



ਜਗਤ ਗੁਰੂ ਨਾਨਕ ਦੇਵ  
ਪੰਜਾਬ ਸਟੇਟ ਓਪਨ ਯੂਨੀਵਰਸਿਟੀ  
ਪਟਿਆਲਾ

The Motto of Our University  
(SEWA)

SKILL ENHANCEMENT

EMPLOYABILITY

WISDOM

ACCESSIBILITY

**JAGAT GURU NANAK DEV**  
**PUNJAB STATE OPEN UNIVERSITY, PATIALA**

(Established by Act No. 19 of 2019 of the Legislature of State of Punjab)

**CERTIFICATE/ DIPLOMA IN STATISTICAL ANALYSIS AND  
RESEARCH METHODOLOGY**

**SEMESTER I**

**SARM 1: INTRODUCTION TO STATISTICS**

**Head Quarter: C/28, The Lower Mall, Patiala-147001**  
**Website: [www.psou.ac.in](http://www.psou.ac.in)**

ALL COPYRIGHTS WITH JGND PSOU, PATIALA

The Study Material has been prepared exclusively under the guidance of Jagat Guru Nanak Dev Punjab State Open University, Patiala, as per the syllabi prepared by the Committee of experts and approved by the Academic Council.

The University reserves all the copyrights of the study material. No part of this publication may be reproduced or transmitted in any form.

**COURSE COORDINATOR AND EDITOR:**

Dr. Pinky Sra

Assistant Professor

JGND PSOU, Patiala.



ਜਗਤ ਗੁਰੂ ਨਾਨਕ ਦੇਵ  
ਪੰਜਾਬ ਸਟੇਟ ਓਪਨ ਯੂਨੀਵਰਸਿਟੀ  
ਪਟਿਆਲਾ



**JAGAT GURU NANAK DEV PUNJAB STATE OPEN UNIVERSITY,  
PATIALA**  
**(Established by Act No. 19 of 2019 of the Legislature of State of Punjab)**

## **PREFACE**

Jagat Guru Nanak Dev Punjab State Open University, Patiala was established in December 2019 by Act 19 of the Legislature of State of Punjab. It is the first and only Open University of the State, entrusted with the responsibility of making higher education accessible to all, especially to those sections of society who do not have the means, time or opportunity to pursue regular education.

In keeping with the nature of an Open University, this University provides a flexible education system to suit every need. The time given to complete a programme is double the duration of a regular mode programme. Well-designed study material has been prepared in consultation with experts in their respective fields.

The University offers programmes which have been designed to provide relevant, skill-based and employability-enhancing education. The study material provided in this booklet is self-instructional, with self-assessment exercises, and recommendations for further readings. The syllabus has been divided in sections, and provided as units for simplification.

The University has a network of 110 Learner Support Centres/Study Centres, to enable students to make use of reading facilities, and for curriculum-based counseling and practicals. We, at the University, welcome you to be a part of this institution of knowledge.

Prof. G.S. Batra  
Dean Academic Affairs



**CERTIFICATE/ DIPLOMA IN STATISTICAL ANALYSIS AND  
RESEARCH METHODOLOGY  
SEMESTER I**

**SARM 1: INTRODUCTION TO STATISTICS**

Max. Marks: 100

External: 70

Internal: 30

Pass: 40%

Credits: 6

**OBJECTIVES:**

The objective of the course is to make the students familiar with various techniques used in summarization and analysis of data. The focus will be on providing basic knowledge of statistics, which deals with data, collection of data, analysis, interpretation and representation of data. It deals with how to analyse statistical data properly and understand the role of formal statistical theory and informal data analytic methods.

**INSTRUCTIONS FOR THE PAPER SETTER/ EXAMINER:**

1. The syllabus prescribed should be strictly adhered to.
2. The Question Paper will have 70 Multiple-choice questions (MCQs) and four choices of answers will be there covering the entire syllabus. Each question will carry 1 mark. All questions will be compulsory; hence candidates will attempt all the questions.
3. Paper-setters/Examiners are requested to distribute the questions from Section A and Section B of the syllabus equally i.e., 35 questions from Section A and 35 questions from Section B.
4. The examiner shall give clear instructions to the candidates to attempt questions.
5. The duration of each paper will be two hours.

**INSTRUCTIONS FOR THE STUDENTS**

The question paper shall consist of 70 Multiple-choice questions. All questions will be compulsory and each question will carry 1 mark. There will be no negative marking. Students are required to answer using OMR (Optimal Mark Recognition) sheets.

**SECTION A**

**Unit 1:** Statistics: definition, importance and Scope, limitations, Distrust

**Unit 2:** Collection of Data: Types and Sources

**Unit 3:** Classification and Tabulation of Data

**Unit 4:** Diagrammatic and Graphical presentation of data (with MS-Excel)

## **SECTION B**

**Unit 5:** Sample, Population, Characteristics of good sample, type of sampling techniques, Sampling errors.

**Unit 6:** Measures of Central Tendency- Mean (Direct, Short cut and step deviation methods), Merits & Demerits.

**Unit 7:** Median (Direct, Short cut and step deviation methods) and Mode: Inspection and grouping method, Merits & Demerits

**Unit 8:** Geometric Mean, Harmonic Mean: Meaning, Merits & Demerits.

Note: Statistical analysis should also be taught with the help of MS Excel, SPSS or any other related software tool.

## **Suggested Readings**

- A.M Goon, M.K Gupta and B. Dasgupta, fundamental of statistics Vol-I, World Press Calcutta
- Gupta SC: Fundamental of statistics, S. Chand & Company. New Delhi
- Gupta, SP: Statistical Methods, S. Chand & Company. New Delhi
- Monga, GS: Mathematics and Statistics for Economics, Vikas Publishing House, New Delhi.
- Singh, D. and Chaudhary, F.S. (1986): Theory and Analysis of Sample Survey Designs. New Age International Publishers.
- Cochran, W.G. (1977): Sampling Techniques (3rd edition), Wiley.



**CERTIFICATE/ DIPLOMA IN STATISTICAL ANALYSIS AND  
RESEARCH METHODOLOGY  
SEMESTER I**

**SARM 1: INTRODUCTION TO STATISTICS**

**EDITOR AND COURSE CO-ORDINATOR- DR. PINKY SRA**

**SECTION A**

<b>UNIT NO.</b>	<b>UNIT NAME</b>
<b>Unit 1</b>	Statistics: definition, importance and Scope, limitations, Distrust
<b>Unit 2</b>	Collection of data: Types and Sources
<b>Unit 3</b>	Classification and Tabulation of data
<b>Unit 4</b>	Diagrammatic and Graphical presentation of data (with MS-Excel)

**SECTION B**

<b>UNIT NO.</b>	<b>UNIT NAME</b>
<b>Unit 5</b>	Sample, Population, Characteristics of good sample, type of sampling techniques, Sampling errors
<b>Unit 6</b>	Measures of Central Tendency- Mean (Direct, Short cut and step deviation methods), Merits & Demerits.
<b>Unit 7</b>	Median (Direct, Short cut and step deviation methods) and Mode: Inspection and grouping method, Merits & Demerits
<b>Unit 8</b>	Geometric Mean, Harmonic Mean: Meaning, Merits & Demerits

**COURSE NAME: STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY**

**SARM 1: INTRODUCTION TO STATISTICS**

**SEMESTER I**

**UNIT 1: STATISTICS: DEFINITION, IMPORTANCE AND SCOPE, LIMITATIONS,  
DISTRUST**

**STRUCTURE**

**1.0 Objectives**

**1.1 Introduction**

**1.2 Meaning of Statistics**

**1.2.1 Statistics in Plural Sense**

**1.2.2 Statistics in Singular Sense**

**1.3 Definitions of Statistics**

**1.4 Importance and Scope of Statistics**

**1.5 Limitations of Statistics**

**1.6 Misuse of Statistics**

**1.7 Distrust**

**1.8 Sum Up**

**1.9 Practice Questions**

**1.10 Suggested Readings**

**1.0 OBJECTIVES**

After reading this unit, learner will be able to learn:

- definitions given by different aspects
- statistics along with its applications
- limitations and misuse of statistics



- various types of data used in statistics will also be explored.

## **1.1 INTRODUCTION**

This module is designed to know about the development of statistics using historical background. Statistics is not a subject that can be studied alone, rather it proves to be the basis for almost all other subjects as data handling is essential in almost all fields of life. Due to this reason, a number of definitions are given to statistics. Some of the major fields where statistics are prominently used are planning, finance, business, agriculture, biology, economics, industry, education, etc. Actually, a country's growth is very much dependent on statistics as without statistics it would not be possible to estimate the requirements of the country. However, statistics are based on probabilistic estimations and therefore not actual (in some cases), therefore can't be believed with 100% guarantee. Also, some people may misuse statistics for their own benefit. But still statistics is very essential and very much needed of life. There are various classifications of the data used in statistics viz., continuous, discrete, nominal, ordinal, etc. The data can be used as per requirement for a particular application.

## **1.2 MEANING OF STATISTICS**

Let us look into the meaning of the word 'Statistics'. It conveys different meaning to different people. A common man may simply interpret it as a mass of figures, graphs or diagrams relating to an economic, business or some other scientific activity. However, for an expert, it may also imply a statistical method of investigation in addition to a mere mass of figures. Let us discuss each of these.

### **1.2.1 Statistics in Plural Sense**

Statistics in plural sense means the mass of quantitative information called 'data'. For example, we talk of information on population or demographic features of India available from the Population Census conducted every ten years by the Government of India. Similarly, we can have statistics (quantitative data or simply data) on Also referred to as Statistical Data, Horace Secrist describes statistics in plural sense as follows: "By Statistics we mean aggregates of facts affected to a marked extent .by multiplicity of causes numerically expressed, enumerated or estimated according to reasonable standard of accuracy, collected in a systematic manner for a pre-determined purpose and placed in relation to each other."

This definition of statistics in plural sense highlights the following features:

- a) **Statistics are numerical facts:** In order that information obtained from an investigation can be called as statistics or data, it must be capable of being represented by numbers. The collected data may be obtained either by the measurement of characteristics (like data on heights, weights, etc.) Or by counting when the characteristics (like honesty, smoking habit, beauty, etc.) is not measurable.
- b) **Statistics are aggregates of facts:** Single and unrelated figures even though expressed as quantities are not statistics. For example, in a university examination Mr. Sharma secures 65% marks does not make statistics or data However, if we find that out of 3 lakh university students whose average marks were 55%, Mr. Sharma secured 65% marks, then these figures are statistics. So, no single figure in any sphere of statistical inquiry, say production, employment, wage and income constitutes statistics.
- c) **Statistics are affected to a marked extent by multiplicity of causes:** In physical sciences it is possible to isolate the effect of various forces on a particular event. But in Statistics' facts and figures, that is, the collected information, are greatly influenced by a number of factors and forces working together. For example, the output of wheat in a year is affected by various factors like the availability of irrigation, quality of soils, method of cultivation, type of seed, amount of fertilizer used, etc. In addition to this there may be certain factors which are even difficult to identify.
- d) **Statistics are numerically expressed:** Statistics are statements of facts expressed numerically or in numbers. Qualitative statements like "the students of a school ABC are more intelligent than those of school XYZ" cannot be regarded statistics. Contrary to this the statement that 'the average marks in school ABC are 90% compared with 60% in school XYZ, and that the former had 80% first division compared with only 50% in the latter", is a statistical statement.
- e) **Statistics are enumerated or estimated with a reasonable degree of accuracy:** While enumerating or estimating statistics, a reasonable degree of accuracy must be achieved. The degree of accuracy needed, in an investigation, depends upon the nature and objective of investigation on one hand and upon the time and resources on the other. Thus, it is necessary to have a reasonable degree of accuracy of data, keeping in mind the nature and objective of investigation and availability of time and resources. The degree of accuracy once decided must be uniformly maintained throughout the investigation.

- f) **Statistics are collected in a systematic manner:** Before the collection of statistics, it is necessary to define the objective of investigation. The objective of investigation must be specific and well defined. The data are then collected in systematic manner by proper planning which involves finding of answers to questions such as: Whether to use sample or census investigation, how to collect, arrange, present and analyse data, etc.)
- g) **Statistics should be placed in relation to one another:** Only comparable data make some sense. 'Unrelated and incomparable data are no data. They are just figures. For example, heights and weights of students of a class do not have any relation with the income and qualification of their parents. For comparability, the data should be homogeneous; that is, it should belong to the same subject or class or phenomenon. For example, pocket money of the students of a class is certainly related to the income of their parents. Prices of onions and potatoes in Delhi can certainly be related to their prices in other cities of India.

Thus, it will not be wrong to say that "all statistics are numerical statements of facts but all numerical statements of facts are not statistics".

### **1.2.2 Statistics in Singular Sense**

In the singular sense, Statistics refers to what is called statistical methods which means the ever-growing body of techniques for collection, condensation, presentation, analysis and interpretation of statistical data/quantitative information. In simple language, it means the subject of Statistics like any other subject such as Mathematics or Economics.

We can now take up definitions given by some famous statisticians.

A. L. Bowley gave a few definitions but none of them was complete and satisfactory. However, his two definitions make some sense even though incomplete. For example, he says, "Statistics may be called the science of counting". Here he is emphasizing on enumeration aspect of statistics, which no doubt is important at another place he describes statistics as "the science of measurement of the social organism...". He is also of the view that "Statistics may rightly be called the science of average". Although measurement, enumeration and averages (Arithmetic, Geometric and Harmonic means; Mode and Median which we will discuss in the next Block) are important, yet they are not the only concern of Statistics, as we shall study in the subsequent units.

Croxton and Cowden have put forward a very simple and precise definition of Statistics as

"Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data". This definition lays emphasis on five important aspects, which in fact, constitute the very scope of the subject called Statistics or Statistical Methods. These are:

**A) Collection of data:** In any statistical inquiry, the collection of data is the first basic step. They form the foundation of statistical analysis, and therefore utmost care should be taken in collecting data. Faulty data will certainly lead to misleading results and can do more harm than good. The data can be drawn from two sources:

- Primary source where data are generated by the investigator himself through various methods
- Secondary source where data are extracted from the existing published or unpublished source, that is, from the data already collected by others. It saves a lot of time, effort and money of the investigator; but then he has to be conscious and judicious in their use.

**B) Arrangement of Data:** Data from the secondary source are already arranged or organised like population data from Census of India. A minor rearrangement to suit our needs can be undertaken. However, primary data are in a haphazard form and need some arrangement so that it makes some sense. The steps involved in this process are: -

- Editing: This involves the removal of omissions and inconsistencies involved in the collected information.
- Classification of data: It follows editing. It involves arranging data according to some common characteristic/s. Normally the raw information received from the respondents is put on the master sheets. For example, we may conduct a survey on, say, metal-based engineering industries of Orissa, from where information are collected on capital structure, output of different types of products, employment of unskilled, semi-skilled and skilled workers, cost and price structure, technology aspects, etc. All this information can be put on master sheets.

**C) Tabulation:** It is the last step in the arrangement process. From the master sheets (or coded sheets) information is tabulated in the form of frequency distributions or tables, where information is arranged in columns and rows.

**D) Presentation of data:** After the data have been arranged and tabulated, they can now be

presented in the form of diagrams and graphs to facilitate the understanding of various trends as well as the process of comparison of various situations. Two different types of presentation of data are normally used, these are:

- Statistical tables
- Graphs including line graphs.

**E) Analysis of data:** It is the most important step in any statistical inquiry. A major portion of this course in Statistics is devoted to the methods used for analyzing the collected data to derive some policy conclusions.

### **1.3 DEFINITIONS OF STATISTICS**

Statistics has been defined by the number of authors in different ways. The main reason for the various definitions is the changes that have taken place in statistics from time to time. Statistics, in general, is defined in two different ways viz., as “statistical data”, i.e., based on the numerical statement of data and facts, and as 'statistical methods, i.e., based on the principles and techniques used in collecting and analyzing such data. Some of the important definitions under these two categories are given below.

Statistics as Statistical Data Webster defines Statistics as "classified facts representing the conditions of the people in a State, especially those facts which can be stated in numbers or in any other tabular or classified arrangement." Bowley defines Statistics as "numerical statements of facts in any department of enquiry placed in relation to each other."

A more exhaustive definition is given by Prof. Horace Secrist as follows: "By statistics we mean aggregation of facts affected to a marked extent by multiplicity of causes numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other."

According to Boddington, "Statistics is the science of estimates and probabilities." Again, this definition is not complete as statistics is not just probabilities and estimates but more than that. Some other definitions are: "The science of Statistics is the method of judging collective, natural or social phenomenon from the results obtained from the analysis or enumeration or collection of estimates."- as provided by King. "Statistics is the science which deals with collection, classification, and tabulation of numerical facts as the basis for explanation, description and

comparison of the phenomenon." as given by Lovitt. But the best definition is the one given by Croxton and Cowden, according to whom Statistics may be defined as "the science which deals with the collection, analysis and interpretation of numerical data."

### **CHECK YOUR PROGRESS (A)**

Q1. Statistics in Plural Sense.

Ans: \_\_\_\_\_

---

Q2. Statistics in Singular Sense.

Ans: \_\_\_\_\_

---

Q3: Collection of data.

Ans: \_\_\_\_\_

---

### **1.4 IMPORTANCE AND SCOPE OF STATISTICS**

Statistics is primarily used either to make predictions based on the data available or to make conclusions about a population of interest when only sample data is available. In both cases, statistics tries to make sense of the uncertainty in the available data. When making predictions statisticians determine if the difference in the data points is due to chance or if there is a systematic relationship. The more the systematic relationship that is observed the better the prediction a statistician can make. The more random error that is observed the more uncertain the prediction. Statisticians can provide a measure of the uncertainty to the prediction. When making inferences about a population, the statistician is trying to estimate how good a summary statistic of a sample really is at estimating a population statistic. For computer students, knowing the basic principles and methods in statistics could help them in doing their research work like comparing the speed of internet connection in different countries and the probability of how many times does each experience the same level of internet connection speed in a week, month or year. It could also be helpful in determining the best operating system to use. Whenever there is the need to compare data and know the best option that we should take statistics can give the answer. Statistics is having applications in almost all sciences - social as well as physical such as biology, psychology,

education, economics, business management, etc. It is hardly possible to think of even a single department of human activity where statistics is not involved. It has rather become indispensable in all phases of human endeavor.

- 1. Statistics and Planning:** Statistics is mother of planning. In the modern age which is termed as 'the age of planning', almost all over the world, particularly of the upcoming economies, are resorting to planning for economic development. In order that planning is successful, it must be based soundly on the correct analysis of complex statistical data.
- 2. Statistics and Economics:** Statistical data and technique of statistical analysis have proved immensely useful in solving various economic problems, such as wages, prices, analysis of time series and demand analysis. A number of applications of statistics in the study of economics have led to the development of new disciplines called Economic Statistics and Econometrics.
- 3. Statistics and Business:** Statistics is an essential tool for production control. Statistics not only helps the business executives to know the requirements of the consumers, but also for many other purposes. The success of a business actually depends upon the accuracy and precision of his statistical forecasting. Wrong analysis, due to faulty and inaccurate analysis of various causes affecting a particular phenomenon, might prove to be a disaster. Consider an examples of manufacturing readymade garments. Before starting one must have an overall idea as to 'how many garments are to be manufactured', 'how much raw material and labour is needed for that', and 'what is the quality, shape, color, size, etc., of the garments to be manufactured'. If these questions are not analysed statistically in a proper manner, the business is bound to be failed. Therefore, most of the large industrial and commercial enterprises are employing trained and efficient statisticians.
- 4. Statistics and Industry:** In industry, statistics is very widely used in 'Quality Control'. In production engineering, to find whether the product is conforming to specifications or not, statistical tools, viz. inspection plans, control charts, etc., are of extreme importance.
- 5. Statistics and Mathematics:** Statistics and mathematics are very intimately related. Recent advancements in statistical techniques are the outcome of wide applications of advanced mathematics. Main contributors to statistics, namely, Bernoulli, Pascal, Laplace, De-Moivre, Gauss, R. A. Fisher, to mention only a few, were primarily talented and skilled mathematicians. Statistics may be regarded as that branch of mathematics which provided us with systematic

methods of analyzing a large number of related numerical facts. According to Connor, " Statistics is a branch of Applied Mathematics which specialise in data."

- 6. Statistics and Biology, Astronomy and Medical Science:** The association between statistical methods and biological theories was first studied by Francis Galton in his work in Regression. According to Prof. Karl Pearson, the whole 'theory of heredity' rests on statistical basis. He said, "The whole problem of evolution is a problem of vital statistics, a problem of longevity, of fertility, of health, of disease and it is impossible to discuss the national mortality without an enumeration of the population, a classification of deaths and knowledge of statistical theory." In astronomy, the theory of Gaussian 'Normal Law of Errors' for the study of the movement of stars and planets is developed by using the 'Principle of Least Squares'. In medical science also, the statistical tools for the collection, presentation, and analysis of observed facts relating to the causes and diseases and the results obtained from the use of various drugs and medicines, are of great importance. Moreover, the efficacy of a manufactured drug or injection, or medicine is tested by analyzing the 'tests of significance'.
- 7. Statistics and Psychology and Education:** In education and psychology, too, statistics have found wide applications, e.g., to determine the reliability and validity of a test, 'Factor Analysis', etc., so much so that a new subject called 'Psychometry' has come into existence.
- 8. Statistics and War:** In war, the theory of 'Decision Functions' can be of great assistance to military and technical personnel to plan 'maximum destruction with minimum effort'. Thus, we see that the science of Statistics is associated with almost all the sciences - social as well as physical. Bowley has rightly said, "A knowledge of Statistics is like a knowledge of foreign language or algebra; it may prove of use at any time under any circumstance."
- 9. Statistics and Physical Sciences:** Statistics has proved to be useful in physical sciences like Physics, Geology, Astronomy, Biology, Medicine, etc. A modern doctor relies heavily on the information on various parameters of a patient in diagnosing his disease. These include his body temperature behaviour, blood pressure and blood sugar level, ECG, etc. Doctor needs this information all the more when performing surgery. Further, before introducing a new drug, data are collected and analysed for its effects on rats, monkeys, rabbits, etc. If found statistically satisfactory, the experiments are then conducted on human beings. The efficacy of the medicine is studied statistically. For example, researchers may be interested in finding whether quinine is still effective in the control of malaria with a new strain of mosquito. They



may conduct the experiment on, say, 1000 patients selected at random. If the percentage of success is quite high, researchers may declare that quinine is still effective in the control of malaria.

## 1.5 LIMITATIONS OF STATISTICS

Statistics, with its wide applications in almost every sphere of human activity; is not without limitations. The following are some of its important limitations:

1. **Statistics is not suited to the study of the qualitative phenomenon:** Statistics, being a science dealing with a set of numerical data, is applicable to the study of only those subjects of enquiry which are capable of quantitative measurement. As such; qualitative phenomena like honesty, poverty, culture, etc., which cannot be expressed numerically, are not capable of direct statistical analysis. However, statistical techniques may be applied indirectly by first reducing the qualitative expressions to precise quantitative terms. For example, the intelligence of a group of candidates can be studied on the basis of their scores on a certain test.
2. **Statistics does not study individuals:** Statistics deals with an aggregate of objects and does not give any specific recognition to the individual items of a series. Individual items, taken separately, do not constitute statistical data and are meaningless for any statistical enquiry. For example, the individual figures of agricultural production, industrial output, or national income of any country for a particular year are meaningless unless, to facilitate comparison, similar figures of other countries or of the same country for different years are given. Hence, statistical analysis is suited to only those problems where group characteristics are to be studied.
3. **Measurement errors:** During the data collection process, measurement errors might happen, resulting in inaccurate results being reported. These errors can be caused by factors such as faulty instruments, human error, respondent bias, or misunderstanding of survey questions. The validity and reproducibility of statistical analysis can be impacted by measurement mistakes.
4. **Sampling bias:** Sampling bias occurs when certain groups or individuals are systematically overrepresented or underrepresented in the sample. This may occur as a result of errors in the sampling procedure or nonresponse bias, where some people decide not to take part in the study. Sampling bias can result in erroneous inferences and unreliable generalizations.
5. **Misuse and misinterpretation:** Statistics can be misused or misinterpreted, leading to

incorrect conclusions. Improper statistical analysis, intentional manipulation of data, or selective reporting of results can distort the findings and mislead the audience. It is important to use statistics appropriately and interpret them critically.

- 6. Statistical laws are not exact:** Unlike the laws of physical and natural sciences, statistical laws are only approximations and not exact. On the basis of statistical analysis, we can talk only in terms of probability and chance and not in terms of certainty. Statistical conclusions are not universally true, rather they are true only on average.

## 1.6 DISTRUST OF STATISTICS

Distrust of statistical data refers to a lack of confidence or skepticism toward the information and findings derived from statistical analysis. This distrust can stem from various factors, including concerns about data collection methods, biases in data interpretation, manipulation or misrepresentation of data, or a general skepticism towards the reliability of statistical methods.

There are several reasons why individuals or groups may exhibit distrust of statistical data:

- 1. Methodological concerns:** Some people question the accuracy and reliability of data collection methods used to gather statistical information. They may believe that the sampling methods are faulty, the sample size is too small, or the data collection process is biased.
- 2. Biases and manipulation:** Skepticism can arise when people suspect that statistical data is manipulated or biased to serve a particular agenda. This can occur through selective data presentation, cherry-picking data, or altering the analysis to support a predetermined conclusion.
- 3. Lack of transparency:** When statistical data is not accompanied by transparent and detailed information about the methodology, sources, and assumptions used, it can lead to distrust. People may be suspicious of hidden biases or vested interests that could influence the results.
- 4. Complexity and misinterpretation:** Statistical analysis can be complex, and the interpretation of data requires expertise. Misinterpretation or misrepresentation of statistical findings by individuals or the media can contribute to distrust. People may feel overwhelmed or confused by the numbers, leading them to question their validity.
- 5. Historical mistrust:** Historical events or instances of statistical manipulation can contribute to a general distrust of statistical data. Past cases of misleading or falsified statistics erode public trust and make people more cautious about accepting statistical information at face

value.

