values of the two middle items) when the data are arranged in an ascending or descending order of magnitude.

a) Median for raw data

If the *n* values are arranged in ascending or descending order of magnitude, the median is the middle value if *n* is odd. When *n* is even, the median is the mean of the two middle values.

Median $= \left(\frac{n+1}{2}\right)^{th}$ position element if n is odd.

$= \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}$ element if n is even.

Example 3.11: Suppose we have the following series: 15, 19, 21, 7, 33, 25, 18 and 5. We have to first arrange it in either ascending or descending order. These figures are arranged in an ascending order as follows: 5, 7, 10, 15, 18, 19, 21, 25, and 33.

Now as the series consists of odd number of items, to find out the value of the middle item, we use the formula **Median** $= \left(\frac{n+1}{2}\right)^{th}$ element if n is odd.

That is the size of the 5^{th} item is the median which is 18.

Suppose the series consists of one more item, 23. We may, therefore, have to include 23 in the above series at an appropriate place, that is, between 21 and 25. Thus, the series is now 5, 7, 10, 15, 18, 19, 21, 23, 25, and 33. Applying the above formula, the median is the size of 5.5^{th} item. Here, we have to take the average of the values of 5^{th} and 6^{th} item. This means an average of 18 and 19, which gives the median as 18.5.

b) Median for ungrouped frequency distribution

Here median is obtained by using less than cumulative frequency distribution; the median is the value of the variable for which the cumulative frequency is just greater than or equal to $\frac{1}{2}N$, where N = $\sum_{i=1}^{k} f_i$

Example 3.12: find median for the following frequency distribution

X	Frequency	Less than cumulative frequency (LCF)
10	3	3
15	5	8
20	10	18
25	7	25
30	5	30

N=30, N/2=15, The LCF greater than or equal 15 is 18 which is corresponds to 20, thus median is 20.

c) Median for continuous frequency distribution

In the case of a continuous frequency distribution, we first locate the median class by procedures under b. In such a way thatclass for which the cumulative frequency is just greater than or equal to $\frac{1}{2}N$, where $N = \sum_{i=1}^{k} f_i$ Then, the median is calculated by:

$$Median(\tilde{X}) = LCL + \frac{\left[\frac{n}{2} - fc_o\right]w_m}{f_m}$$

Where, fc_o is less than cumulative frequency of the class preceding (one before) the median class, f_m is frequency of the median class, LCL is lower class limit of median class and w_m width of the median class and $n = \sum_{i=1}^{k} f_i$

Note: We may use Lower class boundary (LCb) of the median class instead of LCL.

Example 3.13:consider age data of 20 household heads which are obtained during water distribution survey in certain town:30, 25, 23, 41, 39, 27, 41, 24, 32, 29, 35, 31, 36, 33, 36, 42, 35, 37, 41, and 29.Find median from grouped data.

Classes	Class boundary	Frequency	LCF
23 - 26	22.5-26.5	3	3
27 - 30	26.5-30.5	4	7
31 - 34	30.5-34.5	3	10
35 - 38	34.5-38.5	5	15
39 - 42	38.5 – 42.5	5	20
Total		20	

The median class is (31 - 34) which corresponds to $LCF \geq \frac{1}{2}n$ i.e $10 \geq 10$

$$Median(\tilde{X}) = LCL + \frac{\left[\frac{n}{2} - fc_o\right]w_m}{f_m} = 31 + \frac{\left(n/2 - 7\right)4}{3} = 35$$

Properties of median

- Unlike mode it is unique that is like mean there is only one median for a given set of data.
- Easy to calculate and understand.
- It is not affected by extreme value.
- It's especially used for open ended frequency distribution when median is not found in that class.
- It is not based on all observations.
- It is not suitable for further mathematical treatment.

3.4.3 Mode $(\hat{X})$

The mode is another measure of central tendency. It is the value at the point around which the items are most heavily concentrated or the mode is the measurement that occurs in greatest frequency in data set. A given set of data may have

- One mode – uni model e.g. A=3 ,3,7,6,2,1 $\hat{X}$=3
- Two modes – Bimodal e.g. 10,10,9,9,6,3,2,1 $\hat{X}$= 10 and 9
- More than two modes- multimodal. eg. 5,5,5,6,6,6,8,8,8,2,3,2 $\hat{X}$=5,6,8
- May not exist at all e.g. 1,3,2,4,5,6,7,8 no modal value

a) Mode for raw data

For simple series of data mode is the most frequent value in a certain data set.

Example 3.14: Consider the following series: 8, 9, 11, 15, 16, 12, 15, 3, 7, 15

There are ten observations in the series where 15 occurs maximum number of times. Thus, mode is 15.

b) Mode for ungrouped frequency distribution

In case of discrete frequency distribution, mode is the value of the variable corresponding to the maximum frequency. This method can be used conveniently if there is only one value with the highest concentration of observation.

Example 3.15:Consider the following distribution, and then determine modal value of the distribution.

X	1	2	3	4	5	6	7	8	9
f	3	1	18	25	40	30	22	10	6

The maximum frequency is 40 and corresponding value is 5, therefore, the mode is 5.

c) Mode for grouped frequency distribution

First modal class (a class with maximum frequency) could be found, and then mode for grouped frequency distribution is determined by the following formula:

$$\text{Mode}(\hat{X}) = l_o + \left(\frac{f_1 - f_0}{(f_1 - f_0) + (f_1 - f_2)} \right) w$$

Where: l_o = is the lower value of the class in which the mode lies.

f_1 = is the frequency of the class in which the mode lie.

f_0 = is the frequency of the class preceding the modal class.

f_2 = is the frequency of the class succeeding the modal class.

w = is the class width of the modal class.

While applying the above formula, we should ensure that the class-intervals are uniform throughout. If the class-intervals are not uniform, then they should be made uniform on the assumption that the frequencies are evenly distributed throughout the class. In the case of unequal class-intervals, the application of the above formula will give misleading results.

Example 3.15:Consider Example 3.13, find mode for the frequency distribution.

Here we have two modal classes, the last two class with frequencies of 5. Therefore, we called the distribution is bimodal.

Case 1; $\hat{X} = 34.5 + \left(\frac{5-3}{5-3+5-5}\right) 4 = 38.5$

Case 2; $\hat{X} = 38.5 + \left(\frac{5-5}{5-5+5-0}\right) 4 = 38.5$

3.5 Measure of Location (Quintiles)

Median is the value of the middle item which divide the data in to two equal parts and found by arranging the data in an increasing or decreasing order of magnitude, where quintiles are measures which divides a given set of data in to equal subdivision and are obtained by the same procedure to that of median but data must be arranged only in an increasing order. The most commonly used ones include quartiles, deciles and percentiles.

i. Quartiles

Quartiles are measure which divided the ordered data in to four equal parts and usually denoted by Q_1,Q_2, Q_3, and are obtained after arranging the data in an increasing order. Q_1: is first(lower) quartile or value for which 25% of the observation lies below it, Q2: is second quartile or value for which 50%of the observation lies below or above it, and Q_3: is third (upper) quartile or value for which 75% of the arranged item lies below it or 25% lies above it.

a) Quartiles for raw data

For series of data the i^{th} quartiles is the value of the items which is at the

$i*\left(\frac{n+1}{4}\right)^{th}$ position item i.e $Qi = i*\left(\frac{n+1}{4}\right)^{th}$ *position* item where, i=1, 2, 3

$Q1$- is value corresponding to $\left(\frac{n+1}{4}\right)^{th}$ ordered observation.

Q_2 is the value of $2\left(\frac{n+1}{4}\right)^{th}$ ordered observation

Q_3 is the value of the $3\left(\frac{n+1}{4}\right)^{th}$ ordered observation

Example 3.16:Find the value for which 25% of the wages of 7 workers lies below it, second and upper quartiles for the data on the wage distribution of 7 workers.

Wage: 1100, 1150, 1080, 1120, 1200, 1160 and 1400

Arranging the data on ascending order, we have 1080, 1100, 1120, 1150, 1160, 1200, and 1400

Q_1 is the value of $\left(\frac{7+1}{4}\right)^{th}$ *position* item in the ordered observation, $(2)^{nd}$ *position* item =1100. Q_2 is the value of 2* $\left(\frac{7+1}{4}\right)^{th}$ $position\ item = (4)^{th}\ position$ item=1150, Q_3 is the value of 3*$\left(\frac{7+1}{4}\right)^{th}$ *position* item $=(6)^{th}$ *position* item=1200.

b) Quartiles for ungrouped frequency distribution

The i^{th} quartile (Q_i) for the discrete frequency distribution data is the value corresponding to cumulative frequency greater or equal to $i*\frac{n}{4}$ where, $n=\sum_{i=1}^{k} f_i$. Or the value whose cumulative frequency is greater than or equal to $i*\frac{n}{4}$ where, $n=\sum_{i=1}^{k} f_i$

Example 3.17: Consider example 3.12, then compute the first, second and third quartile of the fd.

Solution:we have n=30. The $i*\frac{n}{4}=i*\frac{30}{4}$ $for\ i=1,2,3$, Q_1= the value corresponding to LCF greater or equal to 7.5 which is 15, Q_2= the value corresponding to LCF greater or equal to 15 which is 20, Q_3 = the value corresponding to LCF greater or equal to 22.5 which is 25. i.e. (Q_1=15,Q_1=20, and Q_1=25)

c) Quartiles for grouped frequency distribution

In case of grouped (continuous) frequency distribution quartiles obtained first by obtaining quartile class by the procedure we have applied under ungrouped frequency distribution above,To find i^{th}quartile class compute $\left(\frac{i*n}{4}\right)$ and search for the minimum less than cumulative frequency greater than or equal to this value, then the class corresponding to this cumulative frequency is i^{th} quartile class. Then to determine unique value of i^{th} quartile applies the formula given by:

$$Q_i = L_o + \frac{\left(i*n/4 - cf\right) w}{f}$$

Where, n = the sum of the frequencies of all classes $=\sum f_i$

Lo = the lower class boundary of the i^{th} quartile class.

cf = the cumulative frequencies of class before i^{th} quartile class

f = the frequency of i^{th} quartile class, and w is class width

Example 3.18:Consider example 3.13, find the value for which 75% of the age lies above and/or below it from raw data and grouped dada, and second quartiles.

Solution: To find 1st quartile class, compute $1*\frac{n}{4} = 20/4 = 5$

Then Q_1 lies in the class 27-30, then Lo = 26.5, cf = 3, f = 4 and w = 4. The unique value of Q_1 is computed by $Q_1 = L_o + \frac{\left(n/4 - cf\right) w}{f} = 26.5 + \frac{(20/4 - 3)}{4}4 = 28.5$

The value of Q_1 shows that only 25 % of 20 household headsare with age 28.5 or less.

Similarly, to find 2nd quartile class compute $\left(\frac{2*n}{4}\right) = 2\frac{(20)}{4} = 10$, The class for Q_2 is 31--34, then L_o= 30.5, cf = 7, f = 3 and w=4, the unique value of Q_2 is computed by:

$$Q_2 = L_o + \frac{\left(\frac{2}{4}n - cf\right) w}{f} = 30.5 + \frac{\left(\frac{2}{4}20 - 7\right)}{3}4 = 34.5 = \tilde{X}$$

50 of 20 household heads arewith age 34.5 or less/or more.

ii. Deciles

Deciles are measures which divide a given ordered data in to ten equal parts and each part contains equal no of elements. It has nine points known as 1st, 2nd,…, 9th deciles, denoted by D_1, D_2,… D_9respectively.

a) Deciles for individual data

For series of data set the ith deciles is the value of the item which is at the $i*\left(\frac{n+1}{10}\right)^{th}$ *position* item

$Di = i*\left(\frac{n+1}{10}\right)^{th}$ *position* item where i=1,2,3,3…….,9.

b) Deciles for ungrouped frequency distribution

The ith decile (D_i) for the discrete frequency distribution data is the value corresponding to cumulative frequency greater or equal to $i*\frac{n}{10}$ where, $n = \sum_{i=1}^{k} f_i$. Or the value whose cumulative frequency is greater than or equal to $i*\frac{n}{10}$ where, $n = \sum_{i=1}^{k} f_i$ and $i = 1, 2, 3, ..., 9$

c) Deciles for grouped frequency distribution

For grouped (continuous) data deciles can be obtained first by obtaining the i^{th} deciles classbysearch for the minimum less than cumulative frequency greater than or equal to $i*\frac{n}{10}$, then the class corresponding to this cumulative frequency is i^{th} deciles class. Then, determine the unique decile by the following formula:

$$D_i = L_o + \frac{\left(i\, n/10 - cf\right) w}{f} \text{ , for } \quad i=1, 2, 3\ldots, 9.$$

Where, n = the sum of the frequencies of all classes $= \sum f_i$

L_o = the lower class boundary of the i^{th} deciles class

cf = the cumulative frequencies of class before i^{th} deciles class

f = the frequency of i^{th} deciles class, and w is class width

iii. Percentiles

Percentiles are measures having 99 points which divide a given ordered data in to 100 equal parts and each part consists of equal number of elements. It is denoted by P_1, P_2... P_{99} and known as 1^{st} , 2^{nd} …99^{th} percentiles respectively.

a) Percentile for individual data

For a series of data the i^{th} percentiles is the value of item at the $i*\left(\frac{n+1}{100}\right)^{th}$ *position*

i.e. $P_i = *\left(\frac{n+1}{100}\right)^{th}$ position item where, i=1, 2, 3, 3…, 99.

b) Percentiles for ungrouped frequency distribution

The i^{th}percentile (P_i) for the discrete frequency distribution data is the value corresponding to cumulative frequency greater or equal to $i*\frac{n}{100}$ where, $n = \sum_{i=1}^{k} f_i$. Or the value whose cumulative frequency is greater than or equal to $i*\frac{n}{100}$ where, $n = \sum_{i=1}^{k} f_i$ and $i = 1, 2, 3, ..., 99$

c) Percentile for grouped frequency distribution

First percentile class is obtained; to find i^{th} percentile class compute $\left(\frac{i*n}{100}\right)$ and search less than cumulative frequency less than or equal to this value, then the class corresponding to this cumulative frequency is i^{th} percentile class. Then, for grouped (continuous) frequency distribution data percentile can be obtained by:

$$P_i = L_o + \frac{\left(i\, n/100 - cf\right) w}{f} \quad \text{, for } i=1, 2, 3\ldots, 99.$$

Where, n = the sum of the frequencies of all classes $= \sum f_i$

L_o = the lower class boundary of the i^{th} percentile class

cf = the cumulative frequencies of class before i^{th} percentile class

f = the frequency of i^{th} percentile class, and w is class width

Example 3.19: Consider example 3.11 and compute D_4, D_8, P_{25}, and P_{80}

Example 3.20: Consider example 3.15 and compute D_2, D_9, P_{50}, and P_{95}

Example 3.21: Consider example 3.13 and compute D_1, D_8, P_{60}, and P_{70}

Chapter Four

Measures of Dispersion

4.1 Introduction

In this unit we shall discus the most commonly used measure of dispersion like Range, Quartile Deviation, Mean Deviation, Standard Deviation, coefficient of variation. And measure shapes such as skewness and kurtosis. We have seen that averages are representatives of a frequency distribution. They do not tell anything about the scatterness of observations within the distribution. Suppose that we have the distribution of the yields (kg per plot) of two paddy varieties from 5 plots each.

Variety 1: 45 42 42 41 40

Variety 2: 54 48 42 33 30

The mean yield for both varieties is 42kg. But we cannot say that the yield of the two varieties are the same. The first variety may be preferred since it is more consistent in yield performance. From the above example, it is obvious that a measure of central tendency alone is not sufficient to describe a frequency distribution. In addition to it we should have a measure of scatterness of observations. The scatterness or variation of observations from their average is called dispersion.

Some of the main objectives of dispersion are stated under:

1. ***To determine the reliability of an average***: The measures of dispersion help in determining the reliability of an average. It points out how far an average is representative of a statistical series. If the dispersion or variation is small, the average will closely represent the individual values and it is highly representative. On the other hand, if the dispersion or variation is large, the average will be quite unreliable.
2. ***To compare the variability of two or more series***: The measures of dispersion help in comparing the variability of two or more series. It is also useful to determine the uniformity or consistency of two or more series. A high degree of variation would mean less consistency or less uniformity as compared to the data having less variation.
3. ***For facilitating the use of other statistical measures:*** Measures of dispersion serve the basis of many other statistical measures such as correlation, regression, testing of hypothesis etc.

4. ***Basis of statistical quality control***: The measure of dispersion is the basis of statistical quality control. The extent of the dispersion gives indication to the management as to whether the variation in the quality of the product is due to random factors or there is some defect in the manufacturing process.

Desirable Properties of Measures of Dispersion

1. It should be based on all observations.
2. It should be easy to compute and to understand.
3. It should not be affected much by extreme values.
4. It should not be affected by sampling fluctuation

4.2 Absolute measures of dispersion

Absolute measure is expressed in the same statistical unit in which the original data are given such as kilograms, tones etc. These measures are suitable for comparing the variability in two distributions having variables expressed in the same units and of the same averaging size. These measures are not suitable for comparing the variability in two distributions having variables expressed in different units.

Range: The simplest measure of dispersion is the range. The range is the difference between the two extreme values (highest and lowest value) of data. Range takes only maximum and minimum values into account and not all the values. Hence it is a very unstable or unreliable indicator of the amount of deviation.

-The major area in which range is applied is statistical quality control.

-It is also applicable in the cases where extreme values are important like maximum rainfall, temperature, etc Range =R = X_{max} - Xmin

Example 4.1 Consider the following data on weight of 7 individuals.

24, 25, 30, 15, 47, 43 and 35. R = L-S = 47 - 15 = 32kg.

The relative measures of range, also called *coefficient of range*, is defined as

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

For a grouped data, range is the difference between the upper class boundary of the last class interval and lower class boundary of the first class interval.

Find out range and coefficient of range of the following series

Size	5-10	11-15	16-20	21-25	26-30
Frequency	4	9	15	30	40

Here, $L = Upper\ class\ limit\ of\ the\ largest\ class$

$S = Lower\ class\ limit\ of\ the\ smallest\ class\ \ L = 30\ \ and\ S = 5$

$Range = 30 - 5 = 25$

$$Coefficient\ of\ range = \frac{30-5}{30+5} = 0.7143$$

Properties of range

- It's easy to calculate and to understand
- It can be affected by extreme values
- It can't be computed when the distribution has open ended classes.
- It cannot take the entire data in to account.
- It does not tell any thing about the distribution of values in the series.

Inter Quartile Range: Is the difference between 3rd and 1st quartile and it is a good indicator of the absolute variability than range. IQR = $Q_3 - Q_1$.

Quartile Deviation (semi – inter quartile Range) is a half of inter quartile range

$$QD = \frac{(Q_3 - Q_2) + (Q_2 - Q_1)}{2} = \frac{Q_3 - Q_1}{2}$$

Coefficient of quartile Deviation The relative measure of quartile deviation also called the *coefficient of quartile deviation* is defined as: $Coefficient\ of\ QD = \frac{Q_3 - Q_1}{Q_3 + Q_1}$

Properties of Quartile Deviations

i) The size of quartile deviation gives an indication about the uniformity. If QD is small, it denotes large uniformity. Thus, a coefficient of quartile deviation is used for comparing uniformity or variation in different distribution.

ii) Quartile deviation is not a measure of dispersion in the sense that it doesn't show the scatter around an average but only a distance on scale. As result it is regarded as a measure of partition.

iii) It can be computed when the distribution has an open-ended class. it is quite suitable in the case of open – ended distribution

iv) As compared to range, it is considered a superior measure of dispersion.

v) Since it not influenced by the extreme values in a distribution. It is particularly suitable in highly skewed or irregular distribution.

.

Examples 4.2 For the following frequency distribution find

a) Inter– quartile range.
b) Quartile deviation
c) CQD

Class limit	**Frequency**
21 – 22	10
23 – 24	22
25 – 26	20
27 – 28	14
29 – 30	14
Total	80

$\Longrightarrow$ n/4 = 80/4 = 20, $20^{\text{th ordered}}$ observation

$\Longrightarrow$ The 1^{st} quartile class is 23 -24

$$Q_1 = LCb + \frac{\left(n/4 - cf\right)w}{f} = 22.5 + \frac{(20 - 10)2}{22} = 23.4$$

$$Q_2 = 2\left(\frac{n}{4}\right) = 2\left(\frac{80}{4}\right) = 40, \quad Q_2 \; is \; 40^{th} \; obeservation$$

⟹ The class interval containing Q_2 is 25 – 26.

$$Q_2 = LCb_{Q2} + \frac{\left(2\left(n/4\right) - cf\right)w}{f} = 24.5 + \frac{(40 - 3)\ x2}{20} = \underline{\underline{25.3}}$$

And $Q_3 = 3\left(\frac{n}{4}\right) = 60$, Q_3 is 60th position observation.

⟹ The class limits containing Q_3 is 27 – 28

$$Q_3 = LCb_{Q3} + \frac{\left(3\left(n/4\right) - cf\right)w}{f} = 26.5 + \frac{(60 - 52)}{14} = 27.84$$

a) Inter quartile range $= Q_3 - Q_1 = 27.64 - 23.44 = 4.23$

b) $Q.D = \frac{1}{2}(Q_3 - Q_1) = 4.23/2 = 2.115$

c) CQD =4.23/51.24

The quartile deviation is more stable than the range as it defenses on two intermediate values. This is not affected by extreme values since the extreme values are already removed. However, quartile deviation also fails to take the values of all deviations.

Mean Deviation: Mean deviation is the mean of the deviations of individual values from their average. The average may be either mean or median.

M.D= $\frac{\sum|X - A|}{n}$ for raw data. M.D = $\frac{\sum f|X - A|}{\sum f}$ for grouped data.

Where A is either mean or median.

Example 4.3 Consider the following data and compute mean deviation from mean 53, 56, 57, 59, 63, and 66 $\overline{X} = \sum_{i=1}^{6} Xi \,/\, 6 = 59$

Xi	53	56	57	59	63	66
deviation from mean	6	3	2	0	4	7

Mean deviation = $\frac{\sum \left| X_i - X \right|}{n} = \frac{22}{6} = 3.67$

$\Rightarrow$ The data deviates on average 3.67 from the arithmetic mean

Example 4.4 Calculate the mean deviation for the following data using both mean & median.

X_i :- 14 , 15 , 26 , 20 , 10 , Median 15 , mean = 17

Mean deviation = $\left| 10 - 15 \right| + \left| 14 - 15 \right| + \left| 15 - 15 \right| + \left| 20 - 15 \right| + \left| 26 - 15 \right| = 22/5$

X_i	10	14	15	20	26	Total
$/d_i/ = /x_i - \tilde{X}/$	5	1	0	5	11	22
$/d_i/ = /x_i$ – Mean /	7	3	2	3	9	24

M. D from median = $\frac{\sum \left| x_i - median \right|}{n} = \frac{22}{5} = 4.4$ Mean Deviation from median is 4.4. M. D from mean $= \frac{\sum \left| x_i - mean \right|}{n} = \frac{24}{5} = 4.8$

Example 4.5 Calculate the mean deviation from mean and median

X_i	6	7	8	9	10	11	12
f_i	3	6	9	13	8	5	4
$X_i f_i$	18	42	72	117	80	55	48

Mean= $\frac{\sum f_i x_i}{\sum f_i} = 432/48 = 9$

$$\text{Median}=\frac{\left(\frac{n}{2}\right)^{th}+\left(\frac{n}{2}+1\right)^{th}}{2}=\frac{24^{th}+25^{th}}{2}=\frac{9+9}{2}=9$$

X_i	6	7	8	9	10	11	12	Total
f_i	3	6	9	13	8	5	4	48
$\|di\|$	3	2	1	0	1	2	3	
$fi\|di\|$	9	12	9	0	8	10	12	60

Where di = (Xi – median (or mean)) M. D from median $= \frac{\sum f_i \mid d_i \mid}{\sum f_i} = \frac{60}{48} = 1.25$

Property of Mean Deviation

- The mean deviation takes all values into consideration.
- It is fairly stable compared to range or quartile deviation. But it is not stable as standard deviation. Since, it mathematically ignores signs of deviations.
- It is not possible to use for further statistical investigation.

Variance (S^2 or δ^2)

Variance is the arithmetic mean of square deviation about the mean. When our data constitute a sample , the variance averaging done by dividing the sum of squared deviation from mean by n-1 and it is denoted by $\mathbf{s^2}$ .when our data constitute an entire population variance averaging done by dividing by N and denoted by δ^2 . It is commonly used absolute measure of dispersion

- $\mathbf{s^2} = \sum_{i=1}^{n} (x_i - \bar{x})^2 / n-1$ an unbiased estimator for population variance

- $\delta^2 = \frac{1}{N}\sum_{1}^{N}(X-\mu)^2$; *puplation* var*iance* The computing formula for variance is can be .

simplified as given $S^2 = \frac{\left[\sum_{i=1}^{n} xi^2 - (\sum xi)^2/n\right]}{n-1}$,

Since $\sum(xi - \bar{x})^2 = \sum(x_i - 2x_i\bar{x} + \bar{x}^2) = \sum X_i^2 - 2\bar{x}.\sum x_i + \sum\bar{x}^2$

$= \sum x_{i_2} - 2n\bar{x} + n\bar{x}^2 = \sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum x_i^2 - n\bar{x}^2$

$= \sum x_i^2 - n\left(\frac{\sum x_i}{n}\right)^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$

$\therefore \quad S^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$

∗ Variance for simple frequency distribution

x_i	x_1	x_2	. . .	x_k
f_i	f_1	f_2	. . .	f_k

$S^2 = \frac{\sum_{i-1}^{n} f_i(x_i - \bar{x})^2}{(n-1)}$ *where* $n = \sum f_i$

❖ Determination of variance from grouped frequency distributions.

$s^2 = \sum_{i=1}^{n} f_i\,(m_i - \bar{x})^2 \Big/ (n-1)$ Where m_i is mid value of class

Activity Compute variance for the following frequency distribution

Class interval	1-5	6-10	11-15	16-20
frequency	4	1	2	3

Properties of Variance

1. The variance is always non – negative ($S^2 \geq 0$)
2. If every element in the distributions are multiplied by a constant C the new variance is $S^2_{new} = C^2 S^2_{old}$

Old $x_1, x_2, \ldots, x_n$ $S^2_{old} = \sum (x_i - \bar{x})^2 / n-1$

New $cx_1, cx_2, \ldots, cx_n$ $S^2_{new} = \dfrac{\sum (cx_i - c\bar{x})^2}{n-1}$

$$= \frac{\sum (c\,(x_i - \bar{x}))^2}{n-1} = \frac{\sum c^2 (x_i - \bar{x})^2}{n-1} = \frac{c^2 \sum (x_i - \bar{x})^2}{n-1} = C^2 S^2_{old}$$

3. When a constant c is added to all measurement of the distribution, the variance doesn't change

x_i (old) = x_i, x_2, ... , x_n

x_i (new) = $x_1 + c_1, x_2 + c$, ... , $x_n + c$

$$\bar{X}_{new} = \frac{\sum (x_i + c)}{n} = \frac{\sum x_i + \sum c}{n} = \frac{\sum x_i}{n} + \frac{nc}{n} = \underline{\underline{\bar{X} + c}}$$

$$S^2_{new} = \frac{\sum (x_i + c - (\bar{x} + c))^2}{(n-1)} = \frac{\sum (x_i - \bar{x})}{n-1} = S^2_{old}$$

4. The variance of constant measured n times is zero. $\bar{x} = c$, $S^2 = 0$.

Example 4.6 If the mean & variance of x are 10 & 5 , respectively. Find the mean and variance of y, where **y** = 10x - 5

$$\bar{y} = \frac{\sum_{i=1}^{n} (y_i)}{n} = \frac{\sum (10x - 5)}{n} = \frac{10\sum x - \sum 5}{n} = 10\bar{x} - 5 = 10\,(10) - 5 = 100 - 5 = 95$$

$$\text{var (y)} = \frac{\sum(y_i - \bar{y})^2}{n-1} = \frac{\sum(10x - 5 - 95)^2}{n-1} = \frac{\sum 10^2 (x-10)^2}{n-1} = 10^2 \sum \frac{(x_i - 10)^2}{n-1}$$

$$= 100\,(5) = 500$$

Standard Deviation

The standard deviation is defined as the square root of the mean of the squared deviations of individual values from their mean.