It's essential to remember that critical thinking is necessary for analyzing all data, including statistical data, a total rejection of all statistical data can hinder policy creation, decision-making, and the comprehension of many phenomena. Transparency, clear communication, and strong procedures are needed to address the issues with statistical data in order to deliver accurate and trustworthy information.

### **1.7 MISUSE OF STATISTICS**

Statistics is liable to be misused. As they say, "Statistical methods are the most dangerous tools in the hands of the experts. Statistics is one of those sciences whose adepts must exercise the self-restraint of an artist." The use of statistical tools by inexperienced and untrained persons might lead to very fallacious conclusions. One of the greatest shortcomings of statistics is that by just looking at them one can't comment on their quality and as such can be represented in any manner to support one's way of argument and reasoning. The requirement of experience and judicious use of statistical methods restricts their use to experts only and limits the chances of the mass popularity of this useful and important science. It may be pointed out that Statistics neither proves anything nor disproves anything. It is only a tool which if rightly used may prove extremely useful and if misused might be disastrous. According to Bowley, "Statistics only furnishes a tool necessary though imperfect, which is dangerous in the hands of those who do not know its use and its deficiencies." It is not the statistics that can be blamed but those persons who twist the numerical data and misuse them either due to ignorance or deliberately for personal selfish motives. As King pointed out, "Science of Statistics is the most useful servant but only of great value to those who understand its proper use."

### **CHECK YOUR PROGRESS (B)**

Q1. Explain the importance of the statistics

Ans: \_\_\_\_\_  
\_\_\_\_\_

Q2. Give any two limitations of statistics

Ans: \_\_\_\_\_  
\_\_\_\_\_

Q3. Explain the distrust of statistics

Ans: \_\_\_\_\_

---

Q4. Explain how statistics are misused.

Ans: \_\_\_\_\_

---

## **1.8 SUM UP**

A modern man must possess knowledge of Statistics like that of reading and writing. The word Statistics in singular sense implies statistical methods aimed at collecting, arranging, presenting, analyzing and interpreting data. In the plural sense, it means mass of quantitative information like population data. The word statistic (as against statistics) means an estimator obtained from a sample with a purpose to infer about the population value called parameter. Statistics has utility in almost all branches of knowledge. In Economics and Business, it has a special utility. Combination of Economics, Statistics and Mathematics has led to a new subject called Econometrics. In spite of immense utility, some unscrupulous persons have misused statistics driving it to the level that is worse than damned lies. Because of this, sometimes, it has been termed as unscrupulous science.

## **1.9 QUESTIONS FOR PRACTICE**

1. Give a historical background of statistics.
2. Write various definitions of statistics and discuss these definitions in brief.
3. State and explain various applications of statistics.
4. What are the various limitations of statistics?
5. Provide a few examples which can lead to incorrect conclusions due to wrong analysis of statistics.
6. Give any two examples of collecting data from day-to-day life.

## **1.10 SUGGESTED READINGS**

- A. Abebe, J. Daniels, J.W. Mckean, "Statistics and Data Analysis".
- Clarke, G.M. & Cooke, D., "A Basic course in Statistics", Arnold.
- David M. Lane, "Introduction to Statistics".

- S.C. Gupta and V.K. Kapoor, “Fundamentals of Mathematical Statistics”, Sultan Chand & Sons, New Delhi.

**COURSE NAME: STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY**  
**SARM 1: INTRODUCTION TO STATISTICS**  
**SEMESTER I**

**UNIT 2: COLLECTION OF DATA: TYPES AND SOURCES**

**STRUCTURE**

**2.0 Learning Objectives**

**2.1 Introduction**

**2.2 Types of Data Collection**

**2.3 Sources of Data Collection**

**2.4 Collection of Primary Data Survey Techniques**

**2.5 Limitations of Primary data Collection**

**2.6 Collection of Secondary Data Sources**

**2.7 Limitations of Secondary Data**

**2.8 Precautions to Collect Secondary Data**

**2.9 Sum up**

**2.10 Questions for Practice**

**2.11 Suggested Readings**

**2.0 LEARNING OBJECTIVES**

On going through this unit, you will be able to:

- Explain the concept and types of data collection
- Sources of Data collection
- various survey techniques under primary data and secondary data
- limitations of primary data and secondary data
- Precautions to Collect Secondary Data

**2.1 INTRODUCTION**

We face problems in various fields of our life, which force us to think and discover their

solutions. When we are genuinely serious about the solution of a problem faced, a thinking process starts. Statistical Thinking or Statistical Inquiry is one kind of thinking process that requires evidence in the form of some information, preferably quantitative, which is known as data/statistical information. In a statistical inquiry, the first step is to procure or collect data. Every time the investigator may not start from the very beginning. He must try to use what others have already discovered, this will save us in cost, efforts and time.

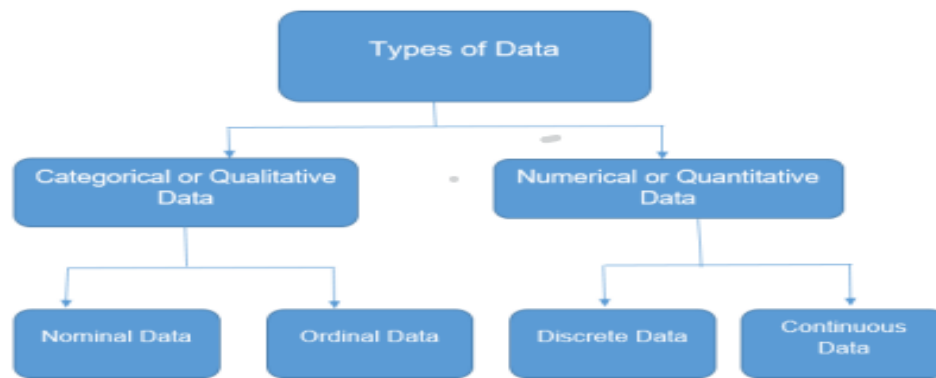
They are collections of any number of related observations with a predetermined goal. We can collect information on the number of T.V. sets sold by a particular salesman or a group of salesmen, on weekdays in different parts of Delhi to study the pattern of sales, lean days, effect of competitive products, income behaviour and other related matters. The information thus collected is called a data set and a single observation a data point. All types of information collected without proper aim or objective is of no use. For example, John's height is 5'6" or monthly wage of Mr. X on 1st January 2004 were Rs.15000/- are not data. Not all quantitative information is statistical. Isolated measurements are not statistical data. Statistics (that is in singular sense) is concerned with collection of data relevant to the solution of a particular problem. According to Simpson and Kafka (Basic Statistics),"Data have no standing in themselves; they have a basis for existence only where there is a problem".

## 2.2 TYPES OF DATA COLLECTION

By now you have known that data could be classified in the following three ways:

- a) Quantitative and Qualitative Data
- b) Sample and Census Data
- c) Primary and Secondary data

**a) Quantitative and Qualitative data:** Quantitative data are those set of information that are quantifiable and can be expressed in some standard units like rupees, kilograms, liters, etc. For example, pocket money of students of a class and the income of their parents can be expressed in so many rupees; the production or import of wheat can be expressed in so many kilograms or lakh quintals; the consumption of petrol and diesel in India as so many lakh liters in one year and so on. In other words, Qualitative data, also known as the categorical data, describes the data that fits into the categories. Qualitative data are not numerical. The categorical information involves categorical variables that describe the features such as a person's gender, home town etc. Categorical measures are defined in terms of natural language specifications, but not in terms of numbers



**1. Qualitative data:** on the other hand, are not quantifiable, that is, cannot be expressed in standard units of measurement like rupees, kilograms, liters, etc. This is because they are 'features', 'qualities', or 'characteristics' like eye colours, skin complexion, honesty, good or bad, etc. These are also referred to as attributes. In this case, however, it is possible to count the number of individuals (or items) possessing a particular attribute.

- **Nominal Data:** Nominal data is one of the types of qualitative information which helps to label the variables without providing the numerical value. Nominal data is also called the nominal scale. It cannot be ordered and measured. But sometimes, the data can be qualitative and quantitative. Examples of nominal data are letters, symbols, words, gender etc. The nominal data are examined using the grouping method. In this method, the data are grouped into categories, and then the frequency or the percentage of the data can be calculated. These data are visually represented using the pie charts.
- **Ordinal Data:** Ordinal data is a type of data that follows a natural order. The significant feature of the nominal data is that the difference between the data values is not determined. This variable is mostly found in surveys, finance, economics, questionnaires, and so on. The ordinal data is commonly represented using a bar chart. These data are investigated and interpreted through many visualization tools. The information may be expressed using tables in which each row in the table shows a distinct category. Quantitative or Numerical Data.

**2. Quantitative Data:** Quantitative data is also known as numerical data which represents the numerical value (i.e., how much, how often, how many). Numerical data gives information about the quantities of a specific thing. Some examples of numerical data are height, length, size, weight, and so on. Quantitative data can be classified into two different types based on the data sets. The two different classifications of numerical data are discrete data and continuous data.



- **Discrete Data:** Discrete data can take only discrete values. Discrete information contains only a finite number of possible values. Those values cannot be subdivided meaningfully. Here, things can be counted in whole numbers e.g., the number of students in the class
- **Continuous Data:** Continuous data is data that can be calculated. It has an infinite number of probable values that can be selected within a given specific range e.g., Temperature range.

The quantitative and qualitative data can be represented as in figure 1.1.

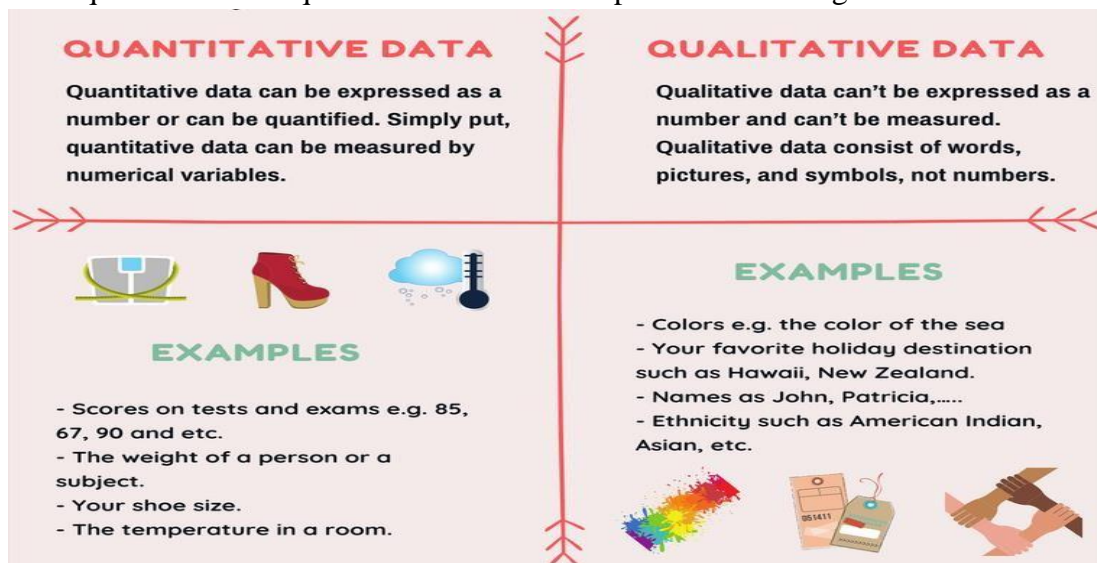




Figure 1.1: Quantitative and Qualitative Data

Figure 1.2 shows the types of qualitative data i.e., discrete and continuous data.

<p><b>DISCRETE</b></p> <p>Discrete data is a count that involves only integers. The discrete values cannot be subdivided into parts. For example, the number of children in a class is discrete data. You can't count 1.5 kids.</p>	<p><b>EXAMPLES</b></p> <ul style="list-style-type: none"> <li>• The number of students in a class.</li> <li>• The number of workers in a company.</li> <li>• The number of home runs in a baseball game.</li> <li>• The number of test questions you answered correctly</li> </ul>	<p><b>PICS</b></p> 
<p><b>CONTINUOUS</b></p> <p>Continuous data could be meaningfully divided into finer levels. It can be measured on a scale or continuum and can have any numeric value. For example, you can measure your height at very precise scales — meters, centimeters, millimeters, etc.</p>	<p><b>EXAMPLES</b></p> <ul style="list-style-type: none"> <li>• The amount of time required to complete a project.</li> <li>• The height of children.</li> <li>• The square footage of a two-bedroom house.</li> <li>• The speed of cars.</li> </ul>	<p><b>PICS</b></p> 

## Figure 1.2: Types of Qualitative Data viz., Discrete and Continuous

Figure 1.3 shows the types of quantitative data i.e. nominal and ordinal data.



Figure 1.3: Types of Quantitative Data viz., Nominal and Ordinal

**b) Sample and Census Data:** Data can be collected either by census method or sample method. Information collected through sample inquiry is called sample data and the one collected through census inquiry is called census data. Population census data are collected every ten years in India.

**c) Primary and Secondary Data:** Primary data are collected by the investigator through field surveys. Such data are in raw form and must be refined before use. On the other hand, secondary data are extracted from the existing published or unpublished sources, that is; from the data already collected by others. The collection of data is the first basic step towards the statistical analysis of any problem. The collected data are suitably transformed and analyzed to draw conclusions about the population.

These conclusions may be either or both of the following:

- (i) To estimate one or more parameters of a population or the nature of the population itself. This forms the subject matter of the theory of estimation.
- (ii) To test a hypothesis. A hypothesis is a statement regarding the parameters or the nature of the population.

## 2.3 SOURCES OF DATA COLLECTION

A pertinent question that arises now is how and from where to get data? Data are obtained through two types of investigations, namely,

1) **Direct Investigation or Primary Data** which implies that the investigator collects information by observing the items of the problem under investigation. As explained above, it is the primary source of getting data or the source of getting primary data and can be done through observation or through inquiry. In the former we watch an event happening, for example, the number and type of vehicles passing through Vijay Chowk in New Delhi during different hours of the day and night. In the latter, we ask questions from the respondents through questionnaire (personally or through mail). It is a costly method in terms of money, time, and effort.

2) Investigation through **Secondary Source** which means obtaining data from the already collected data. Secondary data are the other people's statistics, where other people include governments at all levels, international bodies or institutions like IMF, IBRD, etc., or other countries, private and government research organisations, Reserve Bank of India and other banks, research scholars of repute, etc. Broadly speaking we can divide the sources of secondary data into two categories: published sources and unpublished sources. A) Published Sources

1) Official publications of the government at all levels — Central, State, Union

2) Official publications of foreign countries.

3) Official publications of international bodies like IMF, UNESCO, WHO, etc.

4) Newspapers and Journals of repute, both local and international.

5) Official publications of RBI, and other Banks, LIC, Trade Unions, Stock Exchange, Chambers of Commerce, etc.

6) Reports submitted by reputed economists, research scholars, universities, commissions of inquiry, if made public.

### **Data Collection Methods**

Some main sources of published data in India are Central Statistical Organisation (C.S.O.): It publishes data on national income, savings, capital formation, etc. in a publication called National Accounts Statistics. National Sample Survey Organisation (N.S.S.O.): Under the Ministry of Statistics and Programme Implementation, this organisation provides us with data on all aspects

of the national economy, such as agriculture, industry, labor and consumption expenditure. Reserve Bank of India Publications (R.B.I.): It publishes financial statistics. Its publications are Report on Currency and Finance, Reserve Bank of India Bulletin, Statistical Tables Relating to Banks in India, etc. iv) Labour Bureau: Its publications are Indian Labour Statistics, Indian Labour Year Book, Indian Labour Journal, etc. v) Population Census: Undertaken by the office of the Registrar General India, Ministry of Home Affairs. It provides us with different types of statistics about the population.

#### B) Un-published Sources

- 1) Unpublished findings of certain inquiry committees.
- 2) Research workers' findings.
- 3) Unpublished material found with Trade Associations, Labour Organisations and Chambers of Commerce.

### **CHECK YOUR PROGRESS (A)**

Q1) Explain the term quantitative data?

Ans: \_\_\_\_\_  
\_\_\_\_\_

Q2) What is primary data?

Ans: \_\_\_\_\_  
\_\_\_\_\_

Q3) What is secondary data set?

Ans: \_\_\_\_\_  
\_\_\_\_\_

Q4) Define sample and population.

Ans: \_\_\_\_\_  
\_\_\_\_\_

### **2.4 COLLECTION OF PRIMARY DATA SURVEY TECHNIQUES**

After the investigator is convinced that the gain from primary data outweighs the money cost, effort and time, she/he can go in for this. She/he can use any of the following methods to collect primary data:

- a. Direct Personal Investigation
- b. Indirect Oral Investigation
- c. Use of Local Reports/ agencies to get information
- d. Mailed Questionnaire Method
- e. Schedules sent through enumerators

a) **Direct Personal Investigation:** Here the investigator collects information personally from the respondents. She/ he meets them personally to collect information. This method requires much from the investigator such as:

- She/he should be polite, unbiased and tactful.
- She/he should know the local conditions, customs and traditions
- She/he should be intelligent possessing good observation power. Data Collection Methods
- She/he should use simple, easy and meaningful questions to extract information.

This method is suitable only for intensive investigations. It is a costly method in terms of money, effort and time. Further, the personal bias of the investigator cannot be ruled out and it can do a lot of harm to the investigation. The method is a complete flop if the investigator does not possess the above-mentioned qualities.

**b) Indirect Oral Investigation:** Method This method is generally used when the respondents are reluctant to part with the information due to various reasons. Here, the information is collected from a witness or from a third party who are directly or indirectly related to the problem and possess sufficient knowledge. The person(s) who is/are selected as informants must possess the following qualities:

- They should possess full knowledge about the issue.
- They must be willing to reveal it faithfully and honestly.
- They should not be biased and prejudiced.
- They must be capable of expressing themselves to the true spirit of the inquiry.

**c) Use of Local Reports:** This method involves the use of local newspaper, magazines and journals by the investigators. The information is collected by local press correspondents and not by the investigators. Needless to say, this method does not yield sufficient and reliable data The method is less costly but should not be adopted where high degree of accuracy or precision is required.

**d) Mailed Questionnaire Method:** It is the most important and systematic method of collecting primary data, especially when the inquiry is quite extensive. This method entails creating a questionnaire (a collection of questions pertaining to the research area with a chance for respondents to fill in their replies) and mailing it to the respondents with a deadline for responding quickly. The respondents are asked to extend their full cooperation by providing accurate responses and timely submission of the completed questionnaire. By assuring them that the information they provided in the questionnaire will be kept totally secure and hidden, respondents are also given a sense of security. The investigator typically pays the return postal costs by mailing a self-addressed, stamped envelope to achieve a speedy and better response. Researchers, individuals, non-governmental organisations, and occasionally even the government are involved in this technique.

**e) Schedules sent through Enumerators:** Using enumerators for primary data collection is a common practice in various research studies, surveys, and data collection efforts. Enumerators are individuals responsible for collecting data directly from respondents in the field. This is the method of obtaining answers to the questions in a form that is filled out by the interviewers or enumerators (the field agents who put these questions) in a face-to-face situation with the respondents. The questionnaire is a list of questions that the respondent himself answers in his own handwriting. 'Schedules sent through the enumerators' is the major data collecting technique that is commonly used. This is the case because the earlier ways that have been explained thus far have some drawbacks that this method does not. With the schedule (a list of questions), the enumerators directly contact the respondents, ask them questions, and record their responses.

The questionnaire in primary data is divided into two parts:

- 1) General introductory part which contains questions regarding the identity of the respondent and contains information such as name, address, telephone number, qualification, profession, etc.
- 2) Main question part containing questions connected with the inquiry. These questions differ from inquiry-to-inquiry. Preparation of the questionnaire is a highly specialized job and is perfected with experience. Therefore, some experienced persons should be associated with it.

Drafting and framing a questionnaire is a critical step in primary data collection. A well-designed questionnaire ensures that you gather relevant and reliable data to address your research objectives.

The following few important points should be kept in mind while drafting a questionnaire:

- (i) Clearly outline the research objectives and the specific information you want to collect through the questionnaire. Identify the key research questions that need to be answered.
- (ii) Make sure your questions are easy to understand. Avoid nonsense and complex language. Keep sentences and questions short and to the point.
- (iii) The task of soliciting information from people in the desired form and with sufficient accuracy is the most difficult problem. By their nature people are not willing to reveal any information because of certain fears. Many times they provide incomplete and faulty information. Therefore, it is necessary that the respondents be taken into confidence. They should be assured that their individual information will be kept confidential and no part of it will be revealed to tax and other government investigative agencies. This is very essential indeed. Where providing information is not legally binding, the informant has to be sure and convinced that the results of the survey will help the authorities to frame policies which will ultimately benefit them. It is obvious that some element of good salesmanship is also required in the investigation.
- (iv) Make a decision regarding the questions that will be included in the questionnaire. Typical sorts of queries include:
  - Closed-ended inquiries: Those who respond select from a set of predetermined responses (such as multiple-choice inquiries).
  - Open-ended inquiries: Those that respond provide their own, individual responses.
  - Questions using a Likert scale: Determine if respondents agree or disagree with a statement using a scale (such as 1 to 5).
  - Semantic differential questions: Request a rating on a scale of good to bad or satisfied to dissatisfied from respondents.
- (v) Questions hurting the sentiments of respondents should not be asked. These include questions on his gambling habits, sex habits, indebtedness, etc.
- (vi) Questions involving lengthy and complex calculations should be avoided because they require tedious extra work in which the respondent may lack both interests as well as capabilities.

## **2.5 LIMITATIONS OF PRIMARY DATA COLLECTION**

Primary data refers to data collected firsthand through direct observation, surveys, interviews,

experiments, or other data collection methods. While primary data can be valuable for research and analysis, it also has certain limitations. Here are some common limitations of primary data:

- 1. Cost and time:** Collecting primary data can be a time-consuming and costly process. It requires resources to design research instruments, recruit participants, conduct data collection, and analyze the data. Therefore, primary data collection may be impractical or unaffordable.
- 2. Limited sample size:** Primary data collection often involves a smaller sample size compared to secondary data sources. The sample size may be constrained by factors such as time, budget, or accessibility of the target population. A small sample size may limit the generalizability of the findings to a larger population.
- 3. Sampling bias:** Similar to the limitations of statistics, primary data collection can be susceptible to sampling bias. If the sample is not representative of the population of interest, the findings may not accurately reflect the characteristics or behaviors of the larger population. Careful attention must be given to sampling methods to minimize bias.
- 4. Response bias:** Response bias occurs when participants in a study provide inaccurate or misleading responses. It can be influenced by factors such as social desirability bias (participants providing responses they think are socially acceptable) or recall bias (participants inaccurately remembering past events). Response bias can undermine the validity and reliability of primary data.
- 5. Subjectivity and researcher bias:** Primary data collection methods often involve interaction between the researcher and participants. The subjective interpretation and biases of the researcher can unintentionally influence the data collection process and the responses obtained. Researchers need to be aware of their own biases and take steps to minimize their impact on the data.
- 6. Limited scope:** Primary data collection typically focuses on specific research questions or objectives. While this targeted approach can yield detailed insights into specific areas of interest, it may not capture a broader range of factors or provide a comprehensive understanding of the phenomenon being studied. Using secondary data or employing a mixed-methods approach can help overcome this limitation.
- 7. Ethical considerations:** Primary data collection involves ethical considerations regarding participant privacy, informed consent, and data protection. Researchers must adhere to ethical guidelines and obtain necessary approvals, which can introduce additional time and logistical



constraints.

Understanding these limitations of primary data can help researchers and analysts make informed decisions about data collection methods and consider the strengths and weaknesses of primary data in relation to their research objectives. It may also be beneficial to supplement primary data with secondary data sources to enhance the breadth and depth of the analysis.

### **CHECK YOUR PROGRESS (B)**

Q1. Explain direct personal investigation and indirect oral investigation

Ans. \_\_\_\_\_  
\_\_\_\_\_

Q2. Define the mailed questionnaire method and schedules sent through enumerators

Ans. \_\_\_\_\_  
\_\_\_\_\_

Q3. Give limitations of primary data

Ans. \_\_\_\_\_  
\_\_\_\_\_

## **2.6 COLLECTION OF SECONDARY DATA SOURCES**

The direct investigation, though desirable, is costly in terms of money, time and effort. Alternatively, information can also be obtained through a secondary source. It means drawing or collecting data from the already collected data of some other agency. Technically, the data so collected are called secondary data.

Secondary data sources in statistics refer to existing data that has been collected by someone else or for a different purpose but can be utilized for statistical analysis. These sources provide a wealth of information that can be used to explore research questions, test hypotheses, and derive insights. Here are some common secondary data sources used in statistics:

- 1. Government agencies:** Government agencies at the local, national, and international levels collect and maintain a vast amount of statistical data. Examples include census data, labor statistics, economic indicators, crime rates, health statistics, and demographic information. These datasets are often publicly available and can provide valuable insights into various

social, economic, and demographic trends.

- 2. Research organizations and institutes:** Many research organizations and institutes conduct surveys, studies, and data collection efforts for specific research purposes. These organizations may focus on topics such as education, public health, social issues, or specific industries. Their datasets can provide detailed information on specific domains or research areas.
- 3. International organizations:** International organizations, such as the World Bank, International Monetary Fund (IMF), United Nations (UN), and World Health Organization (WHO), collect and maintain extensive datasets on global development, economics, health, and social indicators. These datasets cover a wide range of countries and can be used for comparative analysis and cross-country studies.
- 4. Academic institutions:** Universities and research institutions often conduct research studies and surveys, resulting in datasets that can be valuable for statistical analysis. These datasets may cover various disciplines, including social sciences, psychology, economics, education, and more. Academic institutions often make their datasets available to researchers, subject to certain restrictions and ethical considerations.
- 5. Nonprofit organizations:** Nonprofit organizations focused on specific causes or social issues often collect data related to their mission. These organizations may conduct surveys, compile reports, or collaborate with other entities to collect data. Their datasets can provide insights into areas such as poverty, environmental issues, human rights, and social justice.
- 6. Commercial data providers:** There are commercial entities that collect, aggregate, and sell datasets on various industries, market trends, consumer behavior, and more. These datasets can be useful for market research, business analytics, and understanding consumer preferences and trends.
- 7. Online platforms and social media:** Online platforms and social media networks generate vast amounts of data. This data includes user-generated content, interactions, behaviors, and demographic information. While accessing and analysing this data may require specific permissions and compliance with privacy regulations, it can offer insights into online behavior, sentiment analysis, and social network analysis.