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

- Its advantage over variance is that it is in the same unit as the variable under consideration.
- It is a measure of average variation in the set of data.

Example 4.7 Compute the variance & S.D. for the data given below.

x_i	32	36	40	44	48	Total
frequency	2	5	8	4	1	20

X_i :	32	36	40	44	48	Total
Fi:	2	5	8	4	1	20
$X_i f_{i:}$	64	180	320	176	48	788
$X_i^2 f_{i:}$	2048	6480	12,800	7,744	2,304	31,376

$$S^2 = \frac{\sum f_i x_i^2 - (\sum x_i f_i)^2/n}{\sum f_i - 1} = \frac{31376 - (788)^2/20}{19} = 328.8/19 = 17.31$$

$\Rightarrow S = \sqrt{S^2} = \sqrt{17.31} = \underline{\underline{4.16}}$

- If the s .d of set of data is small then the values are scattered widely about the mean.

$S^2 = 11, \qquad S = \sqrt{S^2} = \sqrt{11} = \underline{\underline{3.316}}$

Example 4.8 Calculate the S.D for the following grouped frequency distribution.

Class intervals	Frequency(f_i)
1 – 3	1
3 – 5	9
5 – 7	25
7 – 9	35
9 – 11	17
11 – 13	10
13 – 15	3
Totale	100

$$S^2 = \frac{\sum f_i m_i^{\,2} - \left(\sum f_i m_i\right)^2 / n}{\sum f_i - 1} = \frac{7016 - (800)^2/100}{99} = 6.2 \quad S = \sqrt{S^2} = \sqrt{6.22} = \underline{\underline{2.49}}$$

4. 3 Relative Measure of Dispersion

Suppose that the two distributions to be compared are expressed in the same units and their means are equal or nearly equal. Then their variability can be compared directly by using their standard deviations. However, if their means are widely different or if they expressed indifferent units of measurement, we can not use the standard deviation as such for comparing their variability. We have to use the relative measures of dispersion in such situation.

Coefficient of variation (CV): The CV is a unit free measure. It is always expressed as percentage.

$$\textbf{CV} = \frac{SD}{Mean}\,\textbf{100\%}$$

The CV will be small if the variation is small. Of the two groups, the one with less CV is said to be more consistent. The coefficient of variation is unreliable if the mean is near zero. Also it is unstable if the

measurement scale used is not ratio scale. The CV is informative if it is given along with the mean and standard deviation. Otherwise, it may be misleading.

Example 4.9 Consider the distribution of the yields (per plot) of two paddy varieties. For the first variety, the mean and standard deviation are 60kg & 10kg, respectively. For the second variety, the mean and standard deviation are 50kg & 9kg, respectively. Then we have,

CV = (10/60)100%=16.7%, for first variety.

CV = (9/50)100%=18.0%, for second variety.

It is apparent that the variability in first variety is less as compared to that in the second variety.

Coefficient of Mean Deviation*:* The coefficient of mean deviation is founded by dividing the mean deviation by the measure of central tendency about which the deviation is computed.

- It is a relative measure of dispersion, coefficient of mean deviation can be computed as

$$\text{CMD} = \frac{Meandeviation}{Mean} \qquad or\ CMD = \frac{Mean\ deviation}{Median}$$

Example 4.10 Coefficient of the mean deviation from mean & median for above example 4.4 is

Mean Deviation from median is 4.4, M. D from mean $= \frac{24}{5} = 4.8$

C. M. D from median $= \frac{mean\ deviatin\ \ from\ \ median}{median} = \frac{4.4}{15} = 0.293$

C. M. D from mean $= \frac{mean\ \ deviation\ \ \ from\ \ mean}{mean} = \frac{4.8}{17} = 0.283$

The standard Score: The standard score is denoted by Z and defined as $Z = \frac{(x_i - \bar{x})}{S}$ Where S – Standard deviation of the distribution, Xi each observation value. This measures the deviation of individual observation from the mean of the total observation in the unit of standard deviation and termed as Z – Score. The Z – scores of individuals in different groups are then added to give a true Measure of relative performance.

Example 4.11 Compare the performance of the following two students.

Candidate	Marks in economics	Marks in Acct.	Total
A	84	75	159
B	74	85	159

Average mark for Accounting is 50 with standard deviation of 11 and average marks for Economics is 60 with standard deviation 13.Whose performance is better A or B?

Z score for A

$$\begin{cases} Economics & \dfrac{84 - 60}{13} = 1.846 \\ Accounting & \dfrac{75 - 50}{11} = 2.273 \end{cases}$$

Total Z score for A = 1.846 + 2.27 = 4.119

Z score for B

$$\begin{cases} Economics & \dfrac{74 - 60}{13} = 1.077 \\ Accouniting & \dfrac{85 - 50}{11} = 3.182 \end{cases}$$

Total Z – Score for B = 1.077 + 3.182 = 4.25

Since B's Z – score is higher; student B had good performance than student A.

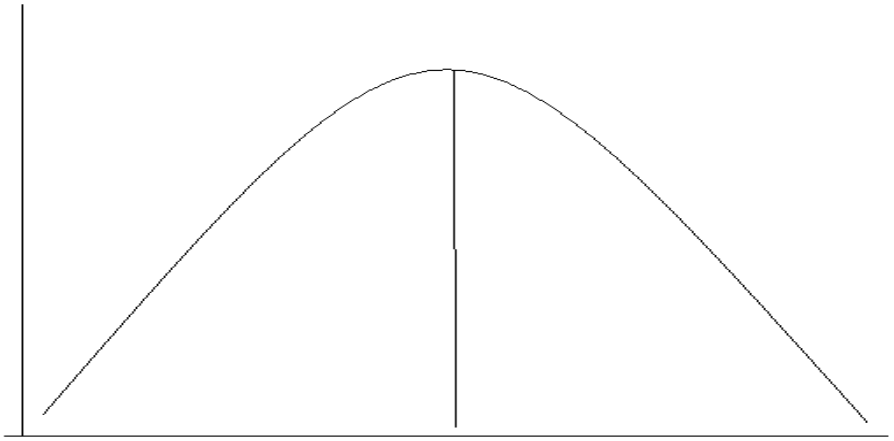
4.4 Measure of shapes

We have seen that averages and measure of dispersion can help in describing the frequency distribution. However, they are not sufficient to describe the nature of the distribution. For this purpose, we use the other concepts known as Skewness and Kurtosis.

Skewness: Skewness means lack of symmetry. When the values are uniformly distributed around the mean a distribution is said to be symmetrical. For example, the following distribution is symmetrical about its mean 3.

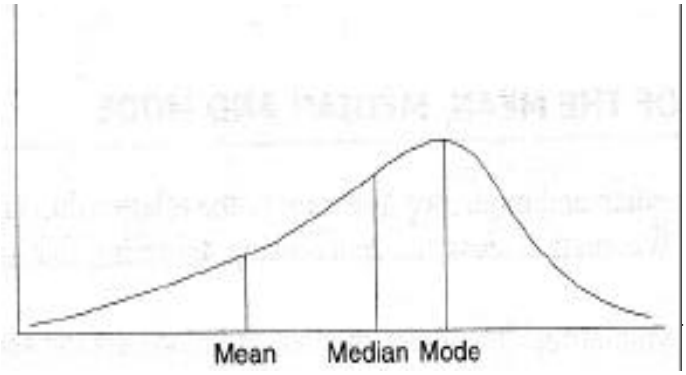
X_i :	1	2	3	4	5
f_i :	5	9	12	9	5

In a symmetrical distribution the mean, median and mode coincide, that is, $\overline{X} = \tilde{X} = \hat{X}$.

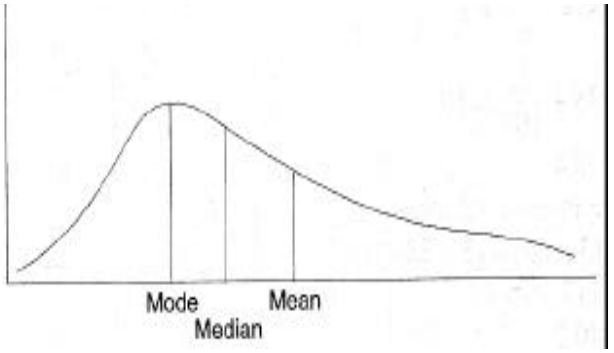


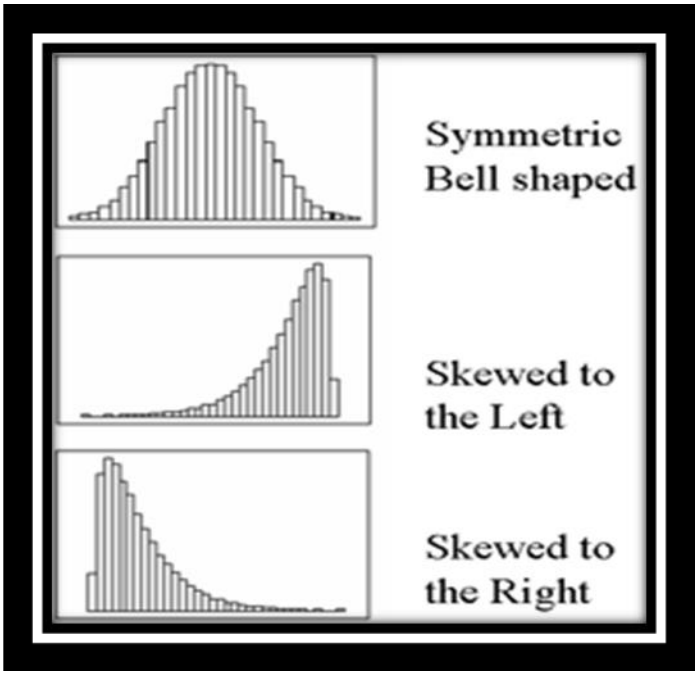
$\overline{X} = \tilde{X} = \hat{X}$ *Symmetrical distribution*

When a distribution is skewed to the right; mean > median > mode. If we take income distribution for different number of families; Income distribution is skewed to the right mean that a large number of families have relatively have low income and a small number of families have extremely high income. In such a case, the mean is pulled up by the extreme high incomes and the relation among these three measures is as shown in figure. Here, we find that mean > median > mode. When a distribution is skewed to the left, then mode > median > mean. This is because here mean is pulled down below the median by extremely low values.



Right(+) skewed Distribution





Left(-)skewedDistribution

Karl person's Measure of skewness: In case the distribution is symmetric we will have

Arithmetic mean. = Median = Mode; unless they will not be equal if the distribution is skewed.

Therefore the distance between the A.M. and the Mode (A.M – Mode) can also be used as a measure of skewness. However since the measure of skewness should be a pure number we define as $Sk = \frac{A.M - Mode}{\delta}$, Where δ is the standard deviation of the distribution.

For distribution which are bell shaped and are moderately skewed, we have an approximate relationship between the A.M, Median and mode.

A. M – Mode = 3 (A. M – Median)

Accordingly we may define skewness as follows $Sk = \frac{3(A.M - Median)}{\delta}$

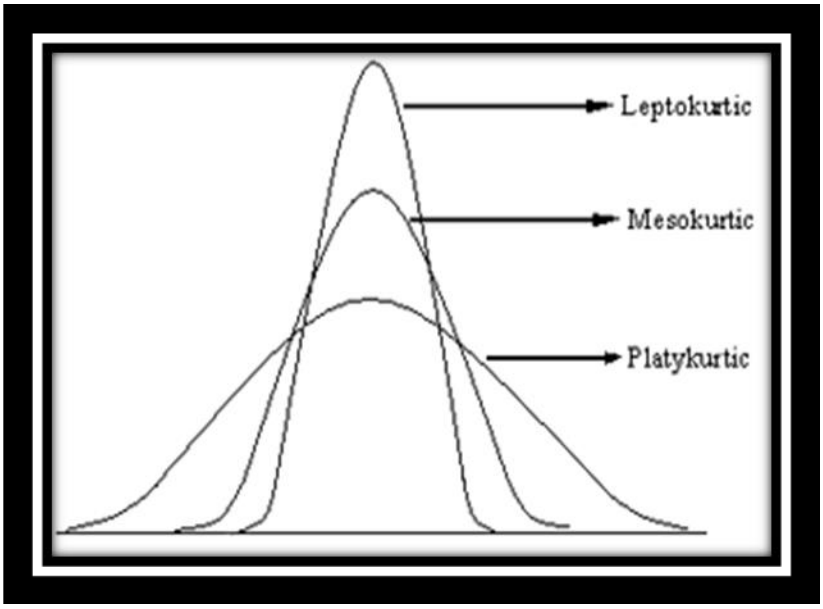
For a symmetrical distribution Sk= 0. If the distribution negatively skewed, then the value of S_k is negative, and if it is positively skewed then S_k is positive. The range for values of S_k is from -3 to 3. The other measure uses the β (beta) coefficient which is given by, $\beta_1 = \mu_3^2/\mu_2^3$. Where μ_2 & μ_3 are the second and the third central moments.

The second central moment is nothing but the variance. The sample estimate of this coefficient is $b_1 = m_3^2/m_2^3$ where m_2& m_3 are sample central moments given by,

$$m_2 = \frac{\sum(X-\bar{X})^2}{n-1} \text{ or } \frac{\sum f(X-\bar{X})^2}{n-1}, \quad m3 = \frac{\sum(X-\bar{X})^3}{n-1} \text{ or } \frac{\sum f(X-\bar{X})^3}{n-1}$$

For a symmetrical distribution b_1 is zero. And also Skewness is positive or negative depending upon whether m_3 is positive or negative.

Example 4.12 The first four moments about mean of the distribution are 0, 2.5, 0.7, and 18.75. Test the Skewness of distribution



Kurtosis: A measure of the peakedness or convexity of a curve is known as Kurtosis.

All the three curves are symmetrical about the mean. Still they are not of the same type. One has different peak as compared to that of others. Curve (1) is known as meso-kurtic (normal curve); curve (2) is known as leptokurtic (leaping curve) and curve (3) is known as platy-kurtic (flat curve). Kurtosis is measured by Pearson's coefficient, β_2. It is given by $\beta_2 = \mu_4/\mu_2^2 = \frac{\mu_4}{\delta_4}$

The sample estimate of this coefficient is $b_2 = m_4/m_2^2$, where m_4 is the 4^{th} central moment given by $m_4 = \frac{\sum(X-\bar{X})^4}{n-1}$. The distribution is called meso-kurtic if the value of $b_2 = 3$. When b_2 is more than 3 the distribution is said to be leptokurtic. And also, if b_2 is less than 3 the distribution is said to be platykurtic.

Example 4.13 The measure of skewness and kurtosis are given below for data in table.

Value(xi)	3	4	5	6	7	8	9	10
Frequency(f)	4	6	10	26	24	15	10	5

Value(xi)	Frequency(f)	d=X- $\overline{X}$	f*d^2	f*d^3
3	4	-3.7	54.76	-202.612
4	6	-2.7	43.74	-118.098
5	10	-1.7	28.90	-49.130
6	26	-0.7	12.74	-8.918
7	24	0.3	2.16	0.648
8	15	1.3	25.35	32.955
9	10	2.3	52.90	121.670
10	5	3.3	54.45	179.685

$m_2=s^2=\frac{\sum f_i\left(X-\bar{X}\right)^2}{n-1}=\frac{275}{99}=2.7777$

$m_4=\frac{\sum f_i\left(X-\bar{X}\right)^4}{n-1}=\frac{2074.13}{99}=20.9508$, $m3=\frac{\sum f_i\left(X-\bar{X}\right)^3}{n-1}=\frac{-43.8}{99}=-0.4424$

$b_1=\frac{m_3^{\,2}}{m_2^{\,3}}=\frac{(-0.4424)^2}{(2.7777)^3}=0.0091$, $b_2=\frac{m_4}{m_2^{\,2}}=\frac{20.9508}{(2.7777)^2}=2.7153$

It is negatively skewed since m_3 is negative. The value of b_2 is 2.7153 which is less than 3. Hence the distribution is platykurtic.

CHAPTER FIVE

PROBABILITY AND PROBABILITY DISTRIBUTIONS

5.1. Basic Concepts of Elementary Probability

Introduction

An experiment is any activity that generates *outcome(s).*If an experiment is conducted under homogeneous and similar conditions, we come across two types of situations or phenomena. The results or the outcome is unique or certain which are called as deterministic phenomena by which the result can be predicted certainly from an experiment performed. For instance, in case of ohm's law voltage, V = I*R, Force, F = m *a , the distance (S) covered by a particle after a time t is given by S=$Vt + \frac{1}{2}at^2$,…, etc are deterministic phenomena because if we know the right hand side, we certainly determine the left hand side.

The results or the outcomes of an experiment are not unique, but can be one of the several possible outcomes; this type of model is called probabilistic or non-deterministic (random) phenomena.

Example -In random tossing of a coin, we are not sure of getting head or tail.

- A sex of new born baby may be male or female
- A manufacturer cannot be ascertained (sure) of the future demand of his product with certainty.

As everybody knows our world is a full of uncertainty ,even ,no one knows exactly what comes after a minute, an hour…etc. but we can guess the chance that something will happen. The word probability or chance is very commonly used in day-to-day conversation, and generally, people have some idea what it means. Terms like possible, probable, or likely and so on, have all similar meanings.

***Probability*:** can be defined as a measure of the likelihood that a particular event will occur or it is a science of decision making with calculated risk in face of uncertainty. It is a numerical measure with a value between 0 and 1 of such likelihood. Where the probability of zero indicates that the given event cannot occur and the Probability of one assures certainty of such an occurrence.

Definitions of some Probability Terms

1. ***Random Experiment***: is statistical process of measurement or observation that can be repeated in any single trial of which the outcome is unpredictable. For example, tossing of a fair coin, rolling a die etc is considered as a statistical experiment.
2. ***Outcome*:** is the result of a random experiment.

Example

Experiment	Outcomes
Tossing of a fair coin	Head, tail
Rolling a die	1, 2, 3, 4, 5, 6
Selecting an item from a production lot	defective (faulty), non-defective (good)
Introducing a new product	Success, failure

3. ***Sample space***: A sample space is the collection of all possible outcomes of an experiment. For example, there are two possible outcomes of a toss of a fair coin, which are a head and a tail. Then the sample space, for this experiment denoted by S would be: S = {H, T}. Each possible outcome in the sample space is called *sample point*.
4. ***Event***is a subset of the sample space or it is asset containing sample points of a certain sample space under consideration. For examples, getting two heads in the trial of tossing three fair coins simultaneously would be an event.
5. ***Elementary event***(simple event) is a single possible outcome of an experiment.
6. ***Composite***(compound) **event** is an event having two or more elementary events in it. For example, rolling a die sample space = {1,2,3,4,5,6) an event having { 5} is simple event where as having even number={2,4,6} is compound (composite)event.
7. ***Mutually exclusive events***: Two events are said to be mutually exclusive, if both events cannot occur at the same time as outcome of a single experiment. In other word two events E_1 and E_2 said to be mutually exclusive events if there is no sample point in common to both events E_1 and E_2. For example, if we roll a fair dice, then the experiment is rolling the die and sample space (S) is $S = \{1,2,3,4,5,6\}$ If we are interested the outcome of event E_1 getting even numbers and E_2 odd numbersE_1= {2, 4, 6} E_2 = {1, 3, 5}. Clearly $E_1 \bigcap E_2 = \emptyset$. Thus E_1 and E_2 are mutually exclusive events.

8. ***Independent Events*:** Two events A and B are said to be independent events if the occurrence of event A has no influence (bearing) on the occurrence of event B. For example, if two fair coins are tossed, then the result of one toss is totally independent of the result of the other toss. The probability that a head will be the outcome of any one toss will always be ½, irrespective of whatever the outcome is of the other toss. Hence, these two events are independent. On the other hand, consider drawing two cards from a pack of 52 playing cards. The probability that the second card will be an ace would depend up on whether the first card was an ace or not. Hence these two events are not independent events.

Another example a bag contains balls of two different colours say yellow and white. Two balls are drawn successively .First ball is drawn from a bag and replaced after notes its colour. Let us assume that it is yellow and denote this event by A. Another ball is drawn from the same bag and its colour is noted let this event denoted by B. Clearly, the result of first draw has no effect on the result of the second draw. Hence, the events A and B are independent events.

9. ***Equally likely outcomes*:** In a certain experiment, if each outcome in the sample space has the same chance to occur, then we say that the outcomes are equally likely outcomes.

Fundamental Principles of Counting Techniques

If the number of possible outcomes in an experiment is small, it is relatively easy to list and count all possible events. When there are large numbers of possible outcomes an enumeration of cases is often difficult, tedious, or both. Therefore, to overcome such problems one can use various counting techniques or rules.

i. ***Addition rule:*** Suppose that a procedure designated by 1, can be performed in n_1 ways. Assume that second procedure designated by 2 can be performed in n_2 ways. Suppose furthermore that it is not possible both procedures 1 and 2 are performed together. The number of ways in which we can perform 1 or 2 procedures is $n_1 + n_2$ ways. This can be generalized as follows if there are k procedures and i^{th} procedure may be performed in n_i ways, i=1, 2, …, k , then the number of ways in which we perform procedure 1 or 2 or … or k is given by $n_1 + n_2 + \ldots + n_k = \sum_{i=1}^{k} n_i$, assuming that no two procedures performed together.

Example 5.1: Suppose that we are planning a trip and deciding between bus and train transportation. If there are 3 bus routes and 2 train routes to go from A to B, find the available routes for the trip.There are 3+2 = 5 possible routes for someone to go from A to B.

ii. ***Multiplication Rule:***Suppose that procedure 1 can be performed in n_1 ways. Let us assume procedure 2 can be performed in n_2 ways. Suppose also that each way of doing procedure 2 may be followed by any way of doing procedure 1, then the procedure consisting of n_1 followed by n_2 may be performed by $n_1 * n_2$ ways. We also generalize to k procedures.

Example 5.2: An airline has 6 flights from A to B, and 7 flights from B to C per day. If the flights are to be made on separate days, in how many different ways can the airline offer from A to C?In operation 1 there are 6 flights from A to B, 7 flights are available to make flight from B to C. Altogether there are 6*7 = 42 possible flights from A to C.

Example5.3: Suppose that in a medical study patients are classified according to their blood type as A, B , AB, and O; according to their RH factors as (+) or (-) and according to their blood pressure as high, normal or low, then in how many different ways can a patient be classified ?

The1st classification has done in 4 ways; the 2nd in 2 ways, and the 3rd in 3 ways. Thus patient can be classified in 4*2*3 = 24 different ways.

iii. ***Permutations***: Permutation is an arrangement of all or parts of a set of objects with regard to order.

Rule 1: The number of permutations of n distinct objects taken all together is n! **Or**

In particular, the number of permutations of n objects taken n at a time is

$_nP_n = \frac{n!}{(n-n)!} = \frac{n!}{0!}$ = n!. In definition 0! = 1! = 1.

Example 5.4: In how many ways 4 people are lined up to get on a bus (or to sit for photo graph)?

In 4! = 4*3*2*1 = 24 Ways.

Rule-2: A permutation of n different objects taken r at a time is an arrangement of r out of the n objects, with attention given to the order of arrangement. The number of permutations of n objects taken r at a time is denoted by ${}_nP_r$, or P (n,r) and is given by

$${}_nP_r = n(n-1)(n-2)\ldots(n-r+1) = \frac{n!}{(n-r)!}$$

Example 5.5: The number of permutations of letters a, b & c taken two at a time is

${}^3P_2 = 6$. These are ab, ba, ac, ca, bc & cb.

Rule-3: The number of permutation of n objects taken all at a time, when n_1 objects are alike of one kind, n_2 objects are alike of second kind, …, n_k objects are alike of k^{th} kind is given by:

$$\frac{n!}{n_1!n_2!n_3!\ldots n_k!} = \frac{\left(\sum_{i=1}^{k} n_i\right)!}{\prod_{i=1}^{k}(n_i!)}$$

Example 5.6 The total number of arrangement of the letters of the word *STATISTICS* taken all at a time is given by $\frac{10!}{3!3!1!2!1!} = 50{,}400$ since there are 3s's, 3t's, 1a, 2i's and 1c.

Note:-The number of arrangements of n distinct objects around circular object (table) is (n-1)! and when the method of selection or arrangement of r objects from n objects with repetition the possible numbers of arrangements are n^r.

Example5.7: Arba Minch University Registrar Office wants to give identity number for students by using 4 digits. The number should be considered by the following numbers only: {0, 1, 2, 3, 4, 5, and 6}. Hence, how many different ID Numbers could be preferred by the Registrar

a. Without repeating the number?
b. With repetition of numbers?

We have 7 possible numbers for 4 digits. But the required number of digits for ID number is 4.Hence n = 7 & r = 4.The possible number of ID numbers given for students without repeating the digit is ${}_nP_{r,} = \frac{n!}{(n-r)!} = \frac{7!}{(7-4)!} = 7*6*5*4 = 840.$

☞ The possible number of ID numbers given for students with repeating the digit is

$n^r = 7^4 = 7*7*7*7 = 2401$

iv. ***Combinations:*** Combination is the selection of objects without regarding order of arrangement. A combination of n different objects taken r at a time is a selection of r out of n objects, with no attention given to the order of arrangement. The number of combinations of n objects taken r at a time is denoted by the symbol $\binom{n}{r}$ or ${}_nC_r$ is given by

$$\binom{n}{r} = \frac{n(n-1)(n-2)...(n-r+1)}{r!} = \frac{n!}{r!(n-r)!} = \frac{{}_nP_r}{r!}$$

Example 5.8: The number of combinations of letters a, b& c taken two at a time is 3C2= $\frac{3!}{2!1!}$ = 3.

These are ab, ac and bc. Note that *ab* is the same combination as *ba*, but not the same in permutation.

Example 5.9: Suppose in the box 3 red, 3 white and 5 black equal sized balls are there. We want to draw 3 balls at a time. How many ways do we have from each type?

→ Solution $\binom{3}{1}\binom{3}{1}\binom{5}{1} = 3(3)5 = 45$ ways.

Different Approaches to Probability

There are four different conceptual approaches to study of probability theory

i. Classical or Mathematical Approach

If a random experiment results in N exhaustive, mutually exclusive and equally likely outcomes; out of which n are favorable to the happening of an event A, then the probability of occurrence of A, usually denoted by P (A) is given by: $P(A) = \frac{favorable\,cases\,to\,A}{exhaustive No.of\ cases} = n/N$

Example 5.10: In a given basket there is 3 yellow, 4 black and 3 white balls. What is the probability of selection of one black ball?