When using secondary data sources, researchers should consider factors such as the data quality,

reliability, representativeness, and potential limitations or biases. It is essential to critically evaluate the data source and ensure that it aligns with the research objectives and analytical requirements

## **2.7 LIMITATIONS OF SECONDARY DATA**

Although the secondary source is cheap in terms of money, time and effort, utmost care should be taken in their use. It is desirable that such data should be vast and reliable and the terms and definitions must match the terms and definitions of the current inquiry. The suitability of the data may be judged by comparing the nature and scope of the present inquiry with that of the original inquiry. Secondary data will be reliable if these were collected by unbiased, intelligent and trained investigators. The time period to which these data belong should also be properly scrutinized.

Secondary data refers to data that is collected by someone else for a different purpose but can be utilized for research or analysis. While secondary data can be convenient and cost-effective, it also has certain limitations. Here are some common limitations of secondary data collection:

- 1. Lack of control over data collection:** Since secondary data is collected by others, researchers have no control over the data collection process. This can result in data that may not perfectly align with the research objectives or may lack specific variables or measures that the researcher requires. The data may not have been collected with the same level of rigor or precision as desired.
- 2. Data relevance and accuracy:** The relevance and accuracy of secondary data can vary. It may be challenging to find secondary data that precisely matches the research needs, as the data may be outdated or collected using different methodologies. In some cases, the data may contain errors, inconsistencies, or missing values, which can affect its reliability and validity.
- 3. Limited contextual information:** Secondary data may lack detailed information about the context in which it was collected. Understanding the specific circumstances, conditions, or nuances surrounding the data collection process may be crucial for accurate interpretation and analysis. Without sufficient contextual information, the researcher may face challenges in fully understanding and interpreting the data.
- 4. Potential bias and validity concerns:** Secondary data may contain inherent biases or limitations introduced by the original data collection process. The biases could be due to the

research design, sampling methods, or data collection instruments used. Researchers must critically evaluate the reliability and validity of the secondary data source to ensure its suitability for their research objectives.

- 5. Incompatibility and inconsistency:** When working with secondary data from multiple sources, researchers may encounter issues of incompatibility and inconsistency. The data may have been collected using different formats, classifications, or units of measurement, making it challenging to combine or compare the data effectively. Harmonization or standardization efforts may be necessary to address these issues.
- 6. Limited control over variables:** Secondary data may not include all the variables of interest to the researcher. Certain variables that are critical for the research objectives may be missing, limiting the scope of analysis or preventing the investigation of specific relationships or factors.
- 7. Data availability and access:** Accessing certain types of secondary data can be challenging due to restrictions, copyright issues, or proprietary considerations. Researchers may face limitations in obtaining the specific data they need or may need to rely on aggregated or summarized data, which may not provide the level of detail required for the research.

Despite these limitations, secondary data can still be a valuable resource for researchers, providing a foundation for analysis, hypothesis generation, and comparison with primary data. Researchers should critically evaluate the quality and relevance of the secondary data and consider its limitations in the interpretation and analysis process.

## **2.8 PRECAUTIONS TO COLLECT SECONDARY DATA**

According to Prof. A.L. Bowley, "It is never safe to take the published statistics at their face value without knowing their meaning and limitations and it is always necessary to criticize the arguments that can be based upon them." In using secondary data, we should take special note of the following factors.

- 1) Reliable, 2) Suitable, and 3) Adequate.

Firstly, the reliability of data has to be the obvious requirement of any data, and more so of secondary data. The user must make himself/herself sure about it. For this (s)he must check whether data were collected by reliable, trained and unbiased investigators from dependable

sources or not.

Second, we should see whether data belong to almost the same type of class of people or not. (1) To look at and compare the given inquiry's objectives, nature, and scope with the original research. To verify that all of the terms and units were uniformly defined throughout the previous investigation and that these definitions are appropriate for the current investigation as well. For instance, a unit can be defined in multiple ways depending on its context, such as a household, wage, price, farm, etc. The secondary data will be considered inappropriate for the present research if the units were identified differently in the original investigation than what we want. lastly, consider the variations in data collecting periods and consistency of conditions comparing the original investigation and the present investigation.

Third, even if the secondary data are reliable and suitable in, it might not be adequate for the particular inquiry's objectives. This happens if the original data refers to an area or a period that is much larger or smaller than the needed one, or when the coverage given in the initial research was too narrow or too wide than what is desired in the current research. Therefore, it is making sure that due to the gap of time, the conditions prevailing then are not much different from the conditions of today in respect of habits, customs, fashion, etc. Of course, we cannot hope to get exactly the same conditions.

The suitability of data is another requirement. The research worker must ensure that the secondary data he plans to use suits his inquiry. He must match the class of people, geographical area, definitions of concepts, unit of measurement, time and other such parameters of the source he wants to use with those of his inquiry. Not only this, the aim and objectives should also be matched for suitability.

Secondary data should not only be reliable and suitable, but also adequate for the present inquiry. It is always desirable that the available data be much more than required by the inquiry. For example, data on, say, the consumption pattern of a state cannot be derived from the data on its major cities and towns.

### **CHECK YOUR PROGRESS (C)**

Q1. Explain the method of collect secondary data.

Ans. \_\_\_\_\_

\_\_\_\_\_

Q2. Define the Mailed questionnaire method and schedules sent through the enumerator? Give two limitations of primary data.

Ans. \_\_\_\_\_  
\_\_\_\_\_

Q3. Give two limitations of secondary data.

Ans. \_\_\_\_\_  
\_\_\_\_\_

## **2.9 SUM UP**

Data Collection Methods Data / Statistics are quantitative information and can be distinguished as sample or census data; primary or secondary data. We require information for an investigation that can be gathered from either a primary source or a secondary source. Both require statistical surveys, which have two stages: planning and execution. The investigator should choose the primary or secondary sources, census or sample inquiry, type of statistical units and measurement units, level of precision desired, and other factors during the design stage. In the execution stage, the chief investigator has to set up administration, select and train field staff and supervise the entire process of data collection. Using secondary data from published or unpublished sources requires caution because they can lead to a number of problems. The questionnaire method is the most crucial of all survey methods. A questionnaire provides a list of relevant inquiries, which should be short, clear, and of the Yes/No variety with illustrative responses. They shouldn't have a lengthy list. Questions that are private or humiliating should be avoided.

## **2.10 QUESTIONS FOR PRACTICE**

1. What are the techniques to collection of data
2. What do you mean by primary data?
3. What are the sources of primary data?
4. What is questioner. What are the points to kept in mind before drafting questioner?
5. Explain the term secondary data with its sources
6. limitations of primary data and secondary data
7. Precautions to Collect Secondary Data

## **2.11 SUGGESTED READINGS**

- A. Abebe, J. Daniels, J.W. Mckean, "Statistics and Data Analysis".

- Clarke, G.M. & Cooke, D., “A Basic course in Statistics”, Arnold.
- David M. Lane, “Introduction to Statistics”.
- S.C. Gupta and V.K. Kapoor, “Fundamentals of Mathematical Statistics”, Sultan Chand & Sons, New Delhi.

**COURSE NAME: STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY**

**SARM 1: INTRODUCTION TO STATISTICS**

**SEMESTER I**

**UNIT 3: CLASSIFICATION AND TABULATION OF DATA**

**STRUCTURE**

**3.0 Learning Objectives**

**3.1 Introduction**

**3.2 Classification of data**

**3.3 Functions of Classification**

**3.4 Basis of Classification**

**3.5 Frequency Distribution**

**3.5.1 Simple Array**

**3.5.2 Discrete or Ungrouped Frequency Distribution**

**3.5.3 Continuous or Grouped Frequency Distribution**

**3.5.4 Various Forms of Frequency Distributions**

**3.6 Tabulation of Data: Meaning**

**3.7 Parts of a Table**

**3.8 Importance of Tables**

**3.9 Questions for Practice**

**3.10 Suggested Readings**

**3.0 OBJECTIVES**

On going through this Unit, you will be able to explain:

- stages of statistical inquiry after data have been collected



- classification of data
- methods of organizing (classification and arrangement) and condensing statistical data
- concepts of frequency distribution for individual, discrete and continuous series
- Tabulation of data
- Parts of table in statistics

### **3.1 INTRODUCTION**

Data collected either from census or sample inquiry, that is from primary source, are always hotchpotch and in rudimentary form. To start with, they are contained in hundreds and thousands of questionnaires. To make a head and tail out of them, they must be organized, (i.e., classified and arranged) summarised. For this purpose, we can use various methods like preparing master sheets in which various information are recorded directly hm the questionnaires. From these sheets small summary tables can be prepared manually. Now-a-days computers can be used for organisation and condensation of data more swiftly, efficiently and in much less time. Some computer softwares are available which help us to construct various types of graphs and diagrams. Data can be summarized numerically also. Here we use summary measures like measures of central tendency (such as Arithmetic, Geometric and Harmonic Means, Mode and Median); measures of dispersion (such as Range, Quartile Deviation, Mean Deviation, and Standard Deviation); measures of association in bivariate analysis (such as Correlation and Regression), Index Numbers, etc. In this Unit, we plan to discuss how data can be summarized using tables and graphs. It must be kept in mind that a good summarization and presentation of data is not undertaken for its own sake. It is not an end in itself. In fact, it sets the stage for useful analysis and interpretation of data. Again, a good presentation helps us to highlight significant facts and their comparisons. Figures can be made to speak out thereby making possible their intelligent use. This module is designed to know about the representation of data in tabular and graphical forms. For analyzing the statistical data, it must be represented in a tabular form and this module does the same, i.e., describe the techniques to convert the data in tabular forms. For the purpose of planning and interpreting the data, visual effects are very useful and necessary. The visual effects in statistics can be obtained by representing the data through graphs.

### **3.2 CLASSIFICATION OF DATA**

According to Tuttle A.M. "A classification is a scheme for breaking a category into a set of parts, called classes, according to some precisely defined differing characteristics possessed by all the elements of the category"

Thus classification impresses upon the arrangement of the data into different classes, which are to be determined depending upon the nature, objectives and scope of the enquiry. This classification is on the basis of sex, age, religion, weight, height, and no. of other factors.

### **3.3 FUNCTIONS OF CLASSIFICATION**

The functions of classification are as follows:

- 1. Summarization:** Classification presents the heavy raw data in a reduced form that is readily comprehensible to the mind and attempts to highlight the significant features contained in the data.
- 2. To make data comparable:** Classification enables us to make meaningful comparisons depending on the basis or criterion of classification. For example, the classification of the students in the university according to sex enables us to make a comparative study of the prevalence of university education among males and females.
- 3. To make relationships among data:** The classification of the given data w.r.t. two or more criteria. say, the sex of the students and the faculty they join in the university will enable us to study the relationship between these two criteria.
- 4. Statistical treatment of the data:** arrangement of the big heterogeneous data into relatively homogeneous groups as per their points of similarities makes it more intelligible, useful and readily willing for further processing like tabulation, analysis, and interpretation of the data heterogeneous data into relatively homogeneous groups or classes according to their points of similarities introduces homogeneity or uniformity for further processing like tabulation, analysis and interpretation of the data.
- 5. Decision Making:** Classification of data informed decision-making by providing structured information about different groups or categories. This is especially valuable in business, healthcare, and other fields where decisions are based on data-driven insights.
- 6. Report Generation:** When presenting statistical results, classification of data simplifies the presentation of complex data by presenting it in a categorized format. This aids in conveying information to different audiences clearly and effectively.

**7. Data Visualization:** in the case of visualization of data, classification facilitates the creation of various types of data visualizations, such as bar charts, pie charts, and histograms. These visualizations help in conveying the distribution and characteristics of data within each category.

### 3.4 BASIS OF CLASSIFICATION

The basis of classification refers to the criteria or attributes used to group and categorize data into distinct classes or categories. The choice of basis for classification depends on the nature of the data and the goals of the analysis. Here are some common bases of classification in statistics:

- **Geographical Basis:** Data can be classified based on geographical regions, such as countries, states, cities, or other specific locations. This is useful when analyzing data that varies across different locations.
- **Chronological Basis:** Data can be classified on the basis of differences in time period. For example, Loss of different years, profit, production, demand and supply, etc. for different periods either by increasing or decreasing time period. This is also called time series data usually used in economics and business.
- **Quantitative Basis:** Data can be classified on the basis of quantitative data which represents the numerical value (i.e., how much, how often, how many). Numerical data gives information about the quantities of a specific thing. Some examples of numerical data are height, length, size, weight, and so on. Quantitative data can be classified into two different types based on the data sets. The two different classifications of numerical data are discrete data and continuous data.
- **Qualitative data:** Data can be classified on the basis of qualitative data, and cannot be expressed in standard units of measurement like rupees, kilograms, liters, etc. This is because they are 'features', 'qualities', or 'characteristics' like eye colors, skin complexion, honesty, good or bad, etc. These are also referred to as attributes. In this case, however, it is possible to count the number of individuals (or items) possessing a particular attribute.

### 3.5 FREQUENCY DISTRIBUTION

**Frequency Distributions** When observations, discrete or continuous, are available on a single characteristic of a large number of individuals, often it becomes necessary to condense the data as far as possible without losing any information of interest. For condensing the data, it is represented using either discrete or continuous frequency distribution tables.

### Terms used in a frequency distribution

**Class Interval:** The whole range of variable values is classified into some groups in the form of intervals. Each interval is called a class interval.

**Class Frequency:** The number of observations in a class is termed the frequency of the class or class frequency.

**Class limits and Class boundaries:** Class limits are the two endpoints of a class interval that are used for the construction of a frequency distribution. The lowest value of the variable that can be included in a class interval is called the lower class limit of that class interval. The highest value of the variable that can be included in a class interval is called the upper-class limit of that class interval.

**3.5.1 Simple Array (Individual observations):** when raw data is arranged in ascending or descending order of magnitude is called arraying of data.

For example, the weekly wages paid to the workers are given below:

300, 240, 150, 160, 145, 120, 320, 140, 130, 175, 143

A simple array in ascending order is :

120, 130, 140, 143, 145, 150, 160, 175, 240, 300, 320

A simple array in descending order is :

320, 300, 240, 175, 160, 150, 145, 143, 140, 130, 120

**3.5.2 Discrete or ungrouped frequency distribution:** In a discrete frequency distribution, values of the variable are arranged individually. The frequencies of the various values are the number of times each value occurs. For example, the weekly wages paid to the workers are given below.

300, 240, 240, 150, 120, 240, 120, 120, 150, 150, 150, 240, 150, 150, 120, 300, 120, 150, 240, 150, 150, 120, 240, 150, 240, 150, 120, 120, 240, 150.

There are various ways to form a frequency distribution for this data. In the first case, let us assume that data is represented in terms of tally marks in a tabular manner as shown below in Table 2.1:

**Table 2.1: Representation of Data using Tally Marks**

Wages Months	Marks	Marks
120	IIII III	8

150	III III II	12
240	III III	8
300	II	2

This data can also be represented without using marks i.e. using frequency only as shown in Table 2.2 which is known as the frequency table.

**Table 2.2: Frequency Table for the Data in Table 2.1**

Weekly Wages (x)	120	150	240	300	<b>Total</b>
No. of Workers (f)	8	12	8	2	<b>30</b>

The frequency table 2.1 is an ungrouped frequency table.

### 3.5.3 Grouped Frequency Distribution:

A grouped frequency distribution is a method used to organize and present quantitative data in a more wide range of datasets. The process involves dividing the data into intervals or groups, each representing a specific range of values. By doing so, the distribution of data becomes clearer, and patterns or trends are easier to identify.

To create a grouped frequency distribution, the first step is to determine the range of the data, which is the difference between the highest and lowest values. Next, the number of intervals is chosen based on the desired level of detail. The interval width is calculated by dividing the range by the number of intervals. A starting point is selected, often slightly below the lowest data value, and intervals are defined accordingly. The frequency of data points within each interval is then counted and recorded in a table.

This table typically includes columns for intervals, frequencies, and sometimes additional columns for cumulative frequencies and midpoints. Cumulative frequencies provide a running total of data points up to a certain interval, while midpoints offer an average value within each interval.

A grouped frequency distribution simplifies the data representation, making it more accessible for analysis, comparison, and interpretation. It condenses the information while retaining the essential characteristics of the data's distribution. When interpreting the table, researchers and analysts can quickly grasp the overall distribution patterns and draw meaningful insights. Careful consideration should be given to the choice of interval width and the number of intervals to ensure that the distribution accurately reflects the data's nature.

We can also draw a grouped frequency table depending on the data we are having. For designing a grouped frequency table, let us consider the following example regarding daily maximum temperatures in a city for 50 days.

28, 28, 31, 29, 35, 33, 28, 31, 34, 29, 25, 27, 29, 33, 30, 31, 32, 26, 26, 21, 21, 20, 22, 24, 28, 30, 34, 33, 35, 29, 23, 21, 20, 19, 19, 18, 19, 17, 20, 19, 18, 18, 19, 27, 17, 18, 20, 21, 18, 19.

Table 2.3 Temperatures in a city for 50 days

<b>Class Interval</b>	<b>Frequency</b>
17-21	17
22-26	9
27-31	13
32-36	11
Total	50

The classes of type 17-21 and 22-26 are inclusive in nature i.e. both the lower bound and upper bound are included in the limit.

Although there are no hard and fast rules that have been laid down for it The following points may be kept in mind for classification:

- (i) The classes should be clearly defined and should not lead to ambiguity.
- (ii) The classes should be exhaustive, i.e., each of the given values should be included in one of the classes.
- (iii) The classes should be mutually exclusive and non-overlapping.
- (iv) The classes should be of equal width. The principle, however, cannot be rigidly followed.
- (v) Indeterminate classes, e.g., the open-end classes such as less than 'a' or greater than 'b' should be avoided as far as possible since they create difficulty in analysis and interpretation.
- (vi) The number of classes should neither be too large nor too small. It should preferably lie between 5 and 15. However, the number of classes may be more than 15 depending upon the total frequency and the details required. But it is 15 desirable that it is not less than 5 since in that case, the classification may not reveal the essential characteristics of the population.

In Table 2.3, the class intervals are 17-21, 22-26, 27-31 and 32-36. Here, say for the class 17-21, the lower-class limit is 17 and the upper-class limit is 21. Both 17 and 21 are part of this class. This is called an inclusive class. Another type of class is an exclusive class as shown below in

### **5.3.4 Continous Frequency Distribution**

A continuous frequency distribution is used to present and analyze quantitative data that takes on a wide range of possible values within a continuous range. Unlike discrete data, which comprises distinct, separate values, continuous data includes values that can take any real number within a given interval. The concept of a continuous frequency distribution is particularly relevant when dealing with measurements such as height, weight, temperature, or time.

Creating a continuous frequency distribution involves dividing the entire range of data into intervals, often referred to as "class intervals". These intervals are constructed in a way that ensures they are non-overlapping and collectively cover the entire range of data. The frequency of data points falling within each interval is then counted, and this information is typically represented in a table or histogram.

Continuous frequency distributions are essential tools in data analysis, enabling us to explore and summarize large datasets while preserving the integrity of the data's continuous nature. The accuracy of the distribution heavily relies on the choice of class intervals and the visualization method used to represent the data, making thoughtful consideration of these factors crucial for meaningful interpretation.

## **Forms of Continuous Frequency Distribution**

### **1. Inclusive Class Interval**

The inclusive type of data has the class interval 20-29, 30-39, 40-49, 50-59, in which the upper limit and the lower limit are included in the class. The fractional values between 29 to 30 cannot be accounted for in such a classification. Therefore inclusive type classification is used in grouped frequency distribution for discrete values like no. of students in the class, no. of road accidents, etc., here the value takes an only integer value.

### **2. Exclusive Class Intervals**

let us consider the temperature in exclusive form.

<b>Temperature</b>	<b>Frequency</b>
17 but less than 21	17
21 but less than 25	7
25 but less than 29	10
29 but less than 33	9
33 but less than 37	7

Total	50
-------	----

In Table 2.4 upper values are excluded from the class i.e., in the class 17-21 only values from 17 to 20 are taken and the values of 21 are considered in the next class. Such a type of distribution is known as an exclusive class.

### 3. Open-ended Frequency distribution

It may be the case that some values in the data set are extremely small compared to the other values of the data set and similarly some values are extremely large in comparison. Then what we do is we do not use the lower limit of the first class and the upper limit of the last class. Such classes are called open end classes. A distribution of open ended frequencies with at least one end open is known as an open-end distribution. The first class's lower limit, the last class's upper limit, or both are not specified. "Below" or "less than" and "above" or "greater than" are used.

**Table 2.5: Open-end Class Grouped Frequency Table**

Marks	No. of students
Less than 20	5
20-40	14
40-60	27
60-80	30
80 & more	35

**Size of the Class:** The length of the class is called the class width. It is also known as class size.

size of the class = Upper Limit-Lower Limit

**Mid-point of the Class:** The midpoint of a class interval is called the Mid-point of the Class. It is the representative value of the entire class.

Mid-point of the class =  $(\text{Upper Limit} + \text{Lower Limit}) / 2$

**Continuous Frequency Distribution:** If we deal with a continuous variable, it is not possible to arrange the data in the class intervals of the above type.

### 4. Unequal Class Frequency Distribution

The classes of a frequency distribution may or may not be of equal width. A frequency distribution with unequal class width is reproduced in the Table below. Here, the width of the 1st, 2nd, and 5th classes is 10, while that of the 3rd is 20 and that of the 4th is 25.

Table of Unequal Class Interval



Class Interval	Frequency
0-10	6
10-20	10
20-40	14
40-65	25
65-75	30

### 5. Cumulative Frequency Distribution

A cumulative frequency distribution is a statistical representation of quantitative data that shows the total number of observations that fall below or within a certain value or interval. It provides a way to understand the distribution of data in terms of cumulative frequencies, allowing for insights into the overall spread and concentration of the data. This cumulative frequency distribution are of two types:

- less than type cumulative frequency
- more than type cumulative frequency

a) **Less Than Type Cumulative Frequency:** less than type cumulative frequency for any value of the variable is obtained by adding successively the frequencies of all the previous values, including the frequency of the variable against which the totals are written, provided the values are arranged in ascending order.

For example

Marks Class Interval	Frequency	Cumulative Frequency
Less than 50	15	15
Less than 100	18	$15+18 = 33$
Less than 150	40	$33+40 = 73$
Less than 200	45	$73 +45 = 118$
Less than 250	30	$118+30 = 148$
Less than 300	25	$148+25 = 173$
Less than 350	20	$173+20 = 193$

**b) More Than Type Cumulative Frequency:** this is obtained similarly by finding the cumulative totals of frequencies starting from the highest value of the variable to the lowest values. For example:

Marks Class Interval	Frequency	Cumulative Frequency
More than 50	5	56
More than 100	10	51
More than 150	11	41
More than 200	13	31
More than 250	11	18
More than 300	5	7
More than 350	2	2

### 3.6 TABULATION OF DATA: MEANING

Tabulation of data means the systematic presentation of the information contained in the data i.e., in the form of rows as well as columns as per the required objective or features. In a table, a row denotes a horizontal arrangement of data, whereas a column denotes a vertical arrangement. A table's rows and columns are indicated by the proper stubs and captions (or headers or subheadings), respectively, to describe the type of information provided. Data should be presented logically, simply, and unambiguously in tabular form.

Professor Bowley in his manual of Statistics refers to tabulation as "the intermediate process between the accumulation of data in whatever form they are obtained, and the final reasoned account of the result shown by the statistics".

Tabulation is a midway process between the collection of the data and statistical analysis. Rather, tabulation is the final stage in the collection and compilation of the data and forms the gateway for further statistical analysis and interpretations. Tabulation makes the data understandable and facilitates comparisons, and the work of further statistical analysis, averaging, correlation, etc different tools of statistics. It makes the data suitable for further representation in the form of diagrams as well as graphics.

There are no rigid rules for tabulating the statistical data. To construct an excellent table, one has

to have a clear understanding of the information to be presented and the points that should be emphasized and familiarity with the table process of creation. To ensure that the relationship between the data of one or more series, as well as the significance of all the figures given in the classification adopted, the organization of data tabulation requires careful consideration. data represents in the table shows comparisons and contrasts. only the ability, knowledge, experience, and common sense of the tabulator, while keeping in mind the nature, scope, and aims of the inquiry can produce a good table.

### **3.8 Parts of a Table**

Based on the type of data and the goal of the study, the various components of a table vary from problem to problem. But the following elements must be present in a decent statistical table:

- Table number
- Title
- Head notes or Prefatory notes
- Captions and Stubs
- Body of the table
- Foot-note
- Source note

**1) Table number:** It is required for the identification of a table mainly when there is more than one table in a particular analysis. The table number is always mentioned in the center at the top or left side depending upon the researcher's choice.

**2) Title of the table:** It indicates the type of information contained in the body of the table. The title of the table provides a brief description of the contents of the table. It exactly describes the nature of the data, place (region or variable), time period, and source of the data. It should be brief and clearly complete to describe the nature of the table.

**3) Head notes:** It is also called prefatory notes are written just below the title. It shows contents and units of measurement like (rupees lakh) or (lakh quintals) or (thousand rupees). It should be written in brackets and should appear on the right side top just below the title. However, every table does not need a head note, like the number of students in each class.

**4) Stubs and Captions:** Stubs are used to designate rows. They appear on the left-hand column of the table. Stubs consist of two parts: a) Stub head describes the nature of stub entry. b) Stub entry is the description of row entries. while captions are called box heads, designate the data presented in the columns of the table.

It may contain more than one column head, and each column head may be subdivided into more than one sub-head. For example, we can divide the students of a college into Section A and Section B and then again into males and females.

**5) Main body of the table:** It is also called the field of the table, and is its most important and immense part. It contains the relevant numerical information which is already contained in the title of the table. For creating as useful it contains row and column totals separately and then a grand total.