Let event A drawing of black ball**,** P (A) = $\frac{favorable\,cases\,to\,A}{exhaustive No.of\ cases} = \frac{4}{10} = 0.4$

ii. Empirical or Frequency Approach

The classic definition of probability has a disadvantage in that of the words equally likely is vague. In fact, since these words seem to be synonymous with "equally probable", the definition is circular because we are essentially defining probability in terms of itself. For this reason, a statistical definition of probability has been advocated by some people. According to this the estimated probability, or empirical probability, of an event is taken to be the relative frequency of occurrence of the event when the number of observations is very large. The probability itself is the limit of the relative frequency as the number of observations increases indefinitely.

Example 5.11: If 1000 tosses of a coin result in 529 heads, the relative frequency of heads is 529/1000 = 0.529. If another 1000 tosses results in 493 heads, the relative frequency in the total of 2000 tosses is $\frac{529+493}{2000}$=0.511.

According to the statistical definition, by counting in this manner we should ultimately get closer and closer to a number that represents the probability of a head in a single toss of the coin. From the results so far presented, this should be 0.5 to one significant figure.

iii. Axiomatic Approach

Both the classical and frequency approaches have serious drawbacks, the first because the words *equally likely* are vague and the second because the *large number* involved is vague. Because of these difficulties, statisticians have been led to an *axiomatic approach* of probability.

Given a sample space of a random experiment S, the probability of the occurrence of any event A is defined as a set function P (A) satisfying the following axioms:

1: P (A) is defined, is real and non-negative i.e. $0 \leq$ P (A)

2: P(S) = 1 where S is the sample space (sure event)

3: If A_1, A2 … A_n is any finite or infinite sequence of disjoint events of S, then

$$P\left\{\bigcup_{i=1}^{k} A_i\right\} = \sum_{i=1}^{k} P(A_i)$$

Results of Axioms

1. $P(\phi) = 0$, for impossible event (null set)
2. $P(A^c) = 1-P(A)$
3. If A and B are two independent events then the chance of occurrence of both two events is

$$P(A \cap B) = P(A)\,P(B)$$

iv. Subjective Approach

A probability derived from an individual's personal judgment about whether a specific outcome is likely to occur. Subjective probabilities contain no formal calculations and only reflect the subject's opinions and past experience. Subjective probabilities differ from person to person. Because the probability is subjective, it contains a high degree of personal bias. An example of subjective probability could be asking Arsenal fan, before the football season starts, the chances of Arsenal winning the world champions. While there is no absolute mathematical proof behind the answer to the example, fans might still reply in actual percentage terms, such as the Arsenal having the 95% chance of winning the world champions.

Events as a Set

If A and B are two events then

- ✓ $A \cup B$: the happening of at least event A or B.
- ✓ $A \cap B$: the simultaneously happening of both events A and B.
- ✓ A^c : A does not happen (complement of event A).
- ✓ $A^c \cap B^c$: neither A nor B happens
- ✓ $A^c \cap B$: B occurs alone or exactly B occurs or only B occurs.
- ✓ $(A \cap B^c) \cup (A^c \cap B)$: exactly one of the two events A and B happens

The probability of occurrence of at least one of the two events A and B is given by:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

If A and B are mutually exclusive events, then

$$P(A \cup B) = P(A) + P(B)$$

Conditional Probability and Independence

Let there be two events A and B. Then the probability of event A given that the outcome of event B is given by: $P[A|B] = \frac{P[AnB]}{P[B]}$ Where: P [A |B] is interpreted as the probability of event A on the condition that event B has occurred. In this case P [A n B] is the joint probability of event A and B, and P [B] is not equal to zero.

And $P(B/A) = \frac{P(B \cap A)}{P(A)}$, where $P(A) \neq 0$

Example 5.12:120 employees of a certain factory are given a performance test and are divided in to two groups as those with good performance(G) and those with poor performance (P) the result is given below

	Good performance (G)	Poor performance(P)	Total
Male (M)	60	25	80
Female (F)	25	15	40
Total	85	35	120

The probability of a person to be male given that it has a good performance is

$$P(M|G) = \frac{P(M \cap G)}{P(G)} = \frac{60/120}{85/120} = \frac{12}{17}$$

The probability of a person to be female given that it has a poor performance is

$$P(F|P) = \frac{P(M \cap P)}{P(P)} = \frac{15/120}{35/120} = \frac{3}{7}$$

Example 5.13: A jar contains black and white marbles. Two marbles are chosen without replacement. The probability of selecting a black marble and a white marble is 0.34, and the probability of selecting a black marble on the first draw is 0.47. What is the probability of selecting white marble on the second draw, given that the first marble drawn is black?

$$P(White \mid Black) = \frac{P(Black \ and \ White)}{P(Black)} = \frac{0.34}{0.47} = 0.72$$

Example 5.14: The probability that it is Friday and that a student is absent is 0.03. Since there are 5 schooldays in a week, the probability that it is Friday is 0.2. What is the probability that a student is absent given that today is Friday?

$$P(Absent \mid Friday) = \frac{P(Friday \ and \ Absent)}{P(Firday)} = \frac{0.03}{0.2} = 0.15$$

It often happens that the knowledge that a certain event A has occurred has no effect on the probability that some other event B has occurred, that is, $P(A \mid B) = P(A)$. One would expect that in this case, the equation $P(B \mid A) = P(B)$ would also be true. If these equations are true, we may say B is *independen*t of A. In another words two events A and B are said to be independent if $P(A \mid B) = P(A)$ $P(B \mid A) = P(B)$ and P (A $\cap$ B) = P (A) P (B)

5.2. Random Variable and Probability Distributions

Random Variables

A variable whose value is determined by random experiment is called random variable. Mathematically a random variable is defined as a function from the sample space S to the line of real numbers R. In other words random variable is a function that assigns a real number to each outcome in the sample space of a random experiment. i.e. X: S$\rightarrow$R

A random variable takes a possible outcome and assigns a number to it.

Example: Flip a coin three times, let X be the number of heads in three tosses.

S= {(HHH),(HHT), (HTH), (HTT), (THH), (THT), (TTH), (TTT)}

X (HHH) =3,

X (HHT) =X (HTH) =X (THH) =2

X (HTT) =X (THT) =X (TTH) =1, X (TTT) =0

X= {0, 1, 2, 3}. X assumes a specific number of values with some probabilities.

Random variables are of two types:

1. ***Discrete Random Variables*:** are variables which can assume only a specific number of values. They have values that can be counted. Examples:
    - ✓ Tossinga coin n times and count the number of heads.
    - ✓ Number of children in a family.
    - ✓ Number of car accidents per week.
    - ✓ Number of defective items in a given company.
    - ✓ Number of bacteria per two cubic centimeter of water.
2. ***Continuous Random Variable:*** are variables that can assume all values between any two given values. Examples:
    - ✓ Height of students at certain college.
    - ✓ Mark of a student.
    - ✓ Life time of light bulbs.
    - ✓ Length of time required to complete a given training.

Probability Distributions

Probability distribution of a random variable is the formula, table or graph that gives the possible value of a random variable X with their associated probability P(X). A probability distribution consists of possible values of the random and the corresponding probabilities of the values. It can be classified as discrete or continuous depending on the nature of the random variable.Discrete probability distributions are commonly called probability mass function (pmf) and denoted by P(X) = P(X=x). Similarly, continuous probability distributions are known as probability density function (pdf) denoted by

$$f(X) = \int_{-\infty}^{\infty} f(x)dx$$

Example: Consider the experiment of tossing a coin three times. Let X be the number of heads. Construct the probability distribution of X.

Solution:First identify the possible value that X can assume then calculate the probability of each possible distinct value of X and express X in the form of frequency distribution.

X=x	0	1	2	3
P(X=x)	1/8	3/8	3/8	1/8

Properties of Probability Distribution

The probability distribution of a random variable should satisfy the following conditions.

1. $P(x) \geq 0, \quad$ *if X is discrete*
   $f(x) \geq 0, \quad$ *if X is continuous*
2. $\sum_x P(X = x) = 1, \quad$ *if X is discrete*
   $\int f(x)dx = 1, \quad$ *if X is continuous.*

Note:

1. If X is a continuous random variable then
   $$P(a < X < b) = \int_a^b f(x)dx$$
2. Probability of a fixed value of a continuous random variable is zero.

$\Rightarrow P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b)$

3. Probability means area for continuous random variable.

The probability distribution can be represented graphically by plotting P(X) against X, just as relative frequency distributions. By cumulating probabilities, we obtain cumulative probability distributions, which are analogous to cumulative relative frequency distributions.The cumulative distribution

is: $$F(X) = P(X \leq x) = \begin{cases} \sum_{X \leq x} P(X = x) & \text{if X is discrete} \\ \int_{-\infty}^{x} f(t)dt & \text{if X is continous} \end{cases}$$

Expectation and Variance of a Random Variable

Expected Value

1. Let a discrete random variable X assume the values $X_1, X_2, \ldots, X_n$ with the probabilities $P(X_1)$, $P(X_2), \ldots, P(X_n)$ respectively. Then the expected value of X , denoted as E(X) is defined as:

$E(X) = X_1P(X_1) + X_2P(X_2) + \ldots + X_nP(X_n) = \sum_{i=1}^{n} X_iP(X = X_i)$

2. Let X be a continuous random variable assuming the values in the interval (a, b) such that $\int_a^b f(x)dx = 1$, *then*
   $$E(X) = \int_a^b x\, f(x)dx$$

Variance of a Random Variable

Let X be given random variable with expected value (mean) of E(x), then the variance of X is given by:

$$Var(X) = E(x - E(X)) = \begin{cases} \sum_{\forall x} (x - E(X))^2 P(X = x) & \text{if X is discrete} \\ \int_{-\infty}^{\infty} (x - E(X)^2) f(x) dx & \text{if X is continous} \end{cases}$$

In the form of expectation variance of X is $Var(X) = E(X^2) - [E(X)]^2$ or
$Var(X) = E(X(X-1)) + E(X) - (E(X))^2$
Where:

$$E(X^2) = \sum_{i=1}^{n} x^2 P(X = x), \qquad \text{if } X \text{ is discrete}$$

$$= \int_a^b x^2 f(X = x), \qquad \text{if } X \text{ is continuous}$$

Example 1: Compute the mean and variance of the random variable X, which denotes the number showing up when a single die is rolled.

Solution: First we have to find the frequency distribution,

$X=x_i$	1	2	3	4	5	6
$P(X=x_i)$	1/6	1/6	1/6	1/6	1/6	1/6

$$E(X) = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = 3.5$$

$$V(X) = (1 - 3.5)^2 \frac{1}{6} + (2 - 3.5)^2 \frac{1}{6} + (3 - 3.5)^2 \frac{1}{6} + (4 - 3.5)^2 \frac{1}{6} + (5 - 3.5)^2 \frac{1}{6} + (6 - 3.5)^2 \frac{1}{6}$$

$= \underline{2.9167}$

Example 2: Compute the mean and variance of the following probability distribution.

$$f(x) = \begin{cases} \frac{1}{4} & for\ x \in [0,4], \\ 0 & otherwise. \end{cases}$$

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx = \int_0^4 x \frac{1}{4} dx = \frac{1}{4}\left[\frac{1}{2}x^2\right]_0^4 = \frac{1}{4}\left(\frac{1}{2} \cdot 4^2 - \frac{1}{2} \cdot 0^2\right) = 2$$

$$V(X)=\int_{-\infty}^{\infty}(x-\mu)^2 f(x)\,dx=\int_0^4(x-2)^2\frac{1}{4}dx=\frac{1}{4}\int_0^4(x^2-4x+4)\,dx=\frac{1}{4}\left[\frac{1}{3}x^3-\frac{4}{2}x^2+4x\right]_0^4$$
$$=\frac{1}{4}\left(\frac{1}{3}4^3-\frac{4}{2}4^2+4\cdot 4-0\right)=\frac{4}{3}$$

5.3. Common Probability Distribution

Simple probabilities can be computed from elementary consideration. However, in dealing with probabilities of whole classes of events, we have to consider more efficient ways of analysis of probability. For this purpose we should know the concept of a probability distribution. In this section we shall examine in detail some probability distribution s that have proved themselves empirically to be useful probability models for any type of practical problems. The probability functions are derived mathematically based on certain assumptions that are assumed to be hold for random phenomenon. We have many probability distributions; discrete probability distributions are Bernoulli distribution, binomial distribution, Poisson distribution, multinomial distribution, geometric distribution, hyper geometric distribution, etc. Continuous probability distributions are uniform distribution, normal distribution, expontial distribution,t-distribution, chi-square distribution, F-distribution, gamma distribution, Weibul distribution, etc.

In practice we have use some of both discrete and continuous probability models.

Common Discrete Probability Distributions

i. Binomial Distribution

It is used to represent the probability distribution of discrete random variables. Binomial means two categories. The successive repetition of an observation (trial) may result in an outcome which possesses or which does not possess a specified character. Our primary interest will be either of these possibilities. Conventionally, the outcome of primary interest is termed as *success.* The alternative outcome is termed as *failure.* These terminologies are used irrespective of the nature of the outcome. For example, non-germination of a seed may be termed as success.

In binomial distribution the experiment consisting the following criteria/assumptions

- The experiment consists of n identical trials
- There is only two outcomes in Bernoulli trials (success or failure)

- The number of trials n is fixed andindependent
- At each trial the probability of success (p) remains the same

The variable X which represents the count of the number of successes in Bernoulli trials will be a discrete random variable. The probability distribution of such discrete random variable X is called the binomial distribution. The binomial distribution is given by the probability mass function (pmf): P(X=x) = $\binom{n}{x} p^x q^{n-x}$ for all possible values of X.

where, n= number of trials, x= number of successes in a trial, n-x = number of failures in a trial, p = probability of success (= x/n), q = 1 - p = probability of failure, and $\binom{n}{x}$ = the possible number of ways in which x successes can occur.

The binomial distribution is determined by two parameters n and p.

The expected value of the binomial distribution is np and the variance of it is npq.

Example 1:A given mid-exam contains 10 multiple choice questions, and each question has four alternatives with one exact answer. Find the probability that the student exactly answered

i. 3 questions
ii. 8 questions
iii. At least 3 questions

Using binomial distribution we can get the probability value easily. That is n = 10,

p = ¼ (the chance of getting answer from 4 alternatives) q = 1- p = 1- ¼ = ¾

The possible marks for a student from 10 questions are X = 0, 1, 2, 3. . . 10.

P(X = x) = $\binom{n}{x} p^x q^{n-x}$

i. P(X = 3) = $\binom{10}{3} (0.25)^3(0.75)^7 = 0.250$

ii. P(X = 8) = $\binom{10}{8} (0.25)^8 (0.75)^2 = 0.00386$

iii. P(X >= x) = 1 - P(X < x). Hence P(X >= 3) = 1 – P(X < 3)
= 1 – {P(X = 0) + P(X = 1) + P(X = 2)}

$P(X = 0) = \binom{10}{0}(0.25)^0(0.75)^{10} = 0.0563$

$P(X = 1) = \binom{10}{1}(0.25)^1(0.75)^9 = 0.1877$

$P(X = 2) = \binom{10}{2}(0.25)^2(0.75)^8 = 0.2816$

.'. $P(X >= 3) = 1 - (0.0563 + 0.1877 + 0.2816) = 0.4744$

The mean = np = 2.5. The variance = npq = 1.875

Example 2: Suppose that a population of size N = 500 consists of 300 dominants and 200 recessive. For a sample of size n = 10, calculate the probabilities:-

a) Exactly 2 individuals will be recessive.
b) At least 2 individuals will be recessive.
c) At most 1 individual will be recessive.(Exercise)
d) At most 5 individuals will be recessive.(Exercise)

Let X = recessive, p = probability of recessives = 200/500 = 2/5.

a) $P[X=2] = \binom{10}{2}(2/5)^2(3/5)^8 = 0.1210$

b) $P[X \geq 2] = 1 - P[X = 0] - P[X = 1]$, but $P[X=0] = \binom{10}{0}q^{10} = 0.006047$

$P[X=1] = \binom{10}{1}pq^9 = 0.040320$.

Hence, $P[X \geq 2] = 1 - (0.006047 + 0.040320) = 0.9536$

Mean = np = 4 V(X) = npq = 12/5

The binomial distribution approaches normal distribution as the number of trials n tends to large ($n \to \infty$) for any fixed value of p. A rule of thumb is that for p < 0.5, the normal approximation is adequate if np

>5. Departures from the given conditions result in less accurate approximations. When n is very large and p is very small ($n\to\infty$ &$p\to 0$) the binomial distribution approaches Poisson distribution.

iii. Poison Distribution

The Poisson distribution is also used to represent the probability distribution of a discrete random variable. It is employed in describing random events that occur rarely over a continuum of time or space, such as number of car accident in certain road corr-section, number of errors in digital communication, number of type fill errors, etc. The Poisson distribution bears a close similarity to the binomial distribution. Suppose that we are interested in the number of occurrences of an event E in a time period of length t. This time period can be split into n equal intervals, each of length t/n. These n intervals can be treated as n trials by Bernoulli process. But there is difficult. Since the event occurs at various points of time, it can occur twice or more in one of the trials of length t/n. In case of binomial distribution the event is dichotomous, and hence there is no possibility of such multiple occurrences within a single trial. In order to overcome this difficulty we make n larger and larger. When n is large, the trials are shorter in terms of length of time. As a result, the probability of occurrence of an event in a single trial would be smaller. It is equivalent of saying that it is a rare event. The binomial distribution can still be used to represent the distribution of such random events. However, the computations become tedious since n is very large. This can be explained by example.

Suppose that the number of insects caught in a trap is being studied and that the data are collected on the number of insects caught per hour. Assume that the probability that an insect will be caught in any single minute is 0.06. Assume further that the events of insects being trapped are mutually independent and the probability p = 0.06 remains same for all the minutes. We may use the binomial distribution to calculate the number of insects caught per hour by considering each minute as a separate Bernoulli trial.

If x is the number of insects caught in a minute then we have: $P[X=x] = \binom{60}{x}(0.06)^x(0.94)^{60-x}$

Instead of dividing the hour into minutes the seconds may be used as basic units. Then the value of p would be reduced to, p=0.06/60=0.001. Considering each second as a Bernoulli trial, we would have a sample size $60\times 60=3600$ for a period of one hour. The binomial distribution would now be:

$$P[X=x] = \binom{3600}{x}(0.001)^x(0.999)^{3600-x}$$

Thus when n becomes larger and larger the computations using binomial become tedious. Fortunately, it has been shown by Poisson that the value of $\binom{n}{x}p^x q^{n-x}$ approaches the value of $\frac{(np)^X e^{-np}}{x!}$, when n becomes large and p becomes small in such a way that the equality, np = λ is maintained.

The probability mass function of Poisson distribution is given by:

P[X=x] = $\frac{e^{-\lambda}\lambda^x}{x!}$. Where, λ = np = mean number of times an event occurs.

x = the number of times an event occur. e = Naperian base = 2.7182…

The value of $e^{-\lambda}$ can be obtained directly from mathematical tables. In case of Poisson distribution the counts of alternative events, i.e., failures are not of interest. This is a contrast between binomial and Poisson distributions. For Poisson distribution all that we need is np, the mean number of successes. We need not know about n and p individually. Thus, the Poisson distribution is determined by the parameter λ. The special property of Poisson distribution is that its mean and variance are same to λ.

i.e. In magnitude; mean = variance = λ.

Example 3: In Black Lion Hospital, the average new born female baby in every 24 hour is 7. What is the probability that

i. No female babies are born in a day?
ii. Only three female babies are born per day?
iii. Two female babies are born in 12 hours?

In this case λ = 7 per day

No female baby born in a day $\Rightarrow$ P(X = 0) = $\frac{e^{-7}7^0}{0!}$ = e^{-7} = 0.0138189

Only three female babies are born $\Rightarrow$ P(X = 3) = $\frac{e^{-7}7^3}{3!}$ = 0.78998

Two female babies are born in 12 hours → in this case λ = 7⁄2 = 3.5

P(X = 2) = $\frac{e^{-3.5}(3.5)^2}{2!}$ = 0.184959

Example 4: In some experiments it was observed that the incidence of stem fly in black gram was 6 percent. Suppose we examine 50 black gram plants in a field at random. What is probability that at most 3 plants will be found to be affected by stem fly?

The probability that a plant is affected by stem fly is given as 0.06. The number of plants observed (n = 50). Hence, λ = np = 3. The required probability is

$P[X \leq 3] = P[X = 0] + P[X = 1] + P[X = 2] + P[X = 3]$

$P[X = x] = \frac{e^{-\lambda}\lambda^x}{x!} P[X = 0] = \frac{e^{-3}3^0}{0!} = e^{-3}$

$P[X = 1] = \frac{e^{-3}3^1}{1!} = 3e^{-3}$

$P[X = 2] = \frac{e^{-3}3^2}{2!} = 4.5e^{-3}$

$P[X = 3] = \frac{e^{-3}3^3}{3!} = \frac{27e^{-3}}{6} = 4.5e^{-3}$

$\Rightarrow P[X \leq 3] = 13e^{-3}$. From mathematical table it can found that $e^{-3} = 0.0498$.

Therefore $P[X \leq 3] = 13 \times 0.0498 = 0.6474$.

Common Continuous Probability Distributions

i. Normal Distribution

The most important and widely used probability distribution is normal distribution. It is also known as Gaussian distribution. Most of the distributions occurring in practice, for instance, binomial, Poisson, etc., can be approximated by normal distribution. Further, many of the sampling distributions like Student's t, F, & χ^2 distributions tend to normality for large samples. Therefore, the normal distribution finds an important place in statistical inference.

The normal distribution is used to represent the probability distribution of a continuous random variable like life expectancies of some product, the volume of shipping container, etc.

Its probability density function is expressed by the relation,

$$f(X) = \frac{1}{\delta\sqrt{2\pi}} \mathrm{e}^{\frac{-1}{2}\left(\frac{X-\mu}{\delta}\right)^2}$$

Where: π = a constant equaling 22/7, e = Naperian base equaling 2.718, μ = population mean, δ = population standard deviation, x = a given value of the rv in the range $-\infty \le x \le \infty$.

For a normal distribution the frequency curve will be symmetrical or bell shaped. However, not all symmetrical curves are normal. The shape of the normal curve is completely determined by two parameters μ & δ. For any given δ, there can be a number of normal curves each with a different μ. Likewise, for any given μ, there can be a number of normal curves each with a different δ. In order to make such all distributions readily comparable with each other, their individuality as expressed by their mean and standard deviation has to be suppressed. This is done by transforming the normal variable into standard normal variable.

The standard normal variable is denoted by Z and is given by Z = $\frac{X-\mu}{\delta}$.

The distribution of the standard normal variable is known as standard normal distribution and given by:

f(x) = $\frac{1}{\sqrt{2\pi}} \mathrm{e}^{\frac{-1}{2}z^2}$, in the range $-\infty \le z \le \infty$.

For standard normal distribution, $\mu = 0$ *and* $\delta = 1$. Tables are readily available for different values of Z. Because of the symmetrical nature of the normal distribution the tables are presented only for the positive values of Z.

Note: Area under curve is equal to one. But area above or below z = 0 is 0.5

Example 5: On a final examination in mathematics, the mean was 72 and the standard deviation was 15.

i) Determine the standard score of the students receiving the grades:

a) 60 b) 93 c) 72

ii) Determine students' grade that have standard score

a) -1 b) 1.6

iii) Find the probability that any student score between 60 & 93. i.e. P[60< X < 93]

Solution: Where X is mark of student

ia) $Z = \frac{X - \overline{X}}{S} = \frac{60 - 72}{15} = -0.8$

ib) $Z = \frac{93 - 72}{15} = 1.4$

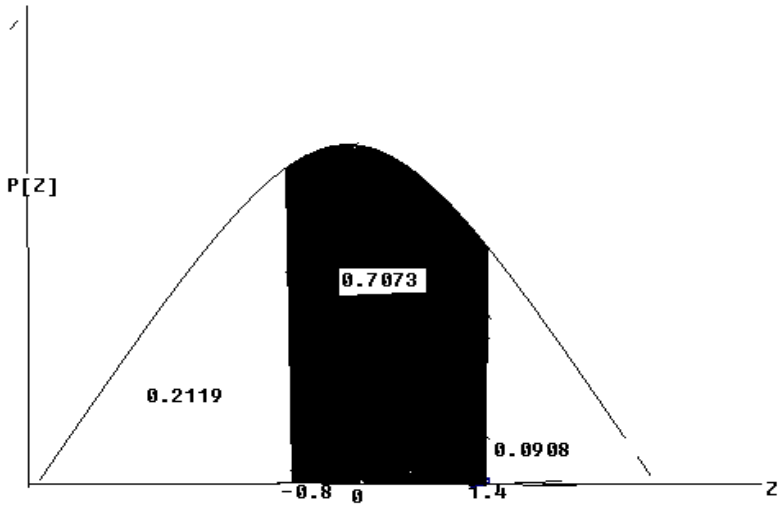
ic) $Z = 0$

iia) $X = \overline{X} + ZS = 72 + -1(15) = 57$

iib) $X = \overline{X} + ZS = 72 + 1.6(15) = 96$

iii) $P[60 \le X \le 93] = P[\frac{60 - \overline{X}}{S} \le \frac{X - \overline{X}}{S} \le \frac{93 - \overline{X}}{S}] = P[-0.8 \le Z \le 1.4] =$

$P[-0.8 \le Z \le 0] + P[0 \le Z \le 1.4] = P[0 \le Z \le 0.8] + P[0 \le Z \le 1.4] = 0.2881 + 0.4192 = 0.7073$ (This is from standard normal table).



Eg. P(0<Z<1.24)= 0.3925

From the table of Normal curves it can be seen that 68.26% of the area lies within the range of $\mu \pm \delta$, 95.46% within the range of $\mu \pm 2\delta$, and 99.74% within the range of $\mu \pm 3\delta$. This is an important property of normal distribution which is frequently used in statistical inference.

The normality of a distribution can be tested using several methods. Three important methods can be mentioned. They are graphic methods, Pearson's test and goodness of fit. On the basis of Pearson's coefficients, β_1 & β_2 we can say whether a distribution is normal or not. If $\beta_1 = 0$ & $\beta_2 = 3$, the distribution is said to be normal. (See chapter four about skewness and kurtosis)

ii. Chi-square Distribution ($\chi^2 - Distribution$)

The square of a standard normal variable is called a chi-square variate with one degree of freedom. Thus if X is a random variable following normal distribution with mean μ and standard deviation σ, then $\frac{(X-\mu)}{\sigma}$ is a standard normal variate. $\left(\frac{X-\mu}{\sigma}\right)^2$ is a chi-square variate with 1 degree of freedom. If $x_1, x_2, \ldots, x_v$ are v independent random variables following normal distribution with means $\mu_1, \mu_2, \ldots, \mu_v$ and standard deviations $\sigma_1, \sigma_2, \ldots, \sigma_v$ respectively then the variate $\chi^2 = \left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 + \cdots + \left(\frac{x_v-\mu_v}{\sigma_v}\right)^2 = Z_1^2 + Z_2^2 + \cdots + Z_v^2 = \sum Z_i^2$ this is the sum of the square of v independent standard normal variates, follows chi-square distribution with v degree of freedom.

iii. Student's t Distribution

In probability and statistics **Student's *t*-distribution** (or simply the ***t*-distribution**) is a probability distribution that arises in the problem of estimating the mean of a normally distributed population when the sample size is small. It is the basis of the popular Student's *t*-tests for the statistical significance of the difference between two sample means, and for confidence intervals for the difference between two population means. The *Student's t-distribution* is a special case of the generalized hyperbolic distribution.