**6) Foot Note:** footnotes used in order to take further elaboration or some additional information. is a qualifying statement put just below the table (at the bottom). Its purpose is to caution about the limitations of the data or certain omissions. It will be used the symbol \*, \*\*, etc.

**7) Source of data:** It may be the last part of a table, yet it is important. It speaks about the authenticity of the data taken into the table. It should be below the footnotes. It is generally required for the secondary data collection, here it explains from where the data has been taken. It contains table no. volume, issue no. page number.

**Format of a Blank Table**  
**Title**  
**(Head Note or Prefatory Note)**

<b>Sub Heading</b> ↓	<b>Caption</b>				<b>Total</b>
	<b>Sub-Heading 1</b>		<b>Sub-Heading 2</b>		
	C1	C2	C1	C2	
R1	<b>Body of Table</b>				Total R1
R2					Total R2
R3					Total R3
<b>Total</b>	<b>Total C1</b>	<b>Total C2</b>	<b>Total C1</b>	<b>Total C2</b>	<b>Grand Total</b>

Foot Note:

Source:

**3.9 Questions for Practice**

Q1. What do you mean by the classification of data?

Q2. What are the functions of classification?

Q3. Explain the basis of classification.

Q4. Explain frequency distribution.

Q5. Explain discrete or ungrouped frequency distribution.

Q6. Discuss continuous or grouped frequency distribution with an example.

Q7. What are the various forms of frequency distributions?

Q8. What is a tabulation of data?

Q9. What are the parts of a table? Explain with example.

### **3.10 Suggested Readings**

- A. Abebe, J. Daniels, J.W. Mckean, “Statistics and Data Analysis”.
- Clarke, G.M. & Cooke, D., “A Basic course in Statistics”, Arnold.
- David M. Lane, “Introduction to Statistics”.
- S.C. Gupta and V.K. Kapoor, “Fundamentals of Mathematical Statistics”, SultanChand & Sons, New Delhi.

**COURSE NAME: STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY**

**SARM 1: INTRODUCTION TO STATISTICS**

**SEMESTER I**

**UNIT 4: DIAGRAMMATIC AND GRAPHICAL PRESENTATION OF DATA**

**STRUCTURE**

**4.0 Learning Objectives**

**4.1 Introduction**

**4.2 Data Processing**

**4.3 Diagrammatic presentation of data processing**

**4.4 Graphical representation of data processing using Excel**

**4.5 Types of Charts**

**4.5.1 Pie Charts**

**4.5.2 Line and Area Charts**

**4.5.3 Column Chart**

**4.5.4 Bar Chart Variations**

**4.6 Apply Chart Layout**

**4.7 Add Labels**

**4.8 Change the Style of a Chart**

**4.9 Data Preserving**

**4.10 Data Preserving vs. Storing Data**

**4.11 Data Preservation vs. Retention of Data**

**4.12 Questions for Practice**

**4.13 Suggested Readings**

**4.0 LEARNING OBJECTIVES**

After studying the Unit, students will be able to:

- Processing the data after collection
- Passes through various stages of Processing
- Plan the data analysis
- Classify the data
- Tabulate the data
- Analyse the data
- Data Preservation, storage and Retention

#### **4.1 INTRODUCTION**

After data collection, the researcher turns his focus of attention to the processing and preserving of the data. "Diagrammatic" and "Graphical" presentation of data both refer to methods of visually representing information to enhance understanding and analysis. While the terms are often used interchangeably, they can have slightly different connotations.

A diagrammatic presentation involves using diagrams or visual aids to represent data. Diagrams are usually simplified and symbolic representations that convey information clearly and straightforwardly. Common types of diagrammatic presentations include bar charts, pie charts, line graphs, histograms, scatter plots, and pictograms. These visual representations help illustrate patterns, trends, comparisons, and relationships within the data.

Graphical presentation refers to the use of graphs, charts, and visual elements to display data in a way that makes it easier to interpret and analyze. Graphs can include various types of charts, plots, and diagrams that present data in a visual format. This type of presentation is particularly useful when dealing with complex data sets or when you want to emphasize relationships and trends. Graphical presentations can include more detailed and sophisticated visualizations, such as 3D graphs, heat maps, area charts, and more.

#### **4.2 DATA PROCESSING**

Data processing refers to certain operations such as editing, coding, computing of the scores, preparation of master charts, etc. A researcher has to make a plan for every stage of the research process. As such, a good researcher makes a perfect plan for processing and analysis of data. To some researchers' data processing and analysis is not a very serious activity.

Data processing occurs when data is collected and translated into usable information. Usually

performed by a data scientist or team of data scientists, it is important for data processing to be done correctly so as not to negatively affect the end product or data output.

Data processing starts with data in its raw form and converts it into a more readable format (graphs, documents, etc.), giving it the form and context necessary to be interpreted by computers and utilized by employees throughout an organization.

### **Stages of Data Processing**

- 1) **Data Collection:** Collecting data is the first step in data processing. Data is pulled from available sources. The data sources available must be trustworthy and well-built so the data collected (and later used as information) is of the highest possible quality.
- 2) **Data Preparation:** Once the data is collected, it then enters the data preparation stage. Data preparation, often referred to as “pre-processing” is the stage at which raw data is cleaned up and organized for the following stage of data processing. During preparation, raw data is diligently checked for any errors. The purpose of this step is to eliminate bad redundant, incomplete, or incorrect data and begin to create high-quality data for the best business intelligence.
- 3) **Data Input:** The data is then entered into its destination and translated into a language that it can understand. Data input is the first stage in which raw data begins to take the form of usable information.
- 4) **Processing:** During this stage, the data inputted to the computer in the previous stage is processed for interpretation. Processing is done using machine learning algorithms, though the process itself may vary slightly depending on the source of data being processed.
- 5) **Data output/interpretation:** The output/interpretation stage is the stage at which data is finally usable to non-data scientists. It is translated, readable, and often in the form of graphs, videos, images, plain text, etc.).
- 6) **Data storage:** The final stage of data processing is storage. After all of the data is processed, it is then stored for future use. While some information may be put to use immediately, much of it will serve a purpose later on. When data is properly stored, it can be quickly and easily accessed by members of the organization when needed.

### **4.3 DIAGRAMMATIC PRESENTATION OF DATA PROCESSING**

As you know, diagrammatic presentation is one of the techniques of visual presentation of data. It



is a fact that diagrams do not add new meaning to the statistical facts but they reveal the facts of the data more quickly and clearly. Because examining the figures from tables become laborious and uninteresting to the eye and also confusing. Here, it is appropriate to state the words of M. J. Moroney, “cold figures are uninspiring to most people. Diagrams help us to see the pattern and shape of any complex situation.” Thus, the data presented through diagrams are the best way of appealing to the mind visually. Hence, diagrams are widely used in practice to display the structure of the data in research work.

#### ➤ **Rules for Preparing Diagrams**

The prime objective of the diagrammatic presentation of data is to highlight their basic hidden facts and relationships. To ensure that the presentation of numerical data is more attractive and effective, therefore, it is essential to keep the following general rules in mind while adapting diagrams in research work. Now, let us discuss them one by one.

1. You must have noted that the diagrams must be geometrically accurate. Therefore, they should be drawn on the graphic axis i.e., the ‘X’ axis (horizontal line) and the ‘Y’ axis (vertical line). However, the diagrams are generally drawn on plain paper after considering the scale.
2. While taking the scale on the ‘X’ axis and ‘Y’ axis, you must ensure that the scale showing the values should be in multiples of 2, 5, 10, 20, 50, etc.
3. The scale should be set up, e.g., millions of tons, persons in Lakhs, value in thousands, etc. On the ‘Y’ axis the scale starts from zero, as the vertical scale is not broken.
4. Every diagram must have a concise and self-explanatory title, which may be written at the top or bottom of the diagram.
5. To draw the readers' attention, diagrams must be attractive and well-proportioned.
6. Different colors or shades should be used to exhibit various components of diagrams and also an index must be provided for identification.
7. It is essential to choose a suitable type of diagram. The selection will depend upon the number of variables, minimum and maximum values, and objects of presentation.

#### **4.4 GRAPHICAL REPRESENTATION OF DATA PROCESSING USING EXCEL**

Excel charts are graphical representations of numeric data. Graphs make it easier for users to compare and understand numbers, so charts have become a popular way to present numerical data. Every chart tells a story. Stories can be simple: “See how our sales have increased” or complex: “This is how our overhead costs relate to the price of our product.” Whether simple or complex,

the story should be readily understandable. If you can't immediately understand what a chart means, then it isn't a good chart.

Graphs are constructed with data points, which are the individual number in a worksheet, and data series, which are the groups of related data points within a column or row. Charts and graphs in Microsoft Excel provide a method to visualize numeric data. While both graphs and charts display sets of data points about one another, charts tend to be more complex, varied, and dynamic. People often use charts and graphs in presentations to give management, client, or team members a quick snapshot of progress or results. You can create a chart or graph to represent nearly any kind of quantitative data — doing so will save you the time and frustration of poring through spreadsheets to find relationships and trends. It's easy to create charts and graphs in Excel, especially since you can also store your data directly in an Excel Workbook, rather than importing data from another program. Excel also has a variety of preset chart and graph types so you can select one that best represents the data relationship(s) you want to highlight. Excel comes with a wide variety of charts capable of graphically representing most standard types of data analysis and even some more exotic numeric interpolations. The type of data you are using and presenting determines the type of chart you will plot the data on.

#### **4.5 TYPES OF CHARTS**

**1. Pie Charts:** These work best for displaying how much each part contributes to a total value. Pie charts can be exploded for greater visual clarity, or turned into doughnut charts, which can represent more than just one set of data.

- Use pie charts to compare percentages of a whole (“whole” is the total of the values in your data). Each value is represented as a piece of the pie so you can identify the proportions. There are five pie chart types: pie, pie of pie (this breaks out one piece of the pie into another pie to show its sub-category proportions), bar of pie, 3-D pie, and doughnut.

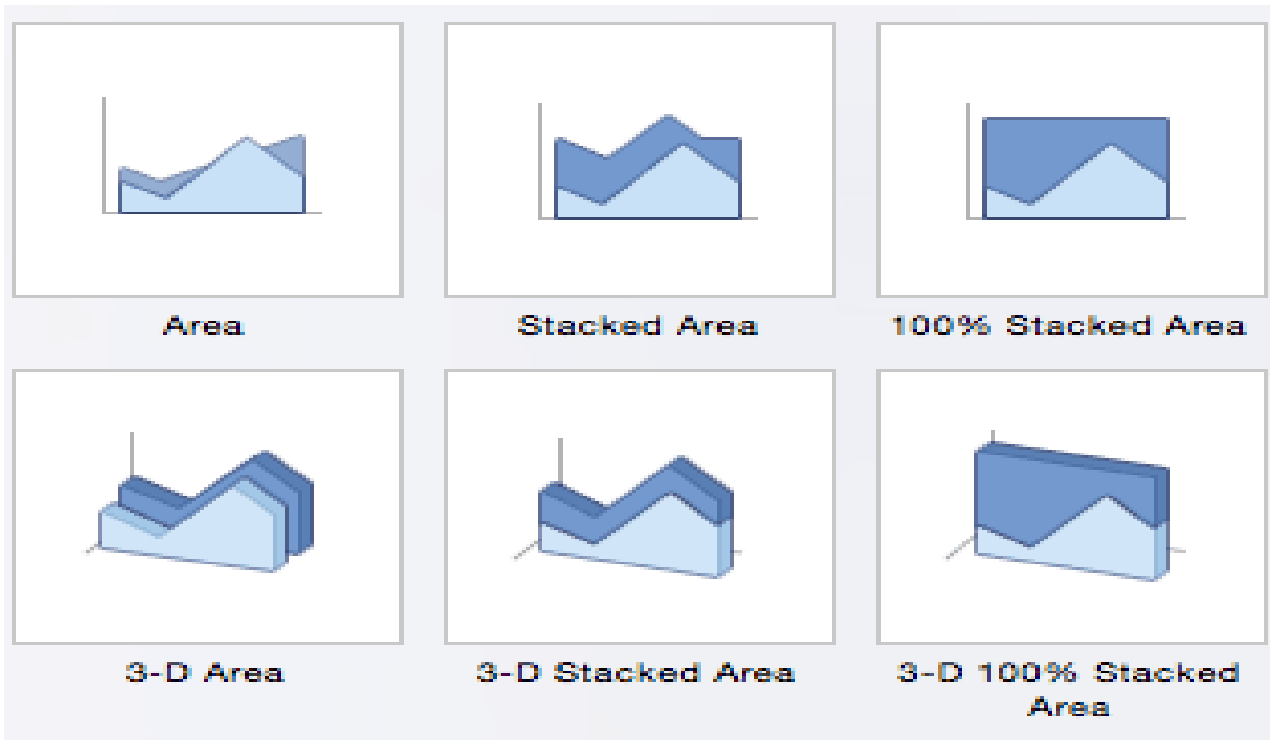


**2. Line and area charts:** These show data points connected with lines, indicating upward or downward trends in value. Area charts show the area below a line filled in. Both types can be combined with column charts to show more data.

- A line chart is most useful for showing trends over time, rather than static data points. The lines connect each data point so that you can see how the value(s) increased or decreased over some time. The seven-line chart options are line, stacked line, 100% stacked line, line with markers, stacked line with markers, 100% stacked line with markers, and 3-D line.

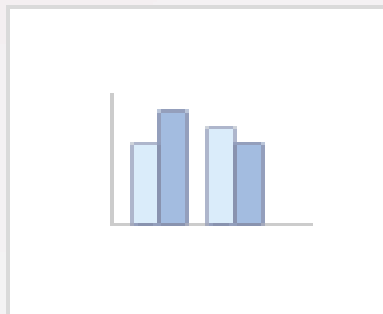


- Area:** Like line charts, area charts show changes in values over time. However, because the area beneath each line is solid, area charts are useful to call attention to the differences in change among multiple variables. There are six area charts: area, stacked area, 100% stacked area, 3-D area, 3-D stacked area, and 3-D 100% stacked area.

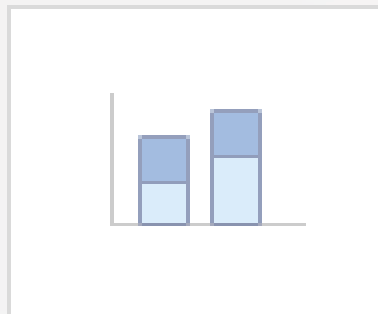


1. **Column and bar charts:** These compare values across categories, with results presented vertically in column charts and horizontally in bar charts, The composition of the column or bar can be stacked in more than one color to represent the contribution of each portion of a category's data to the total for that category.

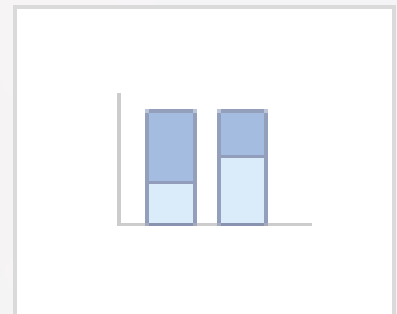
- **Column Charts:** Some of the most commonly used charts, column charts, are best used to compare information or if you have multiple categories of one variable (for example, multiple products or genres). Excel offers seven different column chart types: clustered, stacked, 100% stacked, 3-D clustered, 3-D stacked, 3-D 100% stacked, and 3-D, pictured below. Pick the visualization that will best tell your data's story.



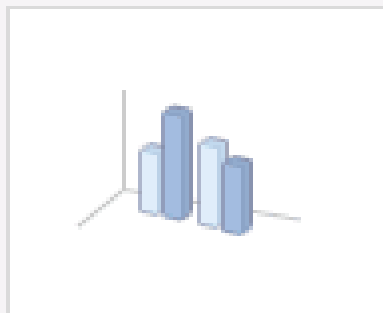
Clustered Column



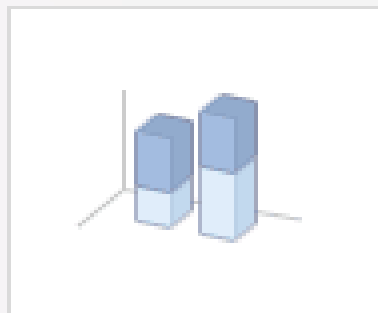
Stacked Column



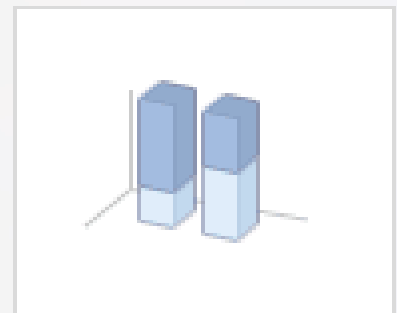
100% Stacked Column



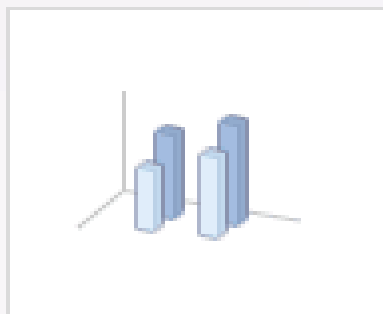
3-D Clustered Column



3-D Stacked Column

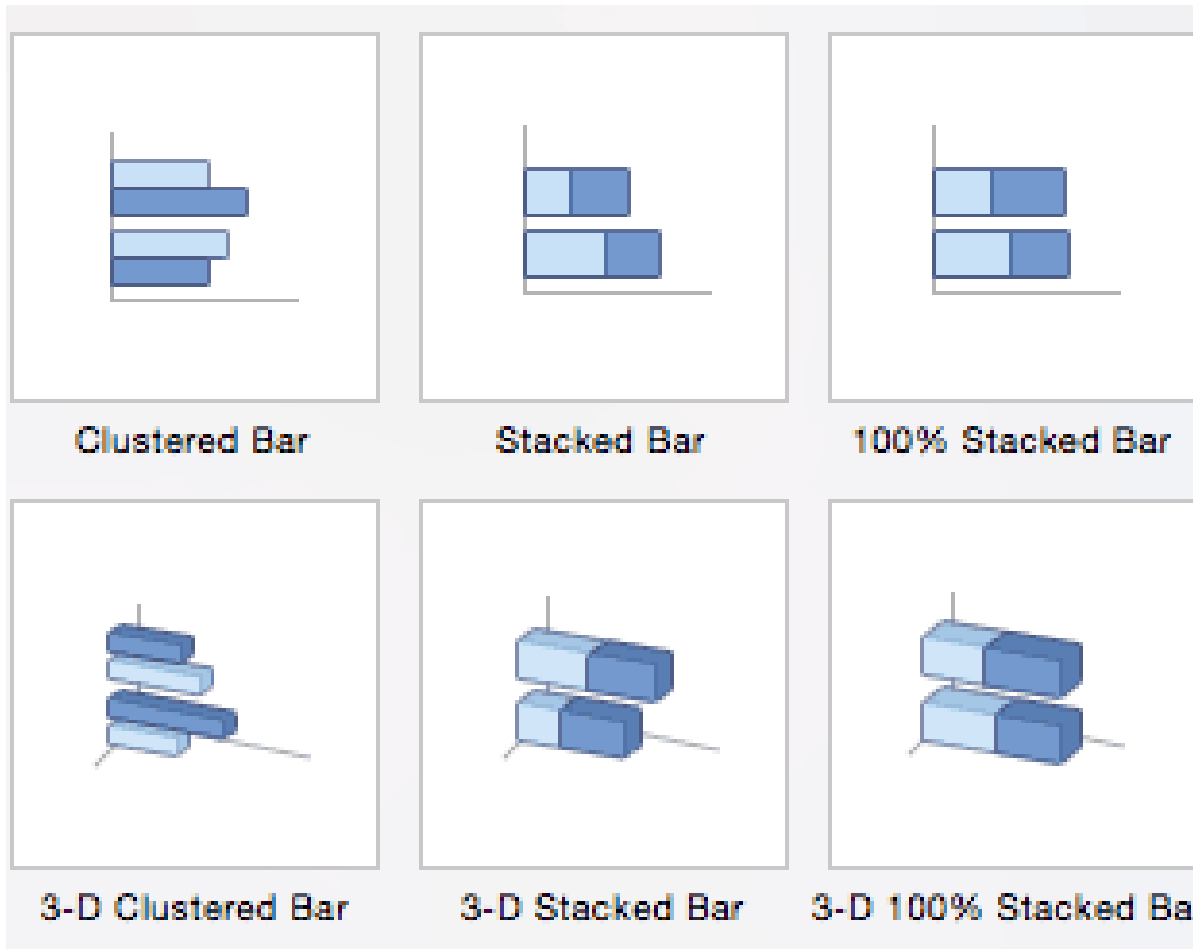


3-D 100% Stacked Column



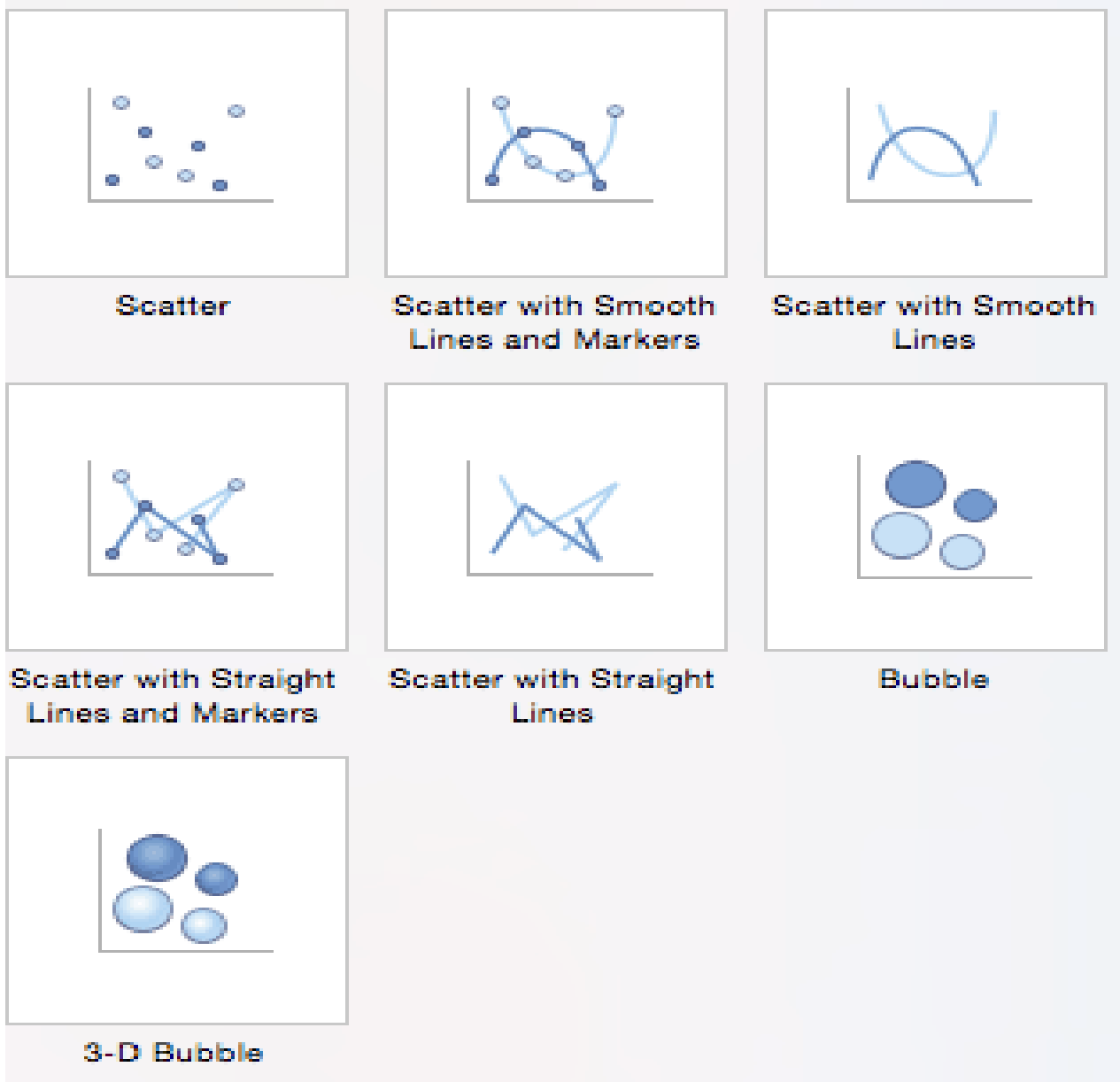
3-D Column

- **Bar Charts:** The main difference between bar charts and column charts is that the bars are horizontal instead of vertical. You can often use bar charts interchangeably with column charts, although some prefer column charts when working with negative values because it is easier to visualize negatives vertically, on a y-axis.

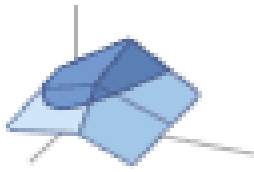


2. **Special charts:** Excel includes several charts suitable for presenting scientific statistical, and financial data. Scatter charts are used to present experimental results. Surface and cone charts are good for presenting 3-D and 2-D changes in data. Radar charts show data values about a single metric. Stock charts present values for between three and five series of data, including open, high, low, close, and volume trading information.
- **Scatter Charts:** Similar to line graphs, because they are useful for showing change in variables over time, scatter charts are used specifically to show how one variable affects another. (This is called correlation.) Note that bubble charts, a popular chart type, are categorized under scatter. There are seven scatter chart options: scatter, scatter with smooth lines and markers,

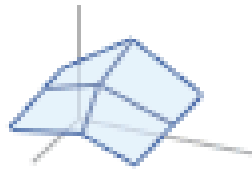
scatter with smooth lines, scatter with straight lines and markers, scatter with straight lines, bubble, and 3-D bubble.