Suppose X_1, ..., X_n are independent random variables that are normally distributed with expected value μ and varianceσ^2. Let the sample mean and sample variances are:

$$\overline{X}_n = (X_1 + \cdots + X_n)/n$$

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

It can be shown that the random variable

$$\frac{(n-1)S_n^2}{\sigma^2}$$

The quantity has a chi-square distribution with $n-1$ degrees of freedom. It is readily shown that the quantity

$$Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

Z is normally distributed with mean 0 and variance 1, since the sample mean $\overline{X}_n$is normally distributed with mean μ and standard error $\sigma/\sqrt{n}$. Moreover, it is possible to show that these two random variables the normally distributed one and the chi-square-distributed one are independent. Consequently the pivotal quantity,

$$T = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}},$$

which differs from *Z* in that the exact standard deviation *σ* is replaced by the random variable S_n, has a Student's t-distribution as defined above. The distribution of *T* is now called the ***t*-distribution**.

CHAPTER SIX

SAMPLING AND SAMPLING DISTRIBUTION

6.1. Basic Concepts of Sampling

When secondary data are not available for the problem under study, a decision may be taken to collect primary data by using any of the methods of data collection discussed in the chapter one. The required information may be obtained by following either the census method or the sample method.

Under the census or complete enumeration survey method, data are collected for each and every unit (person, household, field, shop, factory etc.), as the case may be of the population or universe, which is the complete set of items, which are of interest in any particular situation. If it is not essential to conduct the complete enumeration, then a sample of some suitable size is selected from the population and the study is carried out on the sample. This study is called sample survey. Most of the research work is done through sample surveys.

Sampling is the process or method of selecting sample from the population (universe)

To easily follow sampling remind the terms such as population, sample, statistic, parameter, target population, sampling unit, sampling frame, sample size, etc.

Essentials of Sampling

1. **Saves Time, Energy and Cost:** As the size of the sample is small as compared to the population, the time and cost involved on sample study are much less than the complete counts. For complete count huge funds are required. There is always the problem of finances. A small sample can be studied in a limited time and total cost of sample study is very small. Thus the sample study requires less time and less cost.
2. **Reliability:**If we collect the information about all the units of population, the collected information may be true. But we are never sure about it. We do not know whether the information is true or completely false. Thus we cannot say anything with confidence about the quality of information. We say that the reliability is not possible. This is a very important advantage of sampling. The inference about the population parameters is possible only when the sample data is collected from the selected sample. Sample data is also used to check the accuracy of the census data.

3. **Representativeness:** *A* sample should be so selected that it truly represents the universe otherwise the results obtained may be misleading. To ensure representativeness the random method of selection should be used.
4. **Adequacy:** The size of sample should be adequate; otherwise it may not represent the characteristics of the universe.
5. **Independence:** All items of the sample should be selected independently of one another and all items of the universe should have the same chance of being selected in the sample. By independence of selection we mean that the selection of a particular item in one draw has no influence on the probabilities of selection in any other draw.
6. **Homogeneity:** When we talk of homogeneity we mean that there is no basic difference in the nature of units of the universe and that of the sample. If two samples from the same universe are taken, they should give more or less the same unit.

6.2. Types of Sampling Techniques

The various methods of sampling can be grouped under two broad heads:

1. Probability sampling (also known as random sampling) and
2. Non-probability (or non-random) sampling.

1. ***Random Sampling Techniques*** (*probability sampling techniques*)

Probability sampling methods are those in which every item in the population has a known chance, or probability, of being chosen for sample. This implies that the selection of sample items is independent of the person making the study-that is, the sampling operation is controlled so objectively that the items will be chosen strictly at random. It may be noted that the term random sample is not used to describe the data in the sample but the process employed to select the sample. Randomness is thus a property of the sampling procedure instead of an individual sample. As such, randomness can enter processed sampling in a number of ways and hence random samples may be of kinds.In other words, there is no personal bias regarding the selection. Five common random sampling techniques are:

- Simple Random sampling
- Systematic Random sampling
- Stratified Random sampling
- Cluster Random sampling
- Multi-stage sampling

2. *Non-Random Sampling Technique*

Non-random sampling is a process of sample selection without the use of randomization. In other words, a non-random sample is selected basis other than the probability consideration such as convenience, judgment, voluntary, quota, etc. In non-probability sampling techniques all elements of a population have no known chance of inclusion or some outcomes have a zero chance of being selected as a sample. The most familiar examples of non-random sampling techniques are:

- Judgment (Purposive) sampling
- Convenience sampling
- Quota sampling
- Volunteer sampling

Note: The most important difference between random and non-random sampling is that whereas the pattern of sampling variability can be ascertained in case of random sampling. In non-random sampling, there is no way of knowing the pattern of variability in the process.

Simple Random Sampling

Simple random sampling refers to that sampling technique in which each and every unit of the population has an equal opportunity of being selected in the sample. In simple random sampling which items get selected in the sample is just a matter of chance-personal bias of the investigator does not influence the selection. It should be noted that the word random does not mean 'haphazard' or 'hit-or-miss'-it rather means that the selection process is such that chance only determines which items shall be included in the sample. As pointed out by Chou, when a sample of size n is drawn from a population with *N* elements, the sample is a simple random sample if any of the following is true. And, if any, the following are true:

- All *n* items of the sample are selected independently of one another and all *N* items in the population have the same chance of being included in the sample. By independence of selection we mean that the selection of a particular item in one draw has no influence on the probabilities of selection in any other draw.

- At each selection, all-remaining, items in the population have the same chance of being drawn. If sampling is made with replacement, *i.e.* when each unit drawn from the population is returned prior to drawing the next unit each item has a probability of 1/ N of being drawn at each selection.

If sampling is without replacement, *i.e.* when each unit drawn from the population is not returned prior to drawing the next unit, the probability of selection of each item remaining in the population at the first draw is 1/N, at the second draw is 1/(N-1),at the third draw is 1 / (N-2), and so on. It should be noted that sampling with replacement has very limited and special use in statistics-we are mostly concerned with sampling without replacement.

- All the possible samples of a given size *n* are equally likely to be selected.

What Methods to Ensure Randomness?

To ensure randomness of selection one may adopt either the lottery method or table of random numbers.

Lottery Method: This is a very popular method of taking a random sample. Under this method, all items of the universe are numbered or named on separate slips of paper of identical size and shape. These slips are then folded and mixed up in a container or drum. A blindfold selection then made of the number of slips required to constitute the desired sample size. The selection of items thus depends entirely on chance. This method would be quite clear with the help of an example. If we want to take a sample of 10 persons out of a population of 100, the procedure is to write the names of the 100 persons on separate slips of paper, fold these slips, mix them thoroughly and then make a blindfold selection of 10 slips. The above method is very popular in lottery draws where a decision about prizes is to be made. However, while adopting lottery method it is absolutely essential to see that the slips are of identical size, shape and color, otherwise there is a lot of possibility of personal prejudice and bias affecting the results.

The remaining both random Sampling methods (systematic, stratified, cluster, and multi-stage sampling) and non-random sampling methods (Judgment, Quota, Convenience, and Volunteer sampling) are left for you as reading assignment.

6.3. Sampling Distribution

Sampling distribution is not the distribution of sample, but the distribution of the statistic. That means if we take many samples of equal sizes and get statistic for each of those samples; the distribution of all these statistics is sampling distribution.Often we are interested in sampling distribution of common statistics of the interest such as sample mean, sample proportion, sample variance, their functional form.

Sampling Distribution of the Sample Mean

A sampling distribution of the sample mean is a distribution obtained by using the means computed from random samples of a specific size taken from a population. To illustrate the basic idea of sampling distribution, we will consider very small hypothetical populations of 5 farmers, who use fertilizer in their farming. Suppose the amount of fertilizer used (in kg) by each of the farmer is 70, 78, 80, 80, and 95. Then, we obtain the following characteristics of the population

A. Population Parameters: Let Y_i denotes the amount of fertilizer used by each farmer $i = 1, 2, \ldots, 5$ $N = 5$,

$Y = \sum Y_i = 70 + 78 + 80 + 80 + 95 = 403$ $\mu = \frac{Y}{N} = 80.6$, which is the population mean

$\delta_y^2 = \frac{\sum(Y_i - \bar{Y})^2}{N} = \frac{(70-80.6)^2+(78-80.6)^2+\cdots+(95-80.6)^2}{5} = 65.44$

This population characteristics are mostly unknown and should be estimated from the sample

B. Possible samples and Sampling Distribution

Estimate the population parameters from a sample of three farmers that are randomly select from population of 5 using sampling without replacement. The total number of possible samples is 10, and is shown in the following table with their means. Let F_i represents the i^{th} framer.

No	Sample	Amount in the sample	Sample Mean $\bar{y}_k$
1	$F_1F_2F_3$	70, 78, 80	76.0
2	$F_1F_2F_4$	70, 78, 80	76.0
3	$F_1F_2F_5$	70, 78, 95	81.0
4	$F_1F_3F_4$	70, 80, 80	76.67
5	$F_1F_3F_5$	70, 80, 95	81.67
6	$F_1F_4F_5$	70, 80, 95	81.67
7	$F_2F_3F_5$	78, 80, 80	79.33
8	$F_2F_3F_4$	78, 80, 95	84.33
9	$F_2F_4F_5$	78, 80, 95	84.33
10	$F_3F_4F_5$	80, 80, 95	85.0

The sample mean ($\bar{y}_k$) is obtained by dividing the sum of the three weights of fertilizer in that sample by 3. For example, the mean of the first sample is $\frac{70+78+80}{3} = 76.00$.Using the value of $\bar{y}_k$ as given above, we record the frequency distribution of $\bar{y}_k$. Then, we obtain the probabilities of the random variable $\bar{y}_k$ by dividing the frequency of $\bar{y}_k$ by the sum of the frequencies. These are listed below

Values of $\bar{y}_k$	Frequencies ( )	$\bar{y}_k$	Probability $(\bar{y}_k)$
76.0	2	152.00	0.2
79.67	1	76.67	0.1
79.33	1	79.33	0.1
81.00	1	81.00	0.1
81.67	2	163.34	0.2
84.33	2	168.66	0.2
85.00	1	85.00	0.1
Total	10	806.00	1

This table gives the sampling distribution of $\bar{y}_k$. If we draw just one sample of three framers from the population of five farmers, we may draw any one of the 10 possible farmers. Hence, the sample mean $\bar{y}_k$ can assume any of the values listed above with the corresponding probabilities. For example, the probability of the mean 81.67 is$P(\bar{y}_k = 81.67) = 0.2$. Therefore, the sample average,$\bar{y}_k$, is a random variable that depends on which sample is selected. The value varies from 76.00 to 85 which are lower or higher than the population mean $\mu = 80.6$. The average of the estimates of all possible samples for any sample size is the true population value. That is, the expected value of$\bar{y}_k$, denoted by $E[\bar{y}_k]$, taken over all possible samples equals population mean, i.e., $[\bar{y}_k] = \mu$, in which$E[\bar{y}_k] = \frac{\Sigma f x \bar{y}_k}{\Sigma f} = \frac{806}{10} = 80.6$.

If$y \sim N(\mu, \sigma^2)$, then sample mean $\bar{y} \sim N(\mu, \ \sigma^2/n)$.

Sample Size Determination

An important decision that has to be taken in adopting a sampling technique is about the size of the sample. Size of sample means the number of sampling units selected from the population for investigation. Experts have expressed different opinions on this point. For example, some have

suggested that the sample size should be 5 per cent of the size of population while others are of the opinion that sample size should be at least 10 per cent. However, these views are of little use in practice because no hard and fast rule can be laid down that sample size should be 5 per cent, 10 per cent or 25 per cent of the universe size. It may point out that more size alone does not ensure representativeness. A smaller sample, but well selected sample, may be superior to a larger but badly *selected* sample.

If the size of the sample is small it may not represent the universe and the inference drawn about the population may be misleading. On the other hand, if the size of sample is very large, it may be too burdensome financially, require a lot of time and may have serious problems of managing it. Hence the sample size should neither be too small nor too large. It should be 'optimum'. Optimum size, according to Parten, is one that fulfills the requirements of efficiency, representativeness, reliability and flexibility.

The following factors should be considered while deciding the sample size: the size of the universe,the resources available, the degree of accuracy or precision desired, homogeneity or heterogeneity of the universe, nature of study, and method of sampling adopted, nature of respondents.

To determine sample size a number of formulae have been devised for determining the sample size depending upon the availability of information. A few formulae are given below:

$$n=\left(\frac{zs}{d}\right)^2$$

Where: n = Sample size, z= Value at a specified level of confidence or desired degree of precision, *s*= Standard deviation of the population, *d*= Difference between population mean and sample mean.

Example: Determine the sample size if s = 6, population mean = 25, sample mean = 23 and the desired degree of precision is 99 percent.

$$n=\left(\frac{ZS}{d}\right)^2,$$ s= 6, d= 25-23 = 2, z = 2.576 (at 1% level the z value is 2.576)

Substituting the values: $n=\left[\frac{2.576\times 6}{2}\right]^2 = 7.728^2 = 59.72 \; or \; 60$

CHAPTER SEVEN
REGRESSION ANDCORRELATION ANALYSIS

7.1 Introduction

Regression analysis, in the general sense, means the estimation or prediction of the unknown values of one variable from known values of the other variable .In regression analysis there are two types of variables. The variable whose value is influenced or to be predicted is called dependent (regressed or explained) variable, and the variable which influences the values or is used for prediction is called independent variable (regressor or predictor or explanatory). If the Regression curve is a straight line, we say that there is linear relationship between the variables under study, non-linear elsewhere.