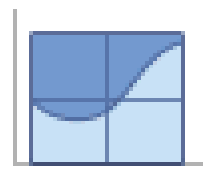
- **Surface:** Use a surface chart to represent data across a 3-D landscape. This additional plane makes them ideal for large data sets, those with more than two variables, or those with categories within a single variable. However, surface charts can be difficult to read, so make sure your audience is familiar with them. You can choose from 3-D surface, wireframe 3-D surface, contour, and wireframe contour.



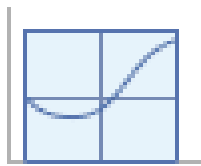
3-D Surface



Wireframe 3-D Surface

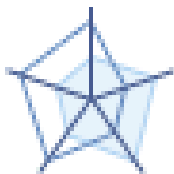


Contour



Wireframe Contour

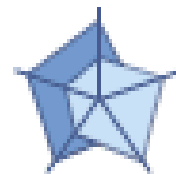
- **Radar:** When you want to display data from multiple variables about each other use a radar chart. All variables begin from the central point. The key with radar charts is that you are comparing all individual variables about each other — they are often used for comparing the strengths and weaknesses of different products or employees. There are three radar chart types: radar, radar with markers, and filled radar.



Radar



Radar with Markers

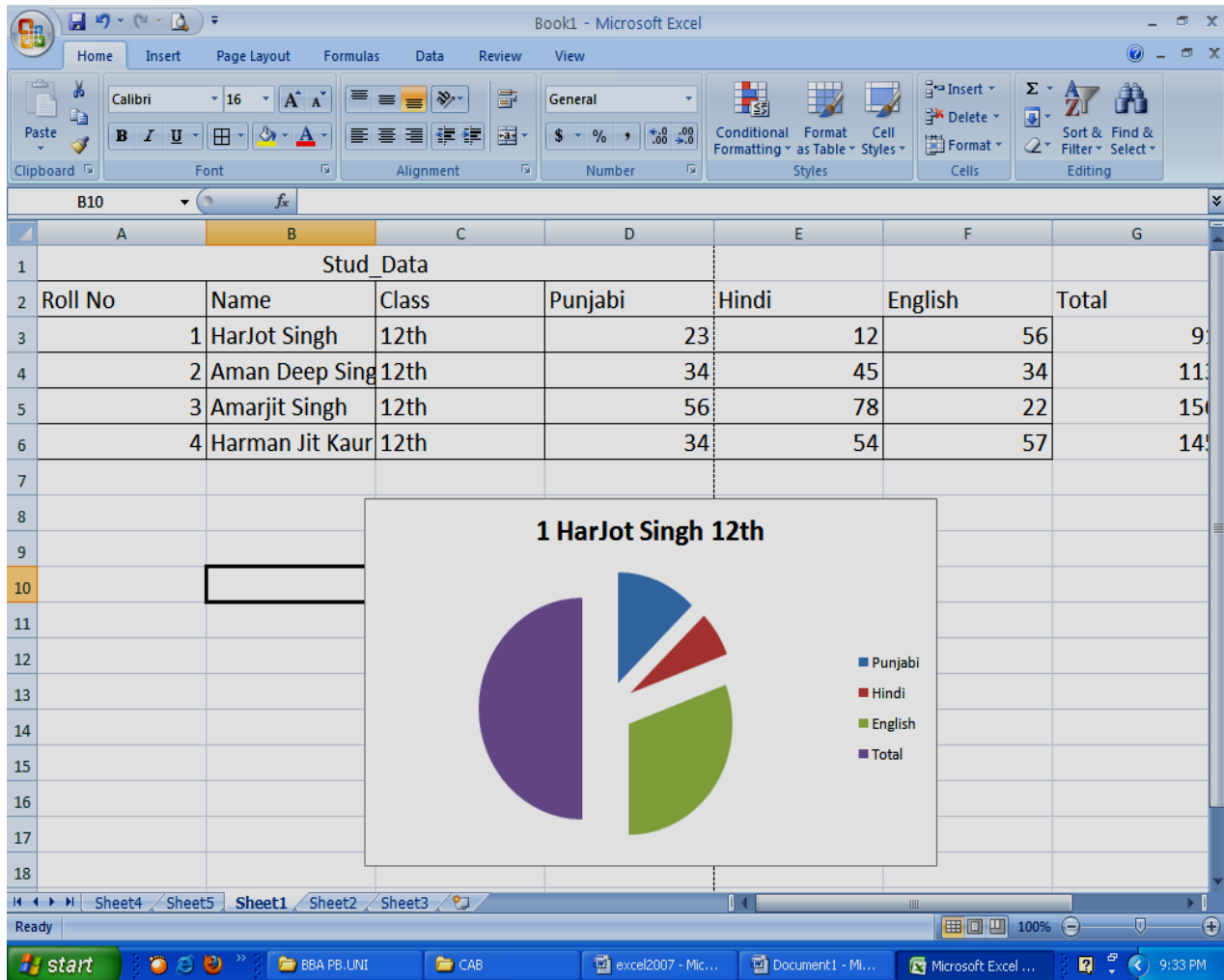


Filled Radar



## 1. Pie Charts

Use pie charts to show the relationships between pieces of an entity. The implication is that the pie includes all or something. The pie chart isn't appropriate for illustrating some of anything, so if there's not an obvious "all" in the data you're charting, don't use a pie.



**Fig.**

A pie chart can only include one data series. If you select more than one data series, Excel uses the first series and ignores all others. No error message appears, so you won't necessarily know that the chart doesn't show the data you intended to include unless you examine the chart carefully. When you create a pie chart, Excel totals the data points in the series and then divides the value of each data point into the series total to determine how large each data point's pie slice should be. Don't include a total from the worksheet as a data point; this doubles the total Excel calculates, resulting in a pie chart with one large slice (50 percent of the pie).

## 2. Line and Area Charts

The series chart shown in the figure is a line chart. In a 2-D version (as shown) or in a 3-D version that is sometimes called a ribbon chart. An area chart is a line chart with the area below the line filled. Line charts and area charts are typically used to show one or more variables (such as sales, income, or price) changing over time.

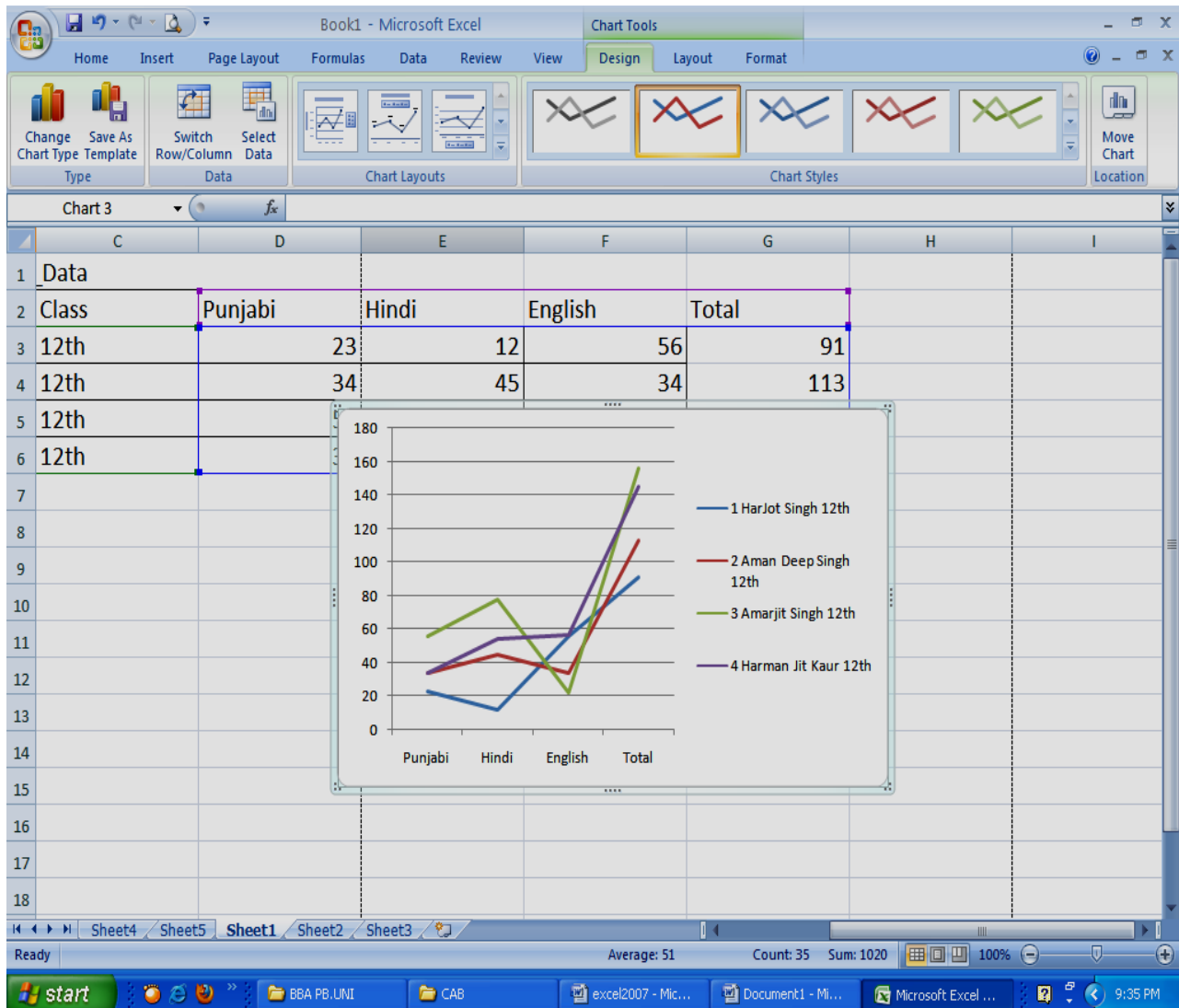


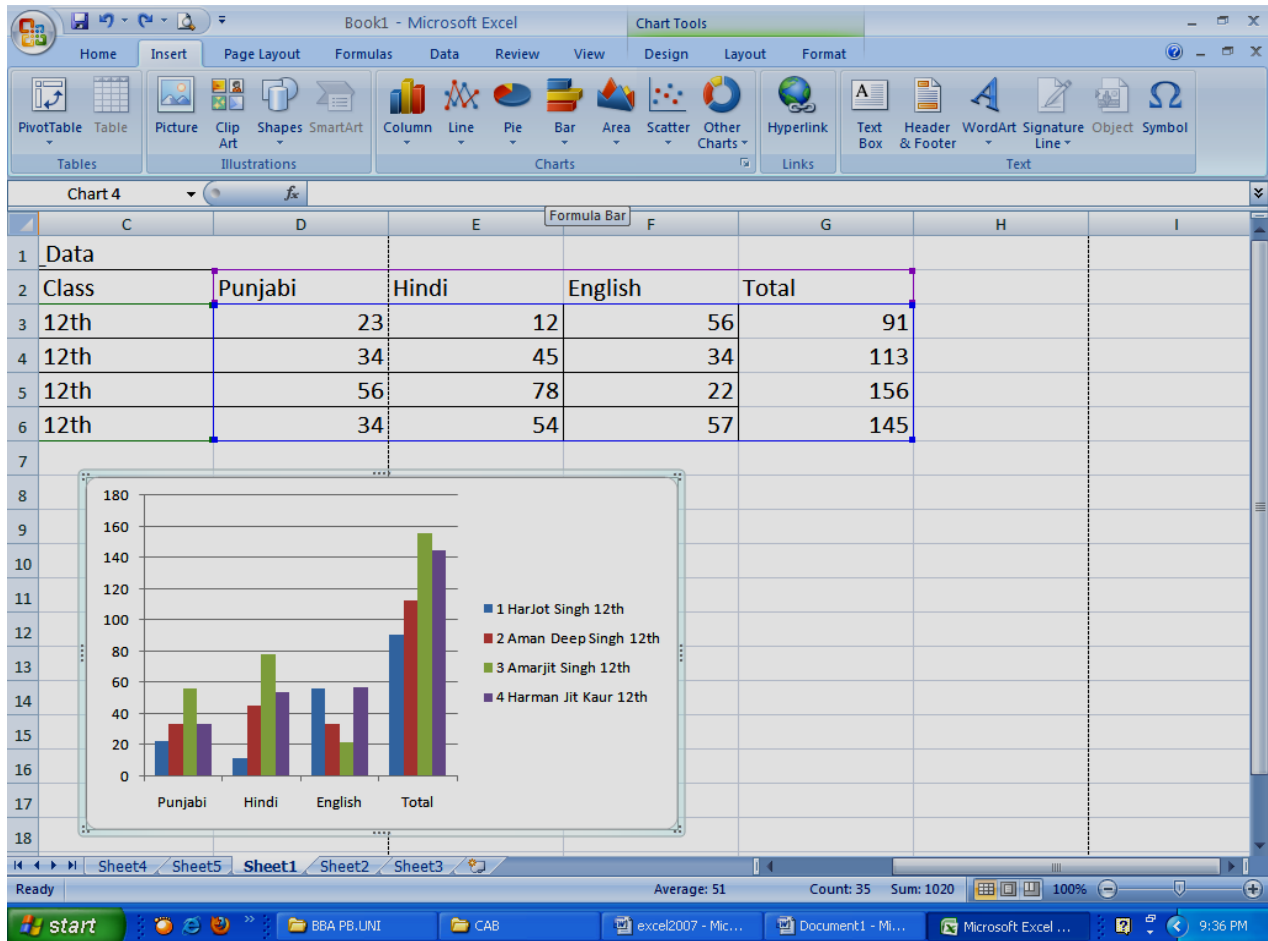
Fig.

## 3. Column Chart

The figure shows the same information presented as a bar chart. The bars give added substance to the chart. In the line chart, what the reader notices is the trend up or down in each line and the gaps between the lines.

Line and area charts share a common layout. The horizontal line is called the X-axis, and the vertical line is the Y-axis (the same x- and y-axis you may have learned about in algebra or geometry class when plotting data points). In a bar chart, however, the axis is turned 90 degrees so that the x-axis is on the left side.

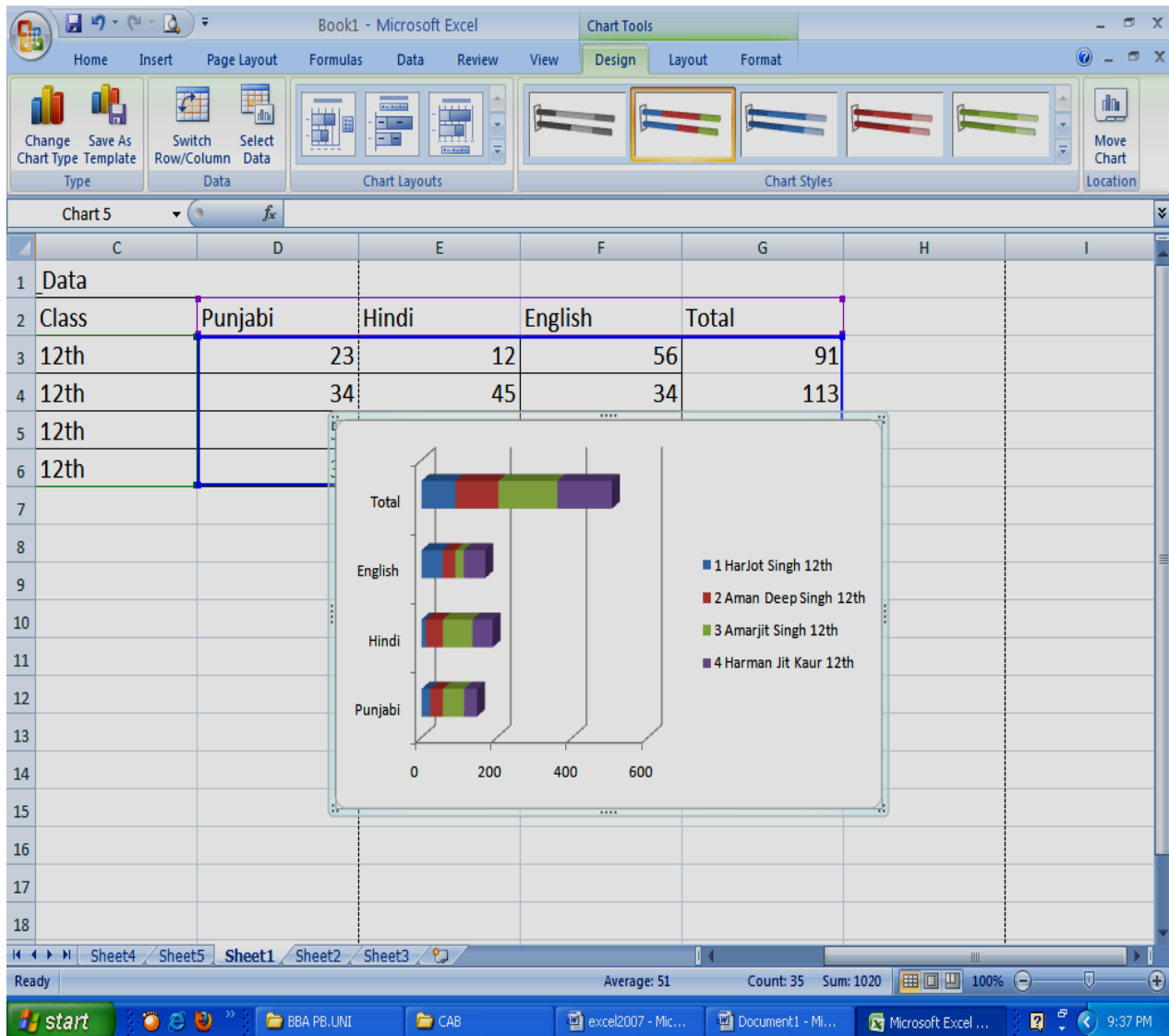
Excel can also combine columns with line or area charts and embellish line or column charts with 3-D effects. You can make the columns and lines on your charts into tubes, pyramids, cones, or cylinders; or transform regular bars into floating 3-D bars. Plotting data on two axes is also possible with column charts.



**Fig.**

#### **4. Bar Chart Variations**

Column charts are the same as bar charts but with the X-axis at the bottom. There are three-dimensional varieties of bar and column charts, which add depth to the regular chart. Cylinders, cones, and pyramids are variations of a column chart.



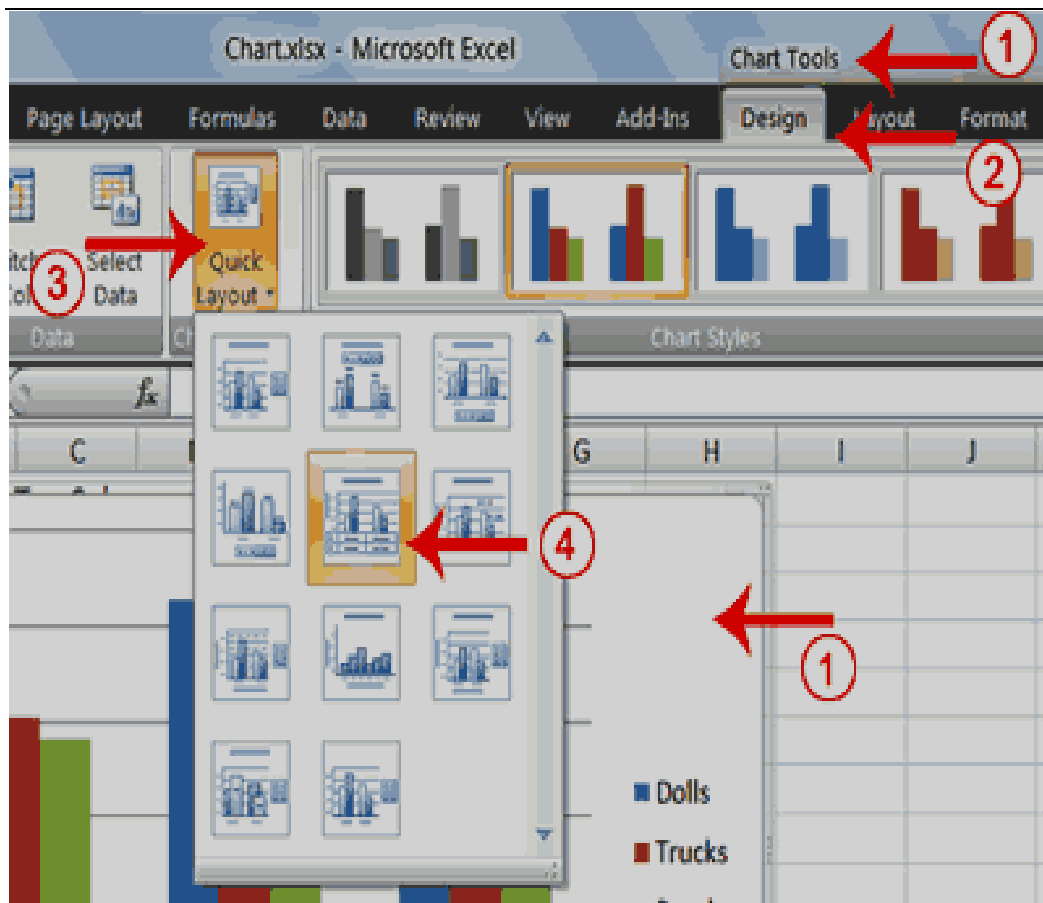
**Fig.**

Excel also offers another style of bar and column chart—the stacked chart. A stacked 3-D column chart, using the same data as Figure. In a stacked chart, parallel data points in each data series are stacked on top or to the right of each other. Stacking adds another dimension to the chart since it allows the user to compare sales between as well as within time periods-like providing a column chart and a pie chart for each period.

The 3-D charts have three axes. In a 3-D column chart, the X-axis is on the bottom. The vertical axis is the Z-axis; the Y-axis goes from front to back, providing the “third dimension” of depth in the chart. Don’t worry about memorizing which axis is which in each chart type; there are ways to know which is which when you’re creating or editing the chart.

## 4.6 APPLY CHART LAYOUT

Context tabs are tabs that only appear when you need them. Called Chart Tools, there are three chart context tabs: Design, Layout, and Format. The tabs become available when you create a new chart or when you click on a chart. You can use these tabs to customize your chart. You can determine what your chart displays by choosing a layout. For example, the layout you choose determines whether your chart displays a title, where the title displays, whether your chart has a legend, where the legend displays, whether the chart has axis labels, and so on. Excel provides several layouts from which you can choose.



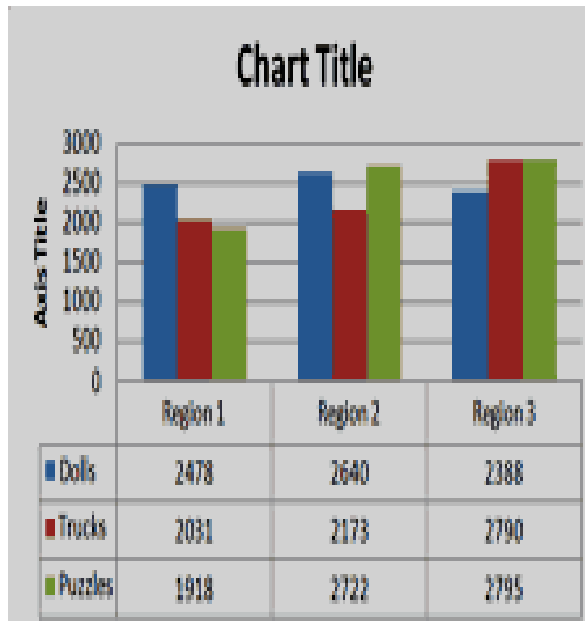
**Fig**

### **Steps to Apply a Chart Layout**

1. Click your chart. The Chart Tools become available.
2. Choose the Design tab.
3. Click the Quick Layout button in the Chart Layout group. A list of chart layouts appears.
4. Click Layout. Excel applies the layout to your chart.

## 4.7 ADD LABELS

When you apply a layout, Excel may create areas where you can insert labels. You use labels to give your chart a title or to label your axes. When you applied layout, Excel created label areas for a title and the vertical axis



Before

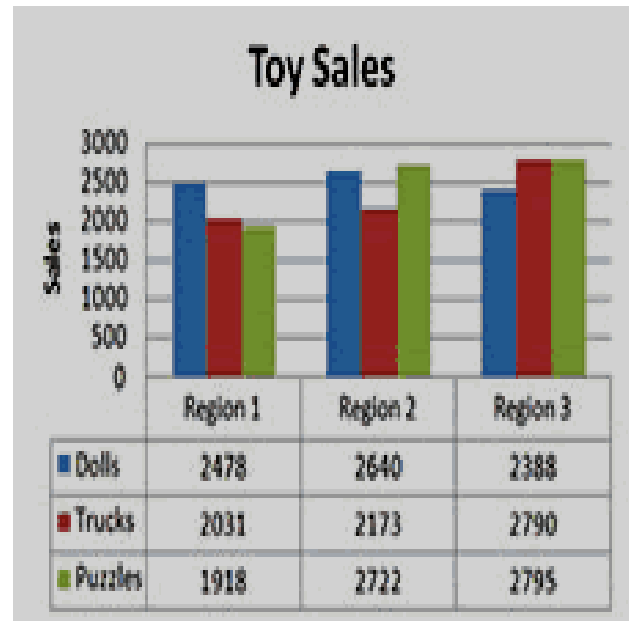


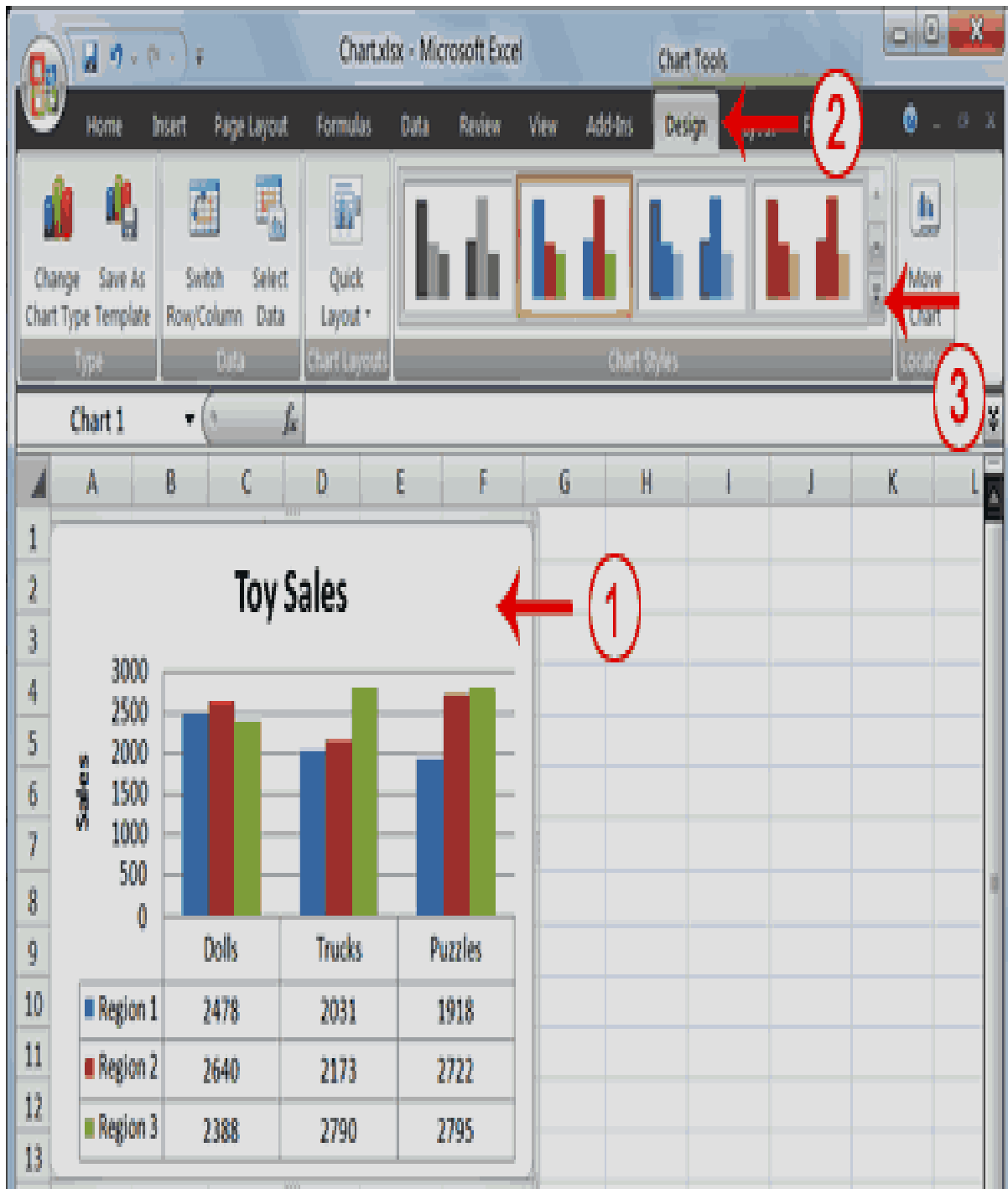
Fig After

### Steps to add labels

1. Select Chart Title. Click on Chart Title and then place your cursor before the C in Chart and hold down the Shift key while you use the right arrow key to highlight the words Chart Title.
2. Type **Toy Sales**. Excel adds your title.
3. Select Axis Title. Click on Axis Title. Place your cursor before the A in Axis. Hold down the Shift key while you use the right arrow key to highlight the words, Axis Title.
4. Type Sales. Excel labels the axis.
5. Click anywhere on the chart to end your entry.

## 4.8 CHANGE THE STYLE OF A CHART

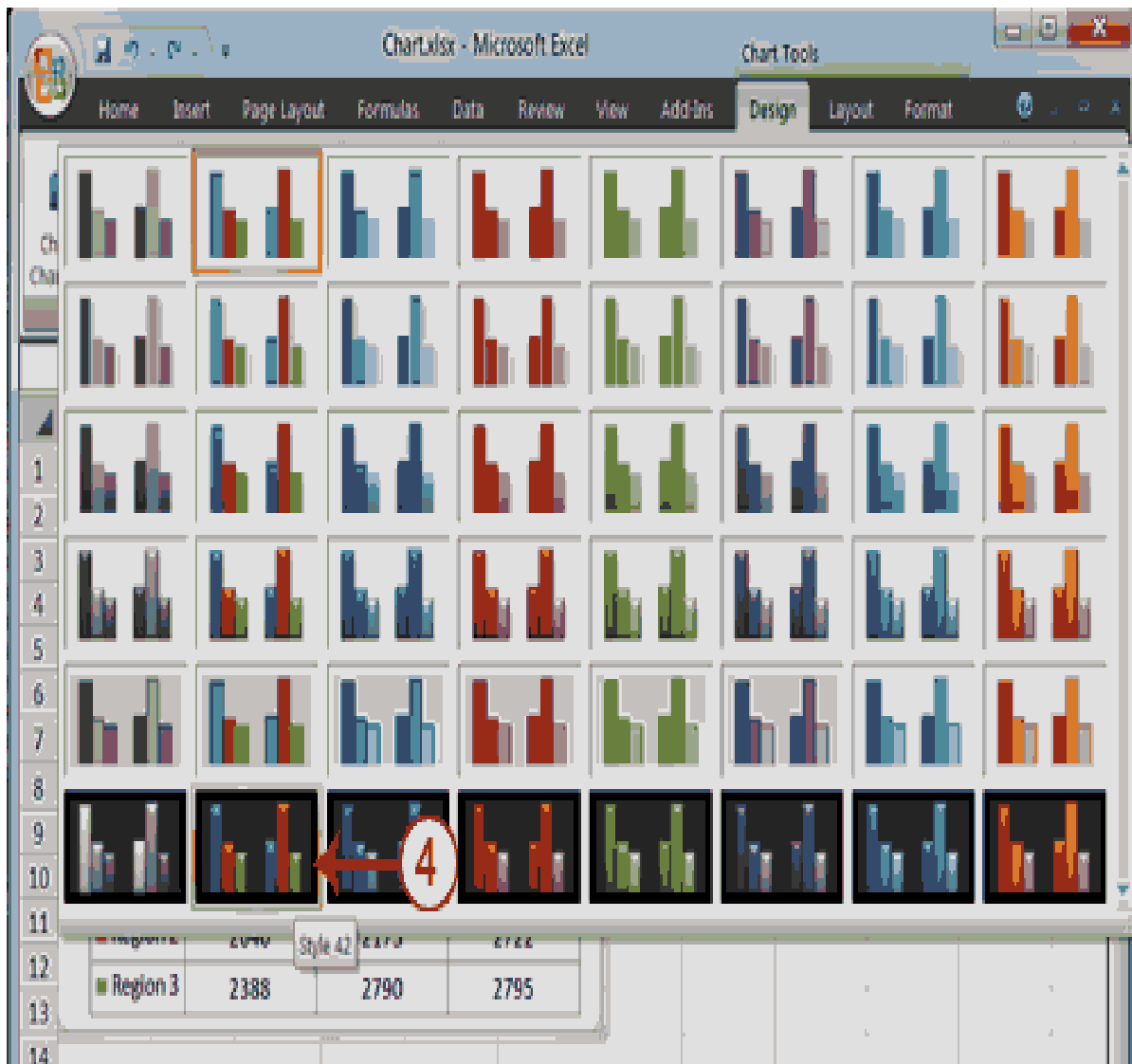
A style is a set of formatting options. You can use a style to change the color and format of your chart. Excel has several predefined styles that you can use. They are numbered from left to right, starting with 1, which is located in the upper-left corner.



**Fig.**

### **Steps to Change the Style of a Chart**

1. Click your chart. The Chart Tools become available.
2. Choose the Design tab.
3. Click the More button in the Chart Styles group. The chart styles appear.



**Fig**

4. Click Style. Excel applies the style to your chart.

#### **4.9 DATA PRESERVATION**

Data preservation is the act of conserving and maintaining both the safety and integrity of data. Preservation is done through formal activities that are governed by policies, regulations, and strategies directed towards protecting and prolonging the existence and authenticity of data. Data preservation refers to maintaining access to data and files over time. For data to be preserved, at minimum, it must be stored in a secure location, stored across multiple locations, and saved in file formats that will likely have the greatest utility in the future. Data preservation provides the usability of data beyond the lifetime of the research activity that generated them.



## **Definition**

Data preservation consists of a series of managed activities necessary to ensure continued stability and access to data for as long as necessary. For data to be preserved, at minimum, it must be stored in a secure location, stored across multiple locations, and saved in open file formats that will likely have the greatest utility in the future. Part of the preservation process can include depositing data in an institutional, discipline-specific, or generalist data repository, all of which allow for publication and preservation.

- The new NIH Data Management and Sharing Policy requires data to be preserved and shared, so medical researcher submits their COVID data to the National COVID Cohort Collaborative (N3C), as listed in the Open Domain-Specific Data Sharing Repository.
- There are many different ways to preserve digital information, including text, photos, audio, and video. Some strategies to preserve digital information include: Refreshing: transferring data to the same format. An example would be the transfer of music from an old CD-ROM to a new CD-ROM.

Digital information is an important source in our knowledge economy, valuable for research and education, science and the humanities, creative and cultural activities, and public policy. New high-throughput instruments, telescopes, satellites, accelerators, supercomputers, sensor networks, and running simulations are generating massive amounts of data. These data are used by decision-makers to improve the quality of life of citizens. Moreover, researchers are employing sophisticated technologies to analyze these data to address questions that were unapproachable just a few years ago. Digital technologies have fostered a new world of research characterized by immense datasets, unprecedented levels of openness among researchers, and new connections among researchers, policymakers, and the public domain. Different types of threats are:

1. Users may be unable to understand or use the data,
2. Lack of sustainable hardware, software, or support of the computer environment may make the information inaccessible,
3. Evidence may be lost because the origin and authenticity of the data may be uncertain,
4. Access and use restrictions (e.g., Digital Rights Management) may not be respected in the future,
5. Loss of ability to identify the location of data,
6. The current custodian of the data, whether an organization or project, may cease to exist at some point in the future.

### ➤ Importance of Data Preservation

Preservation helps protect you from hardware obsolescence. You never want to find yourself in a situation where all of your data is saved on unsupported hardware! Always migrate to new hardware formats so that your data will be available long-term. The most responsible way to preserve your data is to turn it over to a responsible custodian such as a data repository. When possible, try to preserve research data in a repository that provides *data curation services*, not just preservation services. Curated data is more valuable, easier to reuse, easier to locate, and more highly cited. Many data repositories have requirements for deposit - they may only accept certain types of data and have file size limits.



### 4.10 DATA PRESERVING VS. STORING DATA

Preserving is different from storing and backing up data files while your research is still ongoing. The latter typically involves mutable data; the former concerns data (or milestone versions of data) that are ‘frozen’ and not in active use. Long-term preservation requires appropriate actions to prevent data from becoming unavailable and unusable over time, for example, because of:

- Outdated software or hardware
- Storage media degradation
- A lack of sufficient descriptive and contextual information to keep data understandable

In other words, data preservation involves more than simply not deleting the data files created and stored. Maintaining data in a usable form for the longer term takes effort and has a considerable cost. Selecting which (parts of) data to keep, and for how long, is, therefore, an essential component of data preservation.

As a researcher, you have a key role in deciding what to retain and what not, as you know your

data best. Such decisions may depend on factors such as:

- The type of data involved
- The norms in your discipline
- Whether you are keeping data for potential future reuse, verification, or other purposes. Depending on the purpose, you may need to keep the raw data or data in a more processed form.

#### **4.11 DATA PRESERVATION VS. RETENTION OF DATA**

- Data retention is a central component of records management and information governance. Retention refers to the storing of data to meet regulatory and recordkeeping obligations
- The preservation is related to the safekeeping of electronically stored information (ESI) for some anticipated legal matter. In other words, data retention is a proactive ongoing process.
- Retention is usually a mandated requirement for researchers - it's the task that ensures that a bare minimum of data will remain available in some format.
- Preservation refers to having an active plan to ensure that when you do need to access your old data, it's readily available and can be easily accessed and manipulated by whoever needs it. When making a plan for data preservation you should include activities such as:
  - Transferring data from older storage formats to newer ones. This will ensure that the technology required to access your data is still available.
  - Transferring data from older file formats to newer ones. This will ensure that your data can still be opened by current software applications.
  - Having multiple copies of the data in different locations. This will ensure that your data is not lost in an unexpected event, such as theft, flood, or fire.
  - Ensuring your data is well documented, such as making notes on software used for creating codebooks as outlined in the Documentation and Description section of this guide. This will ensure that when you come back to access your data, you'll be able to remember what it all means.

#### **4.11 QUESTIONS FOR PRACTICE**

1. Define data processing. What are the various stages of data processing?
2. What is the purpose of data processing?
3. What is the significance of data processing in the research?
4. Explain the various rules for data processing.

5. What is the role of Excel in the presentation of data processing?
6. What is chart layout? How it can be applied in Excel?
7. What is the utility of the style of the chart in data processing? How it can be applied?
8. Define data preservation. What are the various challenges and threats to the digital preservation of data?
9. Explain data preservation Vs. storage of data.
10. Explain in which situation various types of charts are used in data processing:
  1. Pie chart
  2. Bar chart
  3. Line and area Chart
  4. Surface chart
  5. Radar chart

#### **4.12 SUGGESTED READINGS**

- A. Abebe, J. Daniels, J.W. Mckean, “Statistics and Data Analysis”.
- Clarke, G.M. & Cooke, D., “A Basic Course in Statistics”, Arnold.
- David M. Lane, “Introduction to Statistics”.
- S.C. Gupta and V.K. Kapoor, “Fundamentals of Mathematical Statistics”, SultanChand & Sons, New Delhi.

**COURSE NAME: STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY**

**SARM 1: INTRODUCTION TO STATISTICS**

**SEMESTER I**

**UNIT 5: SAMPLE, POPULATION, CHARACTERISTICS OF GOOD SAMPLE, TYPE OF SAMPLING TECHNIQUES, SAMPLING ERRORS**

**STRUCTURE**

**5.0 Learning objectives**

**5.1 Introduction Sample and Population**

**5.2 Purpose of Sampling**

**5.3 Characteristics of Good Sample**

**5.4 Types/ Procedures/ Methods/ Techniques of Sampling**

**5.4.1 Probability Sampling/ Random Sampling**

**5.4.1.1 Simple Random Sample**

**5.4.1.2 Systematic Random Sample**

**5.4.1.3 Stratified Random Sample**

**5.4.1.4 Cluster/ Multistage Sample**

**5.4.2 Non-Probability Sampling/ Non-random Sampling**

**5.4.2.1 Convenience/ Availability**

**5.4.2.2 Quota Sampling**

**5.4.2.3 Judgment/ Subjective/ Purposive Sampling**

**5.4.2.4 Snowball Sampling**

**5.5 Sampling and Non-sampling Errors**

**5.6 Sum Up**

**5.7 Questions for Practice**

**5.8 Suggested Readings**

**5.0 OBJECTIVE LEARNINGS**

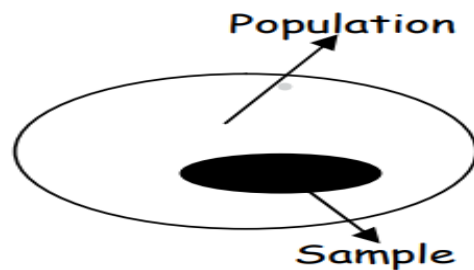
After studying this unit, you should be able to:

- explain the concepts of sample and population
- define the characteristics of good sample
- about types of sampling
- explain different sampling techniques

## 5.1 INTRODUCTION SAMPLE AND POPULATION

The total of items about which information is desired is called population or Universe. Which is further classified into two categories i.e., finite and infinite. A finite population is consisting of a fixed number of elements so that it is possible to count in its totality. A suitable example of a finite population is the population of a city, the number of students in the class, the number of workers in a factory, etc. While an infinite population is a population in which it is theoretically impossible to observe or count all the elements. In the case of an infinite population, the number of items is infinite. A suitable example of an infinite population is the number of stars in the sky and a number of hairs in the head. The use of the term infinite population for a population that cannot be counted in a reasonable period of time.

In statistics, population is the aggregate of objects, animate or inanimate under study in any statistical investigation.



A sample is a finite subset of the population that is selected from it with the objective of analyzing its properties, and the number of units in the sample is known as the sample size. By examining only the objects or items that are part of the sample, sampling is a method that enables us to make inferences about the features of the population. So, sample is a part of the population that represents the characteristics of the population, while Sampling is the process of selecting the sample for estimating the population characteristics. It is the process of obtaining information about an entire population by examining only a part of it.

The main objectives of the sampling theory are

1. To obtain the optimum results, i.e., the maximum information about the characteristics of the population with the available sources at our disposal in terms of time, money and manpower by studying the sample values only.
2. To get the most accurate population parameter estimates.

## **5.2 PURPOSE OF SAMPLING**

Providing an estimate of the population parameter and testing the hypothesis are the two main purposes of sampling. In order to infer information about the entire population and draw conclusions about it, we are gathering data from a smaller subset of a larger population.

- Through sampling a researchers can able to collect data more efficiently and cost-effectively compared to studying the entire population. It is often unfeasible, expensive, and time-consuming to survey or watch every person in a population. Researchers can save money by choosing a representative sample while still obtaining useful information.
- Enable collection of comprehensive data.
- Enable more accurate measurement as it is conducted by trained and experienced investigators.
- Sampling remains the only way when the population contains infinitely many members.
- In certain situations, sampling is the only way of data collection. For example, in testing the pathological status of blood, boiling status of rice, etc.
- Sampling enables statistical analysis and population parameter forecast. Researchers can use statistical methods to analyse data from a representative sample, estimate population characteristics, and compute confidence intervals or margins of error. This offers a degree of accuracy and permits drawing conclusions about the population.
- To improve methodologies, processes, or protocols, sampling is frequently employed during pilot testing or pre-testing phases of the study. Before beginning the larger study, researchers can find any problems or difficulties by gathering data from a smaller sample first. By doing this, the research's validity and quality are enhanced.
- It provides a valid estimation of sampling error.

## **5.3 CHARACTERISTICS OF GOOD SAMPLING**

The process of gathering data from a selected group or sample of the population that is relevant to

a research study is known as sample size. Researchers get data from a manageable group that is representative of the whole population rather than the complete population and solve the objective or purpose of sample collection. The characteristics of the good sample, which is chosen from the population are as:

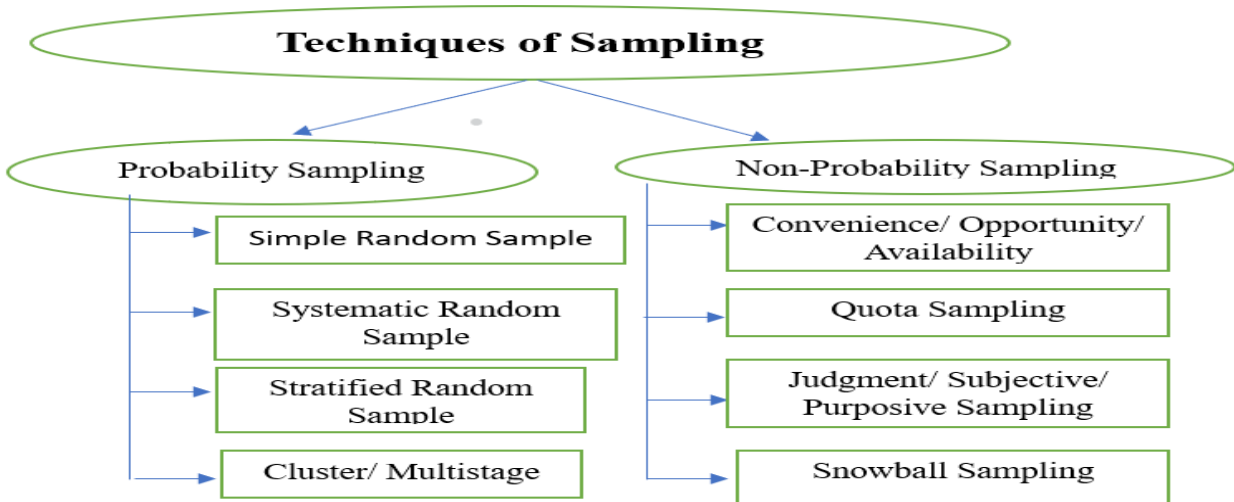
- **Good Representativeness:** The sample should represent the population that serves the purpose. It should represent a large number of the population. This makes it possible to apply the sample's results to the entire population.
- **Sampling Method:** Common sampling techniques include random sampling, stratified sampling, cluster sampling, and convenience sampling. The choice of sampling method depends on the research objectives, available resources, and the desired level of representativeness.
- **Sample Size:** The number of observations that make up the sample is referred to as the sample size. Choosing the right sample size includes the research objectives, desired level of accuracy, anticipated population variability, and required statistical power. Even though a large sample size typically yields accurate and reliable results, it might require more resources.
- **Sampling Frame:** A list or framework known as a sampling frame identifies the population from which the sample will be taken. It serves as the basis for selecting the sample and, from the population. To ensure the sample is representative, the sampling frame needs to be comprehensive and relevant.
- **Data Collection from the Sample:** Data is gathered from the individuals or observations that have been chosen after the sample. Various data collection techniques, including surveys, interviews, observations, and experiments, may be used in this. The study objectives and the characteristics of the variables being measured should be consistent with the method that is to be chosen.
- **Sampling Bias:** Sampling bias is the systematic inaccuracy or misrepresentation in the sample that happens when some people or groups are either overrepresented or underrepresented. Here, the validity and generalizability of the research findings may be affected. In order to minimize the biases during the sampling process, researchers need to be aware of them.

Sample measurement plays a crucial role in research as it allows researchers to collect data efficiently and make inferences about the larger population. Careful consideration of the sampling



method, sample size, and representativeness is important to ensure the reliability and validity of the research findings.

## 5.4 TYPES/ PROCEDURES/ METHODS/ TECHNIQUES OF SAMPLING



There are two basic approaches to sampling:

- Probability Sampling
- Non-probability Sampling

### 5.4.1 PROBABILITY SAMPLING

Probability sampling is also known as random sampling or chance sampling. In this, sample is taken in such a manner that every unit of the population has an equal and positive chance of being selected.

In this way, it is ensured that the sample would truly represent the overall population. Probability sampling can be achieved by random selection of the sample among all the units of the population.

Major random sampling procedures are:

- Simple Random Sample
- Systematic Random Sample
- Stratified Random Sample
- Cluster/ Multistage Sample

**5.4.1.1 Simple Random Sample:** For this, each member of the population is numbered. Then, a given size of the sample is drawn with the help of a random number chart. The other way is to do

a lottery. Write all the numbers on small, uniform pieces of paper, fold the papers, put them in a container, and take out the required lot in a random manner from the container as is done in the kitty parties. It is relatively simple to implement but the final sample may miss out on small sub groups.

**5.4.1.2 Systematic Random Sample:** It also requires numbering the entire population. Then every  $n$ th number (say every 5th or 10th number, as the case may be) is selected to constitute the sample. It is easier and more likely to represent different subgroups.

The sampling interval ( $k$ ) is calculated by dividing the total population size ( $N$ ) by the desired sample size ( $n$ ). Mathematically,  $k = N / n$ , where " $N$ " is the population size, and " $n$ " is the sample size. To introduce randomness into the sample selection process, you randomly select a starting point within the population. You can use random number tables, computer-generated random numbers, or other randomization techniques to pick the initial element. After that select the Sample, once you have the sampling interval ( $k$ ) and a random starting point, you select every " $k^{\text{th}}$ " element in the population as part of your sample until you reach the desired sample size ( $n$ ). If you reach the end of the population, you wrap around and continue the selection process until you complete the sample. The primary advantage of systematic random sampling is that it can generate a representative sample from a large population and is usually simple to apply. To prevent biases in the sample, it is essential to be sure the population is sufficiently randomized before choosing its starting point.

**5.4.1.3 Stratified Random Sample:** At first, the population is first divided into groups or strata each of which is homogeneous with respect to the given characteristic feature. From each strata, then, samples are drawn at random. This is called stratified random sampling. For example, with respect to the level of socio-economic status, the population may first be grouped in such strata as high, middle, low and very low socio-economic levels as per pre-determined criteria, and random sample drawn from each group. The sample size for each sub-group can be fixed to get representative sample. This way, it is possible that different categories in the population are fairly represented in the sample, which could have been left out otherwise in a simple random sample.

As with stratified samples, the population is broken down into different categories. However, the size of the sample of each category does not reflect the population as a whole. The Quota sampling technique can be used where an unrepresentative sample is desirable (e.g., you might want to

interview more children than adults for a survey on computer games), or where it would be too difficult to undertake a stratified sample.

**5.4.1.4 Cluster/ Multistage Sample:** In some cases, the selection of units may pass through various stages, before you finally reach your sample of study. For this, a State, for example, may be divided into districts, districts into blocks, blocks into villages, and villages into identifiable groups of people, and then taking the random or quota sample from each group. For example, taking a random selection of 3 out of 15 districts of a State, 6 blocks from each selected district, 10 villages from each selected block and 20 households from each selected village, totaling 3600 respondents. This design is used for large-scale surveys spread over large areas. The advantage is that it needs a detailed sampling frame for selected clusters only rather than for the entire target area. There are savings in travel costs and time as well. However, there is a risk of missing important sub-groups and not having a complete representation of the target population.

#### **CHECK YOUR PROGRESS (A)**

Q1. What is sampling?

Ans: \_\_\_\_\_  
\_\_\_\_\_

Q2. Define Simple random sample with an example.

Ans: \_\_\_\_\_  
\_\_\_\_\_

Q3. Explain systematic random sample.

Ans: \_\_\_\_\_  
\_\_\_\_\_

Q4. What is stratified random sampling?

Ans: \_\_\_\_\_  
\_\_\_\_\_

Q5. Define Cluster/ Multistage sampling.

Ans: \_\_\_\_\_  
\_\_\_\_\_

#### **5.4.2 NON-PROBABILITY SAMPLING**

Non-probability sampling is any sampling technique where certain population categories have no

possibility of selection, i.e., are not covered, or when it is difficult to calculate the probability of selection. It entails choosing components based on presumptions about the population of interest, which serves as the selection criteria. Non-probability sampling prevents the estimate of sampling errors since the selection of elements is not random. Non-probability sampling is a non-random and subjective sampling technique in which the sampler's discretion or personal judgment is used to choose the population elements that will make up the sample.