When only two variables are involved, the functional relationship is known as simple regression. If the relationship between the two variables is a straight line, it is known as simple linear regression; otherwise it is called as simple non-linear regression. When there are more than two variables and one of them is assumed dependent upon the other, the functional relationship between the variables is known as multiple regressions. Moreover, correlation analysis is concerned with mathematical measure of the extent or degree of relationship between two variables.

Regression analysis is performed if one wants to know relationship between

a) Income –consumption
b) Sales of ice-cream –with temperature of the day
c) Industrial production and consumption of electricity
d) The yield of crops, amount of rainfall, type of fertilizer, humidity.
e) Weight and height, age and strength, blood pressure and time of exercise.

7.2. Simple Linear Regression (SLR)

Simple linear regression is refers to the linear relationship between two quantitative variables in such way that one of them influences the other. The simple linear regression of Y on X in the population is given by $Y = \alpha + \beta X + \varepsilon$

Where, α = y intercept (Value of Y when X is zero)

β = slope of the line or regression coefficient (change in Y for unit change of value of X)

ε = Error term = (y- $\hat{y}$) is unexplained part by the X. where $\hat{y} = \hat{\alpha} + \hat{\beta}x$

Basic Assumptions of SLR

1. There is linear relationship between dependent variable y and explanatory variable x
2. Expected value of error term is zero and its variance is constant (δ2)
3. Error term is approximately normally distributed with mean zero and constant variance (δ^2).
4. The dependent variable has normal distribution with mean $y = \alpha + \beta x$ and variance δ^2
5. Data on independent variables are fixed numbers.

Estimation of the regression Parameters

The y-intercept α and the slope β are the population parameters. We generally obtain the estimates of α *and* β from the sample. The estimator of α *and* β are denoted by a and b, respectively. Thus the fitted regression line is

$\hat{y}$ = a+ bx

The values of a and b are obtained using the method of least squares. According to the principle of least squares, one should select a and b such that $\sum e^2$ will be as small as possible, that is, we minimize

SSE= $\sum e^2$ = S= $\sum [y - (a + bx)]^2$

To minimize this function, first we take the partial derivatives of SSE with respect to a and b. Then the partial derivatives are equal to zero separately. These will result in the equations known as *normal equations.*

For the straight line, y= a+ bx the normal equations are

$$\sum y = na + bx$$

$$\sum xy = a\sum x + b\sum x^2$$

By solving these normal equations, we can get the values of a and b as follow.

The best estimate of β is given by:

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

The best estimate of α is given by:

$a = \bar{y} - b\bar{x}$

The regression line indicates the average value of the dependent variable Y associated with a particular value of the independent variable X. The slope b, hereafter referred to as regression coefficient which indicates the change in Y with a unit change in X.

Example 7.1 The following table gives the ages and blood pressure of 10 women

Age(x)	56	42	36	47	49	42	60	72	63	55
Blood pressure(y)	147	125	118	128	145	140	155	160	149	150

a) Determine the least square regression equation of blood pressure on age of women
b) Estimate the blood pressure of a women whose age is 45 years.

Solution

$\sum x = 522$ $\sum y = 1417$ $\sum xy = 75188$ $\bar{x} = 52.2$

$\sum x^2 = 28348$ $\sum y^2 = 202493$ $\bar{y} = 141.7$

The estimated value of a and b can be obtained as respectively

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{10 * 75188 - 522 * 1417}{10 * 28348 - (522)^2} = 1.11$$

The best estimate of α is given by

$$a = \bar{y} - b\bar{x} = 141.7 - (1.11)(52.5) = 83.76$$

a) The least square regression equation of blood pressure on age of women is given by

$\hat{y}$ =a + bx , where $\hat{y}$ is estimated blood pressure and x is age of woman

=83.76+1.11x

b) Estimated blood pressure of awoman whoseage 45 years is given as follows.

When x= 45

$\hat{y}$ =a + bx

=83.76+1.11*45=133.71

Example 7.2: From the following data obtain the regression equation of Y on X

Sales(X) :	91	97	108	121	67	124	51	73	111	57
Purchase(Y)):	75	75	69	97	70	91	39	61	80	47

Solution

n= 10 , $\sum x = 900$, $\sum y = 700$, $\sum xy = 66900$, $\sum x^2 = 87360$

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{10 * 66900 - 900 * 700}{10 * 87360 - (900)^2} = 0.61$$

$$a = \bar{y} - b\bar{x} = \frac{1}{n}(\sum_{i=}^{n} y_i - b\sum_{i=1}^{n} x_i) = \frac{1}{10}(700 - 0.61 \times 900) = 15.1$$

$\hat{y} = 15.1 + 0.61x$

7.3. Correlation Analysis

Correlation analysis is concerned with measuring the strength (degree) of the relationship between two or more variables. Some of correlations are simple correlation, partial correlation, and autocorrelation. Simple correlation is used if we are interested in knowing the extent of interdependence between two variables or to measure degree of linear relation between two variables.

a) Karl Pearson's Coefficient of Simple Correlation

The Karl Pearson correlation coefficient denoted by $r(x, y)$ or r_{xy} or simply r, is defined as the ratio of the covariance between X and Y to the product of their standard deviations:

$$r = \frac{\text{cov}(x, y)}{\delta_x \delta_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

The simplified formula used for computational purpose is

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}} \sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}} = \frac{n\sum xy - \sum x \sum y}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}}$$

An increase in one variable may cause an increase in the other variable, or a decrease in one variable may cause decrease in the other variable. When the variables move in the same direction like this they are said to be positively correlated. The positive correlation may be termed as direct correlation. If a decrease in one variable causes an increase in the other variable or visa versa, the variables are said to be negatively correlated. The negative correlation may be termed as inverse correlation. In case the two variables are not at all related they are said to be independent or uncorrelated.

Example

i. Amount of rainfall and yield of crop(up to a point) has positive correlation

ii. Price and demand of a commodity has negative correlation.

Properties of Simple Correlation Coefficient

- Coefficient of correlation lies between $-1 \leq r \leq 1$
- If r =0 indicate that there is no linear relationship between two variables.
- If r = -1 or +1 indicate that there is perfect negative (inverse) or positive (direct) linear relationship between two variables respectively.
- A coefficient of correlation(r) that is closes to zero shows the relationship is quite weak, whereas r is closest to +1 or -1, shows that the relationship is strong.

Note that

- ❖ The strength of correlation does not depend on the positiveness and negativeness of r.
- ❖ The slope of simple linear regression (coefficient of regression) and correlation coefficient should be the same in sign.

The correlation between two variables is linear if a unit changes in one variable result in a constant change in the other variable. Correlation can be studied through plotting scattered diagrams

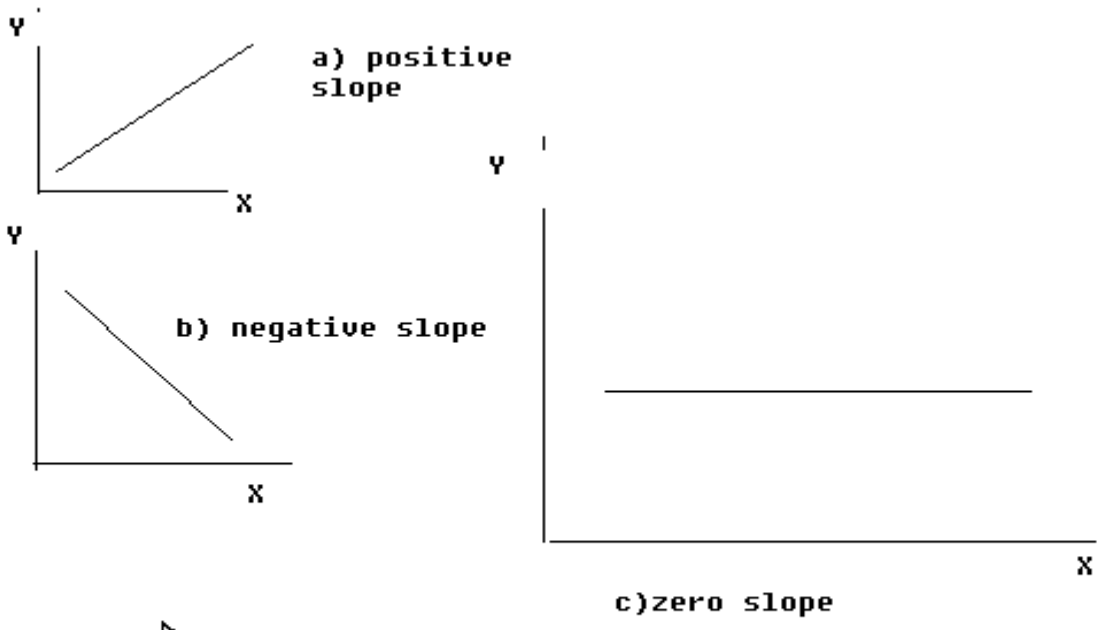


Figure: The slopes of liner regression lines.

Example 7.3 Calculate simple correlation coefficient (r) for the data on advertising and sales expenditure and interpret it.

Advertising(x): 39 65 62 90 82 75 25 98 36 78

Sales (y) : 47 53 58 86 62 68 60 91 51 84

$\sum x = 650, \sum y = 660, \sum xy = 45604, \sum x^2 = 47648, \sum y^2 = 45784$

$$r = \frac{10*45604 - 650*660}{\sqrt{(10*47648 - (650)^2)(10*45784 - (660)^2)}} = \frac{27040}{\sqrt{53980*22240}} = 0.78 \approx 0.8$$

There is strong positive (direct) linear relationship between sales and advertisement since simple correlation coefficient approaches to 1

Example 7.4: Calculate and interpret simple correlation coefficient for data on blood pressure and age of 10 women

Age(x)	56	42	36	47	49	42	60	72	63	55
Blood pressure(y)	147	125	118	128	145	140	155	160	149	150

$\sum x = 522$ $\sum y = 1417$ $\sum xy = 75188$ $\bar{x} = 52.2$

$\sum x^2 = 28348$ $\sum y^2 = 202493$ $\bar{y} = 141.7$

$$r = \frac{10*75188 - 522*1417}{\sqrt{(10*28348 - (522)^2)(10*202493 - (1417)^2)}} = \frac{12206}{\sqrt{10996*17041}} = 0.89 \approx 0.9$$

- There is strong direct linear relationship between blood pressure and age of women, since correlation coefficient approaches to +1.

Coefficient of Determination (R^2)

It is defined as the proportion of the variation in the dependent variable Y that is explained, or accounted for, by the variation of the independent variable X. Its value is the square of the coefficient of correlation, thus we denote it by r^2 and it is usually expressed in the form of percentage.Example compute and interpret coefficient of determination for above example on age and blood pressure.

Given that simple correlation coefficient between blood pressure and age is 0.89, hence coefficient of determination is square of the coefficient of correlation $(r^2)=(0.89)^2=79.21\%$ which implies that 79.21% variation in the blood pressure of women is accounted for, by the variation of the age of women.

Rank Correlation

Sometimes we come across statistical series in which the variables under consideration are not capable of quantitative measurement, but can be arranged in serial order. This happens when we dealing with qualitative characteristics(attributes) such as beauty, efficient ,honest ,intelligence ,….etc., in such case one may rank the different items and apply the spearman method of rank difference for finding out the degree of relationship. The greatest use of this method(rank correlation) lies in the fact that one could use it to find correlation of qualitative variables, but since the method reduces the amount of labor of calculation ,it is sometimes used also where quantitative data is available. It is used when statistical series are ranked according to their magnitude and the exact size of individual item is not known. Spearman's correlation coefficient is denoted by r_s. If the ranks are given, denote the difference R_{1i} - R_{2i} by di and obtain the total of d_i. Then the following formula is applied

$$r_s = 1 - \left[\frac{6\sum d^2}{n\left(n^2 - 1\right)} \right]$$

If the actual data is given, rank it in ascending or descending order and follow the above procedures.

- Note that the values of rank correlation (r_s),.also lies between -1 and +1 inclusive.

Example 7.5 Ten competitors in a beauty contest are ranked by two judges in the following order. Compute and interpret opinion of two judges with regard to beauty out looking.

1^{st} judge(x)	1	6	5	10	3	2	4	9	7	7
2^{nd} judge(y)	3	5	8	4	7	10	2	1	6	9
d=(x-y)	-2	1	-3	6	-4	-8	2	8	1	-2
d^2	4	1	9	36	16	64	4	64	1	4

$\sum d^2 = 203$

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6*203}{10(10^2-1)} = -0.2303$$

- Hence the pair of judges has opposite (divergent) tastes for beauty since rank correlation coefficient is negative.

Example 7.6 Calculate rank correlation coefficient between advertisement cost and sales from the following data and interpret it.

Advertisement(x): 39 65 62 90 82 75 25 98 36 78

Sales (y) : 47 53 58 86 62 68 60 91 51 84

Solution

Rank of X : 8 6 7 2 3 5 10 1 9 4

Rank of Y : 10 8 7 2 5 4 6 1 9 3

d = x - y :-2 -2 0 0 -2 1 4 0 0 1

d^2 :- 4 4 0 0 4 1 16 0 0 1

$\sum d^2 = 30$

$$n = 10 \Rightarrow r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\ x\ 30}{10(10^2-1)} = 1 - \frac{2}{11} = \frac{9}{11} = 0.82 \qquad r_s = 0.82$$

It implies that there is strong positive linear relationship between advertisement cost and sales since correlation coefficient approaches to +1.