Non-probability sampling includes:

- Convenience/ Opportunity/ Availability sampling
- Quota Sampling
- Judgment/ Subjective/ Purposive Sampling
- Snowball Sampling

**5.4.2.1 Convenience/ Opportunity/ Availability Sampling:** Convenience sampling (or opportunity sampling) is a type of non-probability sampling that involves the sample being drawn from that part of the population that is approachable. That is, a sample population is selected because it is readily available and convenient. Such a sample would not be sufficiently representative for the researcher to draw any scientific conclusions about the entire population from it. As an example, if the interviewer were to conduct such a survey at a shopping center early in the morning on a particular day, the people that could interview would be restricted to those present there at that time, which would not represent the views of other members of society in such an area. Pilot testing benefits the most from this kind of sample. This type of sampling is most useful for pilot testing. In convenience sampling, the major problem is that one can never be certain what population the participants in the study represent. The population is unknown, the method for selecting cases is random, and the cases studied probably don't represent any population you could come up with. However, there are some situations in which this kind of design has advantages - for example, survey designers often want to have some people respond to their survey before it is given out in the 'real' research setting as a way of making certain the questions make sense to respondents. For this purpose, availability sampling is not a bad way to get a group to take a survey, though in this case researchers care less about the specific responses given than whether the instrument is confusing or makes people feel bad.

**5.4.2.2 Quota Sampling:** A non-random sampling, which includes to divide the population into

predetermined categories or quotas according to specific criteria, and then choosing those who fit those quotas. When obtaining a random sample is challenging or when access to a sampling frame is constrained, quota sampling is frequently used. In quota sampling, the population is first segmented into mutually exclusive sub-groups, just as similar to stratified sampling. These categorizations can be the demographic basis (e.g., age, gender, education level) or behavioral (e.g., purchasing habits, brand preference). Then judgment is used to select the subjects or units from each segment based on a specified proportion. For instance, a sample of 300 men and 200 women between the ages of 45 and 60 may be asked of the interviewer. The selection of the sample in quota sampling is not random. Interviewers might be drawn to individuals who appear to be most helpful, for instance. Because not everyone is given the opportunity to be chosen, the issue is that these samples may be biased. Its biggest drawback is this random component, and the relative merits of quota and probability have long been a source of debate.

**5.4.2.3 Subjective or Purposive or Judgment Sampling:** In this technique, the sample is selected with a pre-definite purpose in view and the choice of the sampling units depends entirely on the judgment of the researcher. This sampling suffers from drawbacks of favoritism and partiality depending upon the beliefs and prejudgments of the researcher and does not give a representative sample of the population. This method is rarely used and cannot be recommended for general as based on biasness due to an element of subjectivity on the part of the researcher. However, judgment samples may produce useful results if the investigator is knowledgeable and talented and this sampling is used based on judgment.

**5.4.2.4 Snowball Sampling:** snowball sampling is a non-probability sampling used in research to identify and participate people by way of suggestions from original participants. When investigating difficult-to-reach the population or when there isn't a complete list of the target population easily accessible, it is frequently used. Studying social networks, secret populations, or sensitive topics makes good use of snowball sampling.

This sampling technique is used against rare populations. Sampling is a big problem in this case, as the defined population from which the sample can be drawn is not available.

The process of snowball sampling typically first, identify an initial participant. This individual is usually well-connected within the target population and can help in the recruitment process. After that, a researcher conducts an interview or data collection with the initial participant. Afterward,

the participant is asked to provide referrals to others who meet the study's criteria or have relevant experiences or characteristics. Therefore, it creates referral chain, where a newly identified participants become part of the sample. The snowball sampling process continues iteratively until the desired sample size or data fullness is reached.

Therefore, the process sampling depends on the chain system of referrals. Although small sample sizes are the clear advantages of snowball sampling, bias is one of its disadvantages.

## **5.5 SAMPLING AND NON-SAMPLING ERRORS**

Statistically, "error" refers to the discrepancy between the true value and the estimated or approximate value. In other words, error refers to the difference between the true value of a population parameter and its estimate provided by an appropriate sample statistic computed by some statistical device. As a result, the term "error" has a very specific and different meaning in the field of statistics. The existence of these errors the various factors like approximations in measurements, rounding, biases because of incorrect data collection and analysis and personal biasness of the researcher.

Errors are of two types:

- Sampling Errors
- Non-sampling Errors

**Sampling error:** This type of error arises due to the variability in the process of selecting a sample from a larger population. As only a small population is measured, so the results are different from the census. Errors attributed to fluctuations of sampling are called as sampling errors. Reasons for sampling errors are:

- Incorrect sample selection
- Biasness in the estimation method
- Heterogeneity of the population
- Sample size, smaller samples are more susceptible to larger sampling errors, while larger samples tend to have smaller sampling errors, assuming random sampling
- Method of sampling
- Non-Response Bias, when individuals from the sample do not respond to the survey
- Measurement Error, arises from inaccuracies in data collection or measurement

**Non-sampling Errors:** Non-sampling errors are errors that occur in the data collection methods other than those caused by the sampling procedure. These can be because of various sources and can impact the quality as well as accuracy of the collected data. Non-sampling errors can be attributed to factors such as data collection, data processing, measurement, and analysis.

Reasons for non-sampling errors are:

- Respondent Errors, because of misunderstanding on the part of respondent.
- Non-response errors, bias happens when certain individuals selected for the sample do not participate or respond to the survey.
- Processing Errors, occur during data entry, coding, or data cleaning. Human errors as well as software bugs can lead to data inaccuracies during these stages.
- Data Editing Errors, occur when data is modified during the data cleaning process. While data cleaning is essential to ensure data quality, some of the errors introduce inaccuracies.
- Compiling and publishing errors.

### **CHECK YOUR PROGRESS (B)**

Q1. What are the types of non-probability sampling?

Ans: \_\_\_\_\_  
\_\_\_\_\_

Q2. Define convenience sampling.

Ans: \_\_\_\_\_  
\_\_\_\_\_

Q3. What is Quota Sampling?

Ans: \_\_\_\_\_  
\_\_\_\_\_

Q4: Define Judgment Sampling.

Ans: \_\_\_\_\_  
\_\_\_\_\_

Q5. What is Snowball Sampling?

Ans: \_\_\_\_\_  
\_\_\_\_\_

Q6. Errors in sampling.

Ans: \_\_\_\_\_

---

## **5.6 SUM UP**

A sample is a part of the population that represents the characteristics of the population, while Sampling is the process of selecting the sample for estimating the population characteristics. It is the process of obtaining information about an entire population by examining only a part of it. There are different sampling techniques available in statistics depending on the research. These are probability and non-probability sampling techniques. Probability sampling techniques are Simple Random Sample, Systematic Random Sample, Stratified Random Sample, Cluster/ Multistage Sample. While non probability sampling is Convenience/ Opportunity/ Availability sampling, Quota Sampling, Judgment/ Subjective/ Purposive Sampling, and Snowball Sampling. There are two types of error in the sample which are sampling and non-sampling error. Sampling error is the error that arises due to the variability in the process of selecting a sample from a larger population. As only a small population is measured, so the results are obvious. While the non-sampling error is because of various sources and can impact the quality as well as accuracy of the collected data. Non-sampling errors can be attributed to factors such as data collection, data processing, measurement, and analysis.

## **5.7 QUESTIONS FOR PRACTICE**

- Q1. What do you mean by sample and population? Explain with an example.
- Q2. What are the techniques for the collection of data available in statistics?
- Q3. What do you mean by primary data? Give the sources of primary data.
- Q4. What is the questioner? What are the points to keep in mind before drafting questioner?
- Q5. Explain the term secondary data with its sources.
- Q6. Give limitations of primary data and secondary data.
- Q7. what are the precautions to collect secondary data?
- Q8. Explain sampling and non-sampling errors.

## **5.8 SUGGESTED READINGS**

- A. Abebe, J. Daniels, J.W. Mckean, “Statistics and Data Analysis”.



- Clarke, G.M. & Cooke, D., “A Basic Course in Statistics”, Arnold.
- David M. Lane, “Introduction to Statistics”.
- S.C. Gupta and V.K. Kapoor, “Fundamentals of Mathematical Statistics”, Sultan Chand & Sons, New Delhi.

**COURSE NAME: STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY**

**SARM 1: INTRODUCTION TO STATISTICS**

**SEMESTER I**

**UNIT 6: MEASURES OF CENTRAL TENDENCY- MEAN (DIRECT, SHORT CUT AND  
STEP DEVIATION METHODS)**

**STRUCTURE**

**6.0 Learning Objectives**

**6.1 Introduction**

**6.2 Meaning of Average or Central Tendency**

**6.3 Objectives and Functions of Average**

**6.4 Requisites or Features of Good Average**

**6.5 Measures of Central Tendency**

**6.6 Arithmetic Mean**

**6.6.1 Arithmetic Mean in individual series**

**6.6.2 Arithmetic Mean in discrete series**

**6.6.3 Arithmetic Mean in continuous series**

**6.6.4 Arithmetic Mean in cumulative frequency series**

**6.6.5 Arithmetic Mean in unequal series**

**6.6.6 Combined Arithmetic Mean**

**6.6.7 Correcting incorrect Arithmetic Mean**

**6.6.8 Properties of Arithmetic Mean**

**6.6.9 Merits of Arithmetic Mean**

**6.6.10 Limitations of Arithmetic Mean**

**6.7 Sum Up**

**6.8 Questions for Practice**

**6.9 Suggested Readings**

## 6.0 LEARNING OBJECTIVES

After studying the Unit, students will be able to:

- Meaning of Average
- Features of a good measure of Average
- Find different types of averages for various types of data
- Understand the relation that exists between different types of Averages
- Procedure to find out mean of different series
- Merits and limitations of each type of average

## 6.1 INTRODUCTION

We can say that the modern age is the age of Statistics. There is no field in modern life in which statistics is not used. Whether it is Business, Economics, Education. Government Planning or any other field of our life, statistics is used everywhere. Business manager use statistics for business decision making, Economists use statistics for economic planning, Investors use statistics for future forecasting and so on. There are many techniques in statistics that helps us in all these purposes. Average or Central Tendency is one such technique that is widely used in statistics. This technique is used almost in every walk of the life.

## 6.2 MEANING OF AVERAGE OR CENTRAL TENDENCY

Average or Central tendency is the most used tool of statistics. This is the tool without which statistics is incomplete. In simple words we can say that the Average is the single value which is capable of representing its series. It is the value around which other values in the series move. We can define Average as the single typical value of the series which represents the whole series data. Following is the popular definition of average:

According to **Croxton and Cowden** "An average is a single value within the range of data that is used to represent all values in the series. Since an average is somewhere within the range of the data, it is also called a measure of Central Value".

## 6.3 OBJECTIVES AND FUNCTIONS OF AVERAGE

1. **Single Value representing whole Data:** In statistics data can be shown with the help of tables and diagrams. But sometimes data is very larger and it is not easy to present in a table or graph.

So, we want to represent that data in summarised form. Average helps us to represent data in summarised form. For example, the data of national income of India is very large but when we calculate per capita income it gives us idea of the national income.

2. **To Help in Comparison:** In case we want to compare two different series of data, it is very difficult to compare. There are many difficulties like a number of items in the series may be different. In such case, average helps us in making the comparison. For example, if we want to compare income of people living in different countries like India and Pakistan, we can do so by calculating per capita income which is a form of average.
3. **Draw a conclusion about Universe from Sample:** This is one of the important functions of the average. If we take the average of a sample, we can draw certain conclusions about the universe from such Average. For example, the mean of a sample is representative of its universe.
4. **Base of other Statistical Methods:** There are many Statistical Techniques that are based on average. If we don't have an idea about the average, we cannot apply those techniques. For example, Dispersion, Skewness, Index Number are based on average.
5. **Base of Decision Making:** Whenever we have to make certain decision, average plays very crucial role in the decision making. From the average we could have idea about the data and on the basis of that information we can take decision. For example, a company can take decision regarding its sales on the basis of average yearly sales of past few years.
6. **Precise Relationship:** Average helps us to find out if there is precise relation between two variables or two items. It also removes the biasness of the person making analysis. For example, if you say that Rajesh is more intelligent than Ravi it is only our personal observation and does not make any precise relation. If we compare the average marks of both the students, we could have a precise relation.
7. **Helpful in Policy Formulation:** Average helps the government in formulation of the policy. Whenever government has to formulate economic policy they consider various averages like per capita income, average growth rate etc.

#### 6.4 REQUISITE / FEATURES OF GOOD AVERAGE

1. **Rigidly Defined:** A good measure of average is one which is having a clear-cut definition and there is no confusion in the mind of person who is calculating the average. In case person applies his discretion while calculating the average, we cannot say that average is a good

measure. Good average must have fix algebraic formula, so that whenever average of same data is calculated by two different persons, result is always same.

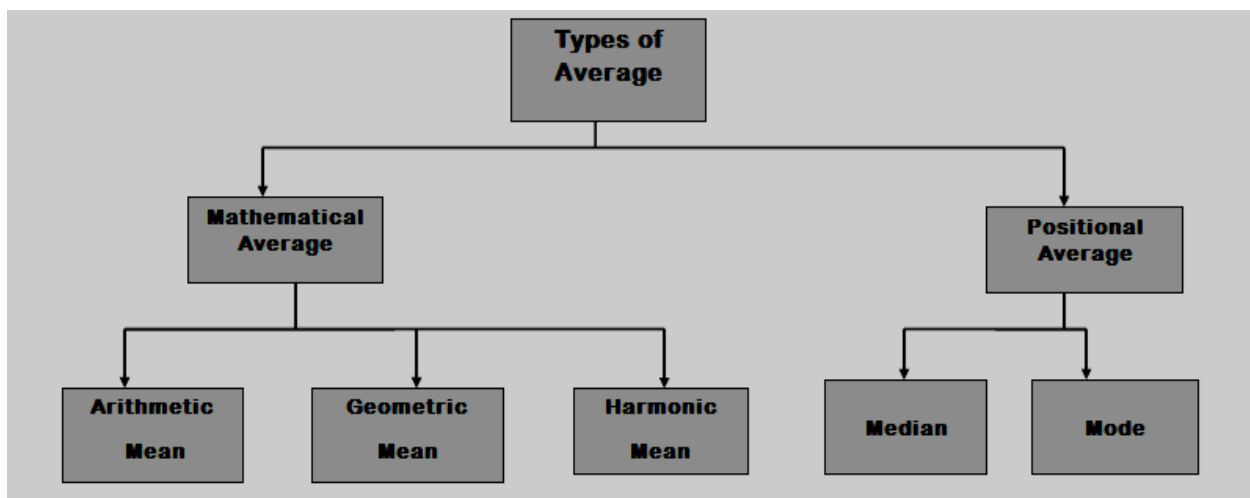
2. **Easy to Compute:** Good average is one which does not involve much calculation and are easy to compute. A good average is one which can be calculated even by a person having less knowledge of Statistics. If it is very difficult to calculate the average, we cannot regard it as a good measure.
3. **Based on all Observations:** Good average must consider all the values or data that is available in the series. If average is based on only few observations of the series, we cannot say that it is a good measure of average.
4. **Not affected by Extreme Values:** A good measure of average is one which is not affected by the extreme values present in the Data. Sometime data contains values which are not within normal limits, these values are called extreme values. If average is affected by these extreme values, we cannot claim that average is a good measure.
5. **Representative of whole Series:** A good measure of average is one which represent characteristics of whole series of the data.
6. **Easy to Understand:** A good measure of average is one that is not only easy to understand but also easy to interpret.
7. **Not Affected by Fluctuations in the Sampling:** If we take one sample from the universe and calculate average, then we draw another sample from the same universe and calculate the average again, there must not be much difference between these two averages. If average significantly change with the change in sample, we cannot treat it as a good measure of average.
8. **Capable of further Algebraic Treatment:** a good average is one on which we can apply further algebraic treatment. In case further algebraic treatment is not possible, we cannot say that it is a good average. Sch further algebraic treatment may be anything like calculating combined average when average of two different series is available.
9. **Located Graphically:** It will be better if we can locate average graphically also. Graphs are easy to understand and interpret, so the average that can be located graphically is a good average.

As no single average has all these features, we cannot say which measure of average is best. Each measure has its own merits and limitation. Moreover, each measure is suitable for

particular situation.

## 6.5 MEASURES OF CENTRAL TENDENCY

There are many methods through which we can calculate average or central tendency. We can divide these methods into two categories that are Algebraic Method and Positional Average. Algebraic methods are those in which the value of average depends upon the mathematical formula used in the average. The mathematical average can further be divided into three categories that are Arithmetic Mean, Geometric Mean and Harmonic Mean. On the other hand, positional averages are those average which are not based on the mathematical formula used in calculation of average rather these depends upon the position of the variable in the series. As these depends upon the position of the variable, these averages are not affected by the extreme values in the data. Following chart shows different types of averages.



## 6.6 ARITHMETIC MEAN

It is the most popular and most common measure of average. It is so popular that for a common man the two terms Arithmetic Mean and Average are one and the same thing. However, in reality these two terms are not same and arithmetic mean is just one measure of the average. We can define the arithmetic mean as:

“The value obtained by dividing sum of observations with the number of observations”.

So arithmetic mean is very easy to calculate, what we have to do is just add up the value of all the items given in the data and then we have to divide that total with the number of items in the data. Arithmetic mean is represented by symbol A. M. or  $\bar{x}$

### 6.6.1 Arithmetic Mean in case of Individual Series

Individual series are those series in which all the items of the data are listed individually. There are two methods of finding arithmetic mean in the individual series. These two methods are Direct method and Shortcut Method.

**1. Direct Method** According to this method calculation of mean is very simple and as discussed above, we have to just add the items and then divide it by number of items. Following are the steps in calculation of mean by direct method:

1. Suppose our various items of the data are  $X_1, X_2, X_3, \dots, X_n$
2. Add all the values of the series and find  $\sum X$ .
3. Find out the number of items in the series denoted by  $n$ .
4. Calculate arithmetic mean dividing sum value of observation with the number of observations using following formula:

$$\bar{x} = \frac{X_1 + X_2 + X_3 + \dots + X_n}{N} = \frac{\sum X}{N}$$

Where  $\bar{x}$  = Mean

$N$  = Number of items

$\sum X$  = Sum of observation

**Example 1. The daily income of 10 families is as given below (in rupees) :**

**130, 141, 147, 154, 123, 134, 137, 151, 153, 147**

**Find the arithmetic mean by direct method.**

**Solution:** Computation of Arithmetic Mean

Serial No.	Daily Income (in Rs.) X
1	130
2	141
3	147
4	154
5	123
6	134
7	137

8	151
9	153
10	147
N = 10	$\sum X = 1417$

A. M.,  $\bar{x} = \frac{X_1 + X_2 + X_3 + \dots + X_N}{N} = \frac{\sum X}{N} = \frac{1417}{10} = \text{Rs. } 141.7$

**2. Short Cut Method:** Normally this method is used when the value of items is very large and it is difficult to make calculations. Under this method we take one value as mean which is known as assumed mean and deviations are calculated from this as you mean. This method is also known as assumed mean method. Following are the steps of this method:

1. Suppose our various items of the data are  $X_1, X_2, X_3, \dots, X_n$
2. Take any value as assumed mean represented by 'A'. This value may be any value among data or any other value even if that is not presented in data.
3. Find out deviations of items from assumed mean. For that deduct Assumed value from each value of the data. These deviations are representing as 'dx'
4. Find sum of the deviations represented by  $\sum dx$ .
5. Find out the number of items in the series denoted by n.
6. Calculate arithmetic mean dividing sum deviations of the observation with the number of observations using following formula:

$$\bar{x} = A + \frac{\sum dx}{N}$$

Where  $\bar{x}$  = Mean

A = Assumed Mean

N = Number of items

$\sum dx$  = Sum of deviations

**Example 2. Calculate A. M. by short - cut method for the following data**

<b>R. No</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>	<b>10</b>
<b>marks</b>	<b>50</b>	<b>60</b>	<b>65</b>	<b>88</b>	<b>68</b>	<b>70</b>	<b>83</b>	<b>45</b>	<b>53</b>	<b>58</b>

**Solution:** Let assumed Mean (A) be 60



R. No.	Marks (X)	dx = X - A
1	50	-10
2	60	0
3	65	5
4	88	28
5	68	8
6	70	10
7	83	23
8	45	-15
9	53	-7
10	58	-2
N = 10		$\sum dx = 40$

As  $\bar{X} = A + \frac{\sum dx}{N}$

$\Rightarrow \bar{X} = 60 + \frac{40}{10} = 60 + 4$

$\Rightarrow \bar{X} = 64$  Marks

### 6.6.2 Arithmetic Mean in case of Discrete Series

In individual series if any value is repeated that is shown repeatedly in the series. It makes series lengthy and make calculation difficult. In case of discrete series, instead of repeatedly showing the items we just group those items and the number of time that item is repeated is shown as frequency. In case of discrete series, we can calculate Arithmetic mean. By using Direct Method and Shortcut Method.

**1. Direct Method:** In indirect method we multiply the value of items (X) with their respective frequency (f) to find out the the product item (fX). Then we take up sun of the product and divide it with the number of items. Following are the steps

1. Multiply the value of items (X) with their respective frequency (f) to find out the the product item (fX)
2. Add up the product so calculated to find  $\sum fX$ .
3. Find out the number of items in the series denoted by n.

4. Calculate arithmetic mean dividing sum of the product with the number of observations using following formula:

$$\bar{x} = \frac{\sum fX}{N}$$

Where  $\bar{x}$  = Mean

N = Number of items

$\sum fX$  = Sum of product of observations.

**Example 3. Find the average income**

<b>Daily Income (in rupees)</b>	<b>200</b>	<b>500</b>	<b>600</b>	<b>750</b>	<b>800</b>
<b>No. of Workers</b>	<b>2</b>	<b>1</b>	<b>4</b>	<b>2</b>	<b>1</b>

**Solution:**

Daily Income (Rs.) X	No. of Workers Frequency (f)	fX
200	2	400
500	1	500
600	4	2400
750	2	1500
800	1	800
	$\sum f = 10$	$\sum fX = 5600$

$$\begin{aligned} \therefore \text{Average Income } \bar{x} &= \frac{\sum fX}{\sum f} = \frac{5600}{10} \\ &= \text{Rs. 560} \end{aligned}$$

- 2. Short Cut Method:** Under this method we take one value as mean which is known as assumed mean and deviations are calculated from this as you mean. Then average is calculated using assumed mean. Following are the steps of this method:

1. Suppose our items of the data are 'X' and its corresponding frequency is 'f'.
2. Take any value as assumed mean represented by 'A'.
3. Find out deviations of items from assumed mean. For that deduct Assumed value from each value of the data. These deviations are representing as 'dx'

4. Multiply the values of dx with corresponding frequency to find out product denoted by fdx
5. Find sum of the product so calculated represented by  $\sum fdx$ .
6. Find out the number of items in the series denoted by n.
7. Calculate arithmetic mean dividing sum deviations of the observation with the number of observations using following formula:

$$\bar{x} = A + \frac{\sum fdx}{N}$$

Where  $\bar{x}$  = Mean

A = Assumed Mean

N = Number of items

$\sum fdx$  = Sum of product of deviation with frequency.

**Example 4. From the following data find out the mean height of the students.**

<b>Height (in cms.)</b>	<b>154</b>	<b>155</b>	<b>156</b>	<b>157</b>	<b>158</b>	<b>159</b>	<b>160</b>	<b>161</b>	<b>162</b>	<b>163</b>
<b>No. of Students</b>	<b>1</b>	<b>6</b>	<b>10</b>	<b>22</b>	<b>21</b>	<b>17</b>	<b>14</b>	<b>5</b>	<b>3</b>	<b>1</b>

**Solution:** Let the Assumed Mean (A) be 150

Height in cms. X	No. of students f	dX = (X – A) = X – 150	fdX
154	1	4	4
155	6	5	30
156	10	6	60
157	22	7	154
158	21	8	168
159	17	9	153
160	14	10	140
161	5	11	55
162	3	12	36
163	1	13	13

	$\sum f = 100$		$\sum fdX = 813$
--	----------------	--	------------------

Applying the formula

$$\bar{x} = A + \frac{\sum fdX}{\sum f}$$

We get

$$\begin{aligned}\bar{x} &= 150 + \frac{813}{100} \\ &= 150 + 8.13 = 158.13\end{aligned}$$

$\therefore$  Mean Height = 158.13 cm

### 6.6.3 Arithmetic Mean in case of Continuous Series

Continuous series is also known as Grouped Frequency Series. Under this series the values of the observation are grouped in various classes with some upper and lower limit. For example, classes like 10-20, 20-30, 30-40 and so on. In the class 10-20 lower limit is 10 and upper limit is 20. So, all the observations having values between 10 and 20 are put in this class interval. Similar procedure is adopted for all class intervals. The procedure of calculating Arithmetic Mean in continuous series is just like discrete series except that instead of taking values of observations we take mid value of the class interval. The mid value is represented by 'm' and is calculated using following formula:

$$m = \frac{\text{Lower Limit} + \text{Upper Limit}}{2}$$

1. **Direct Method:** In indirect method we multiply the mid values (m) with their respective frequency (f) to find out the product item (fm). Then we take up sum of the product and divide it with the number of items. Following are the steps
  1. Multiply the mid values (m) with their respective frequency (f) to find out the product item (fm)
  2. Add up the product so calculated to find  $\sum fm$ .
  3. Find out the number of items in the series denoted by n.
  4. Calculate arithmetic mean by dividing sum of the product with the number of observations using following formula:

$$\bar{x} = \frac{\sum fm}{N}$$

Where  $\bar{x}$  = Mean

N = Number of items

$\sum fm$  = Sum of product of observations of mean and frequencies.

**Example 5. Calculate the arithmetic mean of the following data:**

Class Intervals C. I.	100 – 200	200 – 300	300 – 400	400 – 500	500 – 600	600 – 700
f	4	7	16	20	15	8

**Solution:**

Class Intervals (C. I.)	Mid Value (m)	Frequency (f)	fm
100 – 200	150	4	600
200 – 300	250	7	1750
300 – 400	350	16	5600
400 – 500	450	20	9000
500 – 600	550	15	8250
600 – 700	650	8	5200
		$\sum f = 70$	$\sum fm = 30,400$

As 
$$\bar{x} = \frac{\sum fm}{\sum f}$$

$$\therefore \bar{x} = \frac{30,400}{70} = 434.3$$

**2. Short Cut Method:** This method of mean is almost similar to calculation in the discrete series but here the assumed mean is selected and then the deviation is taken from mid value of the observations. Following are the steps of this method:

1. Calculate the Mid Values of the series represented by 'm'.
2. Take any value as assumed mean represented by 'A'.
3. Find out deviations of items from assumed mean. For that deduct Assumed value from mid values of the data. These deviations are representing as 'dm'

4. Multiply the values of  $dm$  with corresponding frequency to find out product denoted by  $fdm$
5. Find sum of the product so calculated represented by  $\sum fdm$ .
6. Find out the number of items in the series denoted by  $n$ .
7. Calculate arithmetic mean dividing sum deviations of the observation with the number of observations using following formula:

$$\bar{x} = A + \frac{\sum fdm}{N}$$

Where  $\bar{x}$  = Mean,  $A$  = Assumed Mean,  $N$  = Number of items

$\sum fdm$  = Sum of product of deviation from mid values with frequency.

**Example 6. Calculate the mean from the following data**

Daily Wages (Rs.)	0-100	100-200	200 – 300	300 – 400	400-500	500-600	600-700	700 – 800	800-900
No. of Workers	1	4	10	22	30	35	10	7	1

**Solution:** Let the assumed mean,  $A = 150$

Daily Wages (Rs.) C.I.	No. of Workers $f$	Mid Value $m$	$dm = m - A$ ( $m - 150$ )	$fdm$
0 – 100	1	50	-100	-100
100 – 200	4	150	0	0
200 – 300	10	250	100	1000
300 – 400	22	350	200	4400
400 – 500	30	450	300	9000
500 – 600	35	550	400	14,000
600 – 700	10	650	500	5000
700 – 800	7	750	600	4200
800 – 900	1	850	700	700
	$\sum f = 120$			$\sum fdm = 38,200$

As 
$$\bar{x} = A + \frac{\sum fdm}{\sum f}$$

$$= 150 + \frac{38,200}{120} = 150 + 318.33 = 468.33$$

$\Rightarrow \bar{x} = 468.33$

**3. Step Deviation Method:** Step Deviation method is the most frequently used method of finding Arithmetic Mean in case of continuous series. This method is normally used when the

class interval of the various classes is same. This method makes the process of calculation simple. Following are the steps of this method:

1. Calculate the Mid Values of the series represented by 'm'.
2. Take any value as assumed mean represented by 'A'.
3. Find out deviations of items from assumed mean. For that deduct Assumed value from mid values of the data. These deviations are representing as 'dm'.
4. Find out if all the values are divisible by some common factor 'C' and divide all the deviations with such common factor to find out dm' which is dm/c
5. Multiply the values of dm' with corresponding frequency to find out product denoted by fdm'
6. Find sum of the product so calculated represented by  $\sum fdm'$ .
7. Find out the number of items in the series denoted by n.
8. Calculate arithmetic mean dividing sum deviations of the observation with the number of observations using following formula:

$$\bar{x} = A + \frac{\sum fdm'}{\sum f} \times C$$

Where  $\bar{x}$  = Mean

A = Assumed Mean, N = Number of items, C = Common Factor

$\sum fdm'$  = Sum of product of deviation after dividing with common factors and multiplying it with frequency.

**Example 7. Use step deviation method to find  $\bar{x}$  for the data given below:**

<b>Income (Rs.)</b>	<b>1000</b>	<b>2000</b>	<b>3000</b>	<b>4000</b>	<b>5000</b>	<b>6000</b>
	<b>- 2000</b>	<b>- 3000</b>	<b>- 4000</b>	<b>- 5000</b>	<b>- 6000</b>	<b>- 7000</b>
<b>No. of Persons</b>	<b>4</b>	<b>7</b>	<b>16</b>	<b>20</b>	<b>15</b>	<b>8</b>

**Solution:** Let the assumed mean A = 4500

Income (Rs.) C.I.	No of Persons f	Mid Value m	dm = m - A = (m - 4500)	dm' = $\frac{dm}{C}$ C = 1000	fdm'
1000 - 2000	4	1500	-3000	-3	-12
2000 - 3000	7	2500	-2000	-2	-14
3000 - 4000	16	3500	-1000	-1	-16
4000 - 5000	20	4500	0	0	0

5000 – 6000	15	5500	1000	1	15
6000 – 7000	8	6500	2000	2	16
	$\sum f = 70$				$\sum f dm' = -11$

As  $\bar{x} = A + \frac{\sum f dm'}{\sum f} \times C$

$\therefore \bar{x} = 4500 + \frac{(-11)}{70} \times 1000 = 4500 - \frac{1100}{7}$

$= 4500 - 157.14 = 4342.86$

$\bar{x} = 4342.86$

### CHECK YOUR PROGRESS (A)

1. Following data pertains to the monthly salaries in rupees of the employees of a Mohanta Enterprises. Calculate the average salary per employ

3000, 4100, 4700, 5400, 2300, 3400, 3700, 5100, 5300, 4700

2. Calculate mean for the following data using the shortcut method.

700, 650, 550, 750, 800, 850, 650, 700, 950

3. Following is the height of students of class tenth of a school. Find out the mean height of the students.

Height in Inches	64	65	66	67	68	69	70	71	72	73
No. of students	1	6	10	22	21	17	14	5	3	1

4. Calculate A.M for the following frequency distribution of Marks.

Marks	5	10	15	20	25	30	35	40
No of students	5	7	9	10	8	6	5	2

5. Calculate mean for the following data

Marks	5-15	15-25	25-35	35-45	45-55	55-65
No of Students	8	12	6	14	7	3

6. Calculate mean for the given data by step deviation method

C.I	0-10	10-20	20-30	30-40	40-50	50-60	60-70
-----	------	-------	-------	-------	-------	-------	-------



f	8	12	14	16	15	9	6
---	---	----	----	----	----	---	---

### Answers

1) 4170	3) 68.13 inches	5) 31.8
2) 733.30	4) 20.48	6) 33.625

### Other Special case of Continuous Series

#### 6.6.4 Arithmetic Mean in case of Cumulative Frequency Series:

The normal continuous series give frequency of the particular class. However, in case of cumulative frequency series, it does not give frequency of particular class rather it gives the total of frequency including the frequency of preceding classes. Cumulative frequency series may be of two types, that are 'less than' type and 'more than' type. For calculating Arithmetic mean in cumulative frequency series, we convert such series into the normal frequency series and then apply the same method as in case of normal series.

#### Less than Cumulative Frequency Distribution

**Example 8. Find the mean for the following frequency distribution:**

<b>Marks Less Than</b>	<b>10</b>	<b>20</b>	<b>30</b>	<b>40</b>	<b>50</b>	<b>60</b>
<b>No. of Students</b>	<b>5</b>	<b>15</b>	<b>40</b>	<b>70</b>	<b>90</b>	<b>100</b>

**Solution:** Convert the given data into exclusive series:

Marks C. I.	No. of Students f	Mid Value m	dm = m - A A = 25	dm' = $\frac{dm}{C}$ C = 10	fdm'
0 - 10	5	5	-20	-2	-10
10 - 20	15 - 5 = 10	15	-10	-1	-10
20 - 30	40 - 15 = 25	25	0	0	0
30 - 40	70 - 40 = 30	35	10	1	30
40 - 50	90 - 70 = 20	45	20	2	40
50 - 60	100 - 90 = 10	55	30	3	30
	$\sum f = 100$				$\sum fdm' = 80$

As  $\bar{x} = A + \frac{\sum fdm'}{\sum f} \times C$

$\Rightarrow \bar{x} = 25 + \frac{80}{100} \times 10 = 33$

$$\Rightarrow \bar{x} = 33$$

### More Than Cumulative Frequency Distribution

**Example 9.** Find the mean for the following frequency distribution

<b>Marks More Than</b>	<b>0</b>	<b>10</b>	<b>20</b>	<b>30</b>	<b>40</b>	<b>50</b>	<b>60</b>	<b>70</b>	<b>80</b>	<b>90</b>
<b>No. of Students</b>	<b>80</b>	<b>77</b>	<b>72</b>	<b>65</b>	<b>55</b>	<b>43</b>	<b>28</b>	<b>16</b>	<b>10</b>	<b>8</b>

**Solution:** Convert the given data into exclusive series

Marks C.I.	No. of Students f	Mid Value	dm = m - A A = 55	dm' = $\frac{dm}{C}$ C = 10	fdm'
0 - 10	80 - 77 = 3	5	-50	-5	-15
10 - 20	77 - 72 = 5	15	-40	-4	-20
20 - 30	72 - 65 = 7	25	-30	-3	-21
30 - 40	65 - 55 = 10	35	-20	-2	-20
40 - 50	55 - 43 = 12	45	-10	-1	-12
50 - 60	43 - 28 = 15	55	0	0	0
60 - 70	28 - 16 = 12	65	10	1	12
70 - 80	16 - 10 = 6	75	20	2	12
80 - 90	10 - 8 = 2	85	30	3	6
90 - 100	8	95	40	4	32
	$\sum f = 80$				$\sum fdm' = -26$

As 
$$\bar{x} = A + \frac{\sum fdm'}{\sum f} \times C$$

$$\begin{aligned} \therefore \bar{x} &= 55 + \frac{(-26)}{80} \times 10 \\ &= 55 - \frac{13}{4} = \frac{220-13}{4} = \frac{207}{4} = 51.75 \end{aligned}$$

$$\Rightarrow \bar{x} = 51.75$$

### 6.6.5 Arithmetic Mean in case of Unequal Class Interval Series:

Sometime the class interval between two classes is not same, for example 10-20, 20-40 etc.

These series are known as unequal class interval series. However, it does not affect the finding of arithmetic mean as there is not precondition of equal class interval in case of arithmetic mean.

So, mean will be calculated in usual manner.

**Example 10.** Calculate  $\bar{x}$  if the data is given below:

C. I.	4 – 8	8 – 20	20 – 28	28 – 44	44 – 68	68 – 80
f	3	8	12	21	10	6

**Solution:**

C. I.	f	Mid Value m	dm = m – A A = 26	fdm
4 – 8	3	6	–20	–60
8 – 20	8	14	–12	–96
20 – 28	12	24	–2	–24
28 – 44	21	36	+10	210
44 – 68	10	56	+30	300
68 – 80	6	74	+48	288
	$\sum f = 60$			$\sum fdm = 618$

$$\text{As } \bar{x} = A + \frac{\sum fdm}{\sum f}$$

$$\Rightarrow \bar{x} = 26 + \frac{618}{60} = 26 + 10.3 = 36.3$$

$$\Rightarrow \bar{x} = 36.3$$

### 6.6.6 Combined Arithmetic Mean:

Sometime we have the knowledge of mean of two or more series separately but we are interested in finding the mean that will be obtained by taking all these series as one series, such mean is called combined mean. It can be calculated using the following formula.

$$\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

Where  $N_1$  = Number of items in first series,  $N_2$  = Number of items in second series

$\bar{X}_1$  = Mean of first series, and  $\bar{X}_2$  = Mean of second series

**Example 11. Find the combined mean for the following data**

	Firm A	Firm B
No. of Wage Workers	586	648
Average Monthly Wage (Rs.)	52.5	47.5

**Solution:** Combined mean wage of all the workers in the two firms will be

$$\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

Where  $N_1$  = Number of workers in Firm A

$N_2$  = Number of workers in Firm B

$\bar{X}_1$  = Mean wage of workers in Firm A

and  $\bar{X}_2$  = Mean wage of workers in Firm B

We are given that

$$N_1 = 586 \quad N_2 = 648$$

$$\bar{X}_1 = 52.5 \quad \bar{X}_2 = 47.5$$

$\therefore$  Combined Mean,  $\bar{X}_{12}$

$$= \frac{(586 \times 52.5) + (648 \times 47.5)}{586 + 648} = \frac{61,545}{1234} = \text{Rs. } 49.9$$

### 6.6.7 Correcting Incorrect Mean

Many a time it happens that we take some wrong items in the data or overlook some item. This results in wrong calculation of Mean. Later we find the correct values and we want to find out correct mean. This can be done using the following steps:

1. Multiply the incorrect mean of the data (incorrect  $\bar{x}$ ) with number of items to find out incorrect  $\sum \bar{x}$ .
2. Now subtract all the wrong observation from the above values and add the correct observation to the above value to find out correct  $\sum \bar{x}$ .
3. Now divide the correct  $\sum \bar{x}$  with the number of observations to find correct mean.

**Example 12.** Mean wage of 100 workers per day found to be 75. But later on, it was found that the wages of two laborer's Rs. 98 and Rs. 69 were misread as Rs. 89 and Rs. 96. Find out the correct mean wage.

**Solution:** We know that, Correct  $\sum X =$  Incorrect  $\sum X -$  (Incorrect items) + (Correct Items)

Also  $\bar{X} = \frac{\sum X}{N}$

$$\Rightarrow \text{Incorrect } \sum X = 100 \times 75 = 7500$$

$$\therefore \text{Correct } \sum X = 7500 - (89 + 96) + (98 + 69) = 7482$$

$$\Rightarrow \text{Correct } \bar{X} = \frac{\text{Correct } \sum X}{N} = \frac{7482}{100} = 74.82$$

### Determination of Missing Frequency

**Example 13.** Find the missing frequencies of the following series, if  $\bar{X} = 33$  and  $N = 100$

<b>X</b>	<b>5</b>	<b>15</b>	<b>25</b>	<b>35</b>	<b>45</b>	<b>55</b>
<b>f</b>	<b>5</b>	<b>10</b>	<b>?</b>	<b>30</b>	<b>?</b>	<b>10</b>

**Solution:** Let the missing frequencies corresponding to  $X = 25$  and  $X = 45$  be ' $f_1$ ' and ' $f_2$ ' respectively.

X	f	fX
5	5	25
15	10	150
25	$f_1$	$25f_1$
35	30	1050
45	$f_2$	$45f_2$
55	10	550
	$\sum f = 55 + f_1 + f_2$	$\sum fX = 1775 + 25f_1 + 45f_2$

Now,  $N = 100$  (Given)

$$\therefore 55 + f_1 + f_2 = 100$$

$$\Rightarrow f_1 + f_2 = 45 \quad \dots(i)$$

Also  $\bar{X} = \frac{\sum fX}{N}$

$$\Rightarrow 33 = \frac{1775 + 25f_1 + 45f_2}{100}$$

$$\Rightarrow 3300 = 1775 + 25f_1 + 45f_2$$

$$\Rightarrow 25f_1 + 45f_2 = 1525 \quad \dots(ii)$$

Solving (i) and (ii), we get

$$25 \times (f_1 + f_2 = 45) \quad \Rightarrow 25f_1 + 25f_2 = 1125$$

$$1 \times (25f_1 + 45f_2 = 1525) \quad \Rightarrow 25f_1 + 45f_2 = 1525$$

$$\begin{array}{r} (-) \quad (-) \quad (-) \\ \hline \end{array}$$

$$-20f_2 = -400$$

$$f_2 = \frac{400}{20} = 20$$

$$\therefore f_2 = 20$$

Put  $f_2 = 20$  in (i)

$$f_1 + 20 = 45$$

$$\Rightarrow f_1 = 45 - 20 = 25$$

$$\therefore f_1 = 25$$

∴  $f_1 = 25, f_2 = 20$

### **6.6.8 Properties of Arithmetic mean**

1. If we take the deviations of the observations from its Arithmetic mean and then sum up such deviations, then sum of such deviations will always be zero.
2. If we take the square of the deviations of items from its Arithmetic mean and then sum up sum of squares, the value obtained will always be less than the square of deviation taken from any other values.
3. If we have separate mean of two series, we can find the combined mean of the series.
4. If the value of all items in that data is increased or decreased by some constant value say 'k', then the Arithmetic mean is also increased or decreased by same 'k'. In other words, if k is added to the items, then actual mean will be calculated by deducting that k from the mean calculated.
5. If value of all items in the series is divided or multiplied by some constant 'k' then the mean is also multiplied or divided by the same constant 'k'. In other words, if we multiply all observations by 'k' then actual mean can be calculated by dividing the mean to obtained by the constant 'k'.

### **6.6.9 Merits of Arithmetic Mean**

1. Arithmetic mean is very simple to calculate and it is also easy to understand.
2. It is most popular method of calculating the average.
3. Arithmetic mean is rigidly defined means it has a particular formula for calculating the mean.
4. Arithmetic mean is comparatively less affected by fluctuation in the sample.
5. It is most useful average for making comparison.
6. We can perform further treatment on Arithmetic mean.
7. We need not to have grouping of items for calculating Arithmetic mean.
8. Arithmetic mean is based on all the values of the data.

### **6.6.10 Limitations of Arithmetic Mean**

1. The biggest limitation of Arithmetic mean is that it is being affected by extreme values.
2. If we have open end series, it is difficult to measure Arithmetic mean.
3. In case of qualitative data, it is not possible to calculate Arithmetic mean.

4. Sometime it gives absurd result like we say that there are 20 students in one class and 23 students in other class then average number of students in a class is 21.5, which is not possible because student cannot be in fraction.
5. It gives more importance to large value items than small value items.
6. Mean cannot be calculated with the help of a graph.
7. It cannot be located by just inspections of the items.

### CHECK YOUR PROGRESS (B)

1. From the following data, find the average sale per shop.

Sales in '000 (units)	10-12	13-15	16-18	19-21	22-24	25-27	28-30
No. shops	34	50	85	60	30	15	7

2. For the following data (Cumulative Series), find the average income.

Income Below in (Rs.)	30	40	50	60	70	80	90
No. of persons	16	36	61	76	87	95	100

3. Calculate the average marks for the following cumulative frequency distribution.

Marks Above	0	10	20	30	40	50	60	70	80	90
No of students	80	77	72	65	55	43	28	16	10	8

4. For a group of 50 male workers, their average monthly wage Rs.6300 and for a group of 40 female workers this average is Rs.5400. Find the average monthly wage for the combined group of all the workers.
5. The average marks of 100 students is given to be 45. But later on, it was found that the marks of students getting 64 was misread as 46. Find the correct mean.
6. Find missing frequency when mean is 35 and number is 68.

X:            0-10   10-20   20-30   30-40   40-50   50-60

F:            4        10        12        ?        20        ?

7. The mean age of combined group of men and women is 30 years. The mean age of group of men is 32 years and women is 27 years. Find the percentage of men and women in the group

## Answers

1) 17.8 (in 000 units)	4) 5900	7) Men 60%
2) 48	5) 45.18	
3) 51.75	6) 10,12	

### 6.7 SUM UP

Average is the single value that represent its series. Average is also known as Central Tendency. Five types of average Arithmetic Mean, Geometric Mean, Harmonic Mean, Median, Mode. Average or Central Tendency is one such technique that is widely used in statistics. It can be calculated with different ways likewise short cut method, direct method, step deviation method. One can also able to determine combined average for more than one series. Incorrect mean can be corrected with the help of correct arithmetic mean formula.

### 6.8 QUESTIONS FOR PRACTICE

- Q1. What is central tendency? What are uses of measuring central tendency.
- Q2. What are the objectives and functions of average?
- Q3. Explain the features of good average
- Q4. What is average? Give uses and limitations of average.
- Q5. What is arithmetic mean? How it is calculated.
- Q6. Give properties, advantages and limitations of Arithmetic mean.
- Q7. How you can calculate combined arithmetic mean.

### 6.9 SUGGESTED READINGS

- J. K. Sharma, Business Statistics, Pearson Education.
- S.C. Gupta, Fundamentals of Statistics, Himalaya Publishing House.
- S.P. Gupta and Archana Gupta, Elementary Statistics, Sultan Chand and Sons, New Delhi.
- Richard Levin and David S. Rubin, Statistics for Management, Prentice Hall of India, New Delhi.



**COURSE NAME: STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY**

**SARM 1: INTRODUCTION TO STATISTICS**

**SEMESTER I**

**UNIT 7: MEDIAN (DIRECT, SHORT CUT AND STEP DEVIATION METHODS) AND**

**MODE: INSPECTION AND GROUPING METHOD**

**STRUCTURE**

**7.0 Learning Objectives**

**7.1 Introduction**

**7.2 Median for different series**

**7.2.1 Median in Individual Series**

**7.2.2 Median in Discrete Series**

**7.2.3 Median in Continuous Series**

**7.2.4 Merits of Median**

**7.2.5 Limitations of Median**

**7.3 Mode**

**7.3.1 Mode in Individual Series**

**7.3.2 Mode in Discrete Series**

**7.3.3 Mode in Continuous Series**

**7.3.4 Merits of Mode**

**7.3.5 Limitations of Mode**

**7.4 Relation between Mean, Median and Mode**

**7.5 Sum Up**

**7.6 Questions for Practice**

**7.7 Suggested Readings**

**7.0 LEARNING OBJECTIVES**

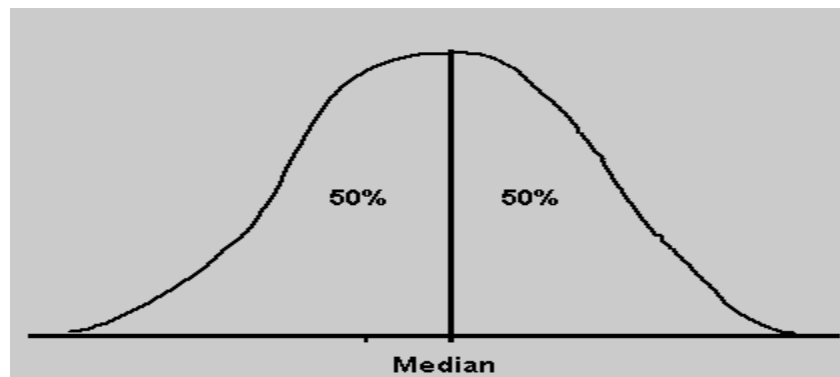
After studying the Unit, students will be able to:

- Define the meaning of median
- Calculate medium for different series

- Define the meaning of mode
- Know merits and limitations of Mean and Median
- Relationship between Mean, Mode and Median
- Meaning of Quartile, Decile and Percentile

## 7.1 INTRODUCTION MEDIAN

Median is the positional measure of Central tendency. It means the median does not depend upon the value of the item under the observation, rather it depends on the position of the item in the series. Median is a value that divide the series exactly in two equal parts, it means 50% of the observation lies below the median and 50% of the observations lies above the median. However, it is important to arrange the series either in ascending order or in descending order before calculation of Median. If series is not arranged, then Median cannot be calculated



For calculating Median

1. Series should be in ascending or descending order.
2. Series should be exclusive, not inclusive.

### 7.2.1 Median in case of Individual series

For calculating the median in individual series, following are the steps:

1. Arrange the series in ascending or descending order
2. Calculate the number of observations. It is denoted by N
3. Calculate the  $\left(\frac{N+1}{2}\right)^{\text{th}}$  term
4. Corresponding value to this item is the median of the data

5. In case there are even number of items in the series, this value will be in fraction. In that case take the arithmetic mean of the adjacent items in which Median is falling. For example, if it is 4.5 than take arithmetic mean of 4<sup>th</sup> item and 5<sup>th</sup> item

$$\text{Median} = \text{value of } \left(\frac{N+1}{2}\right)^{\text{th}} \text{ term}$$

When the number of observations N is odd

**Example 1. Calculation median from the following observations:**

**15, 17, 19, 22, 18, 47, 25, 35, 21**

**Solution:** Arranging the given items in ascending order, we get

15, 17, 18, 19, 21, 22, 25, 35, 47

Now Median,  $M = \text{Size of } \left(\frac{N+1}{2}\right)^{\text{th}} \text{ item}$

$$M = \text{Size of } \left(\frac{9+1}{2}\right)^{\text{th}} \text{ item}$$

$$= \text{Size of } 5^{\text{th}} \text{ item}$$

$$= 21$$

$$\Rightarrow M = 21$$

When the number of observations N is even

**Example 2. Find median from the following data**

**28, 26, 24, 21, 23, 20, 19, 30**

**Solution:** Arranging the given figures in ascending order, we get

19, 20, 21, 23, 24, 26, 28, 30

Now Median,  $M = \text{Size of } \left(\frac{N+1}{2}\right)^{\text{th}} \text{ item}$

$$M = \text{Size of } \left(\frac{8+1}{2}\right)^{\text{th}} \text{ item}$$

$$= \text{Size of } 4.5^{\text{th}} \text{ item}$$

$$= \frac{4^{\text{th}} \text{ item} + 5^{\text{th}} \text{ item}}{2} = \frac{23+24}{2} = \frac{47}{2} = 23.5$$

$$\Rightarrow M = 23.5$$

### 7.2.2 Median in case of Discrete series

Following are the steps in case of discrete series:

1. Arrange the data in ascending or descending order.
2. Find the cumulative frequency of the series.
3. Find the  $\left(\frac{N+1}{2}\right)^{\text{th}}$  term
4. Now look at this term in the cumulative frequency of the series.
5. Value against which such cumulative frequency falls is the median value.

$$\text{Median} = \text{value of } \left(\frac{N+1}{2}\right)^{\text{th}} \text{ term}$$

**Example 3.** Calculate the value of median, if the data is as given below:

<b>Height (in cms.)</b>	<b>110</b>	<b>125</b>	<b>250</b>	<b>200</b>	<b>150</b>	<b>180</b>
<b>No. of Students</b>	<b>8</b>	<b>12</b>	<b>3</b>	<b>10</b>	<b>13</b>	<b>15</b>

**Solution:** Arranging the given data in ascending order, we get

Height (in cms.)	No. of Students f	Cumulative Frequency C · f
110	8	8 (1 – 8)
125	12	20 (9 – 20)
150	13	33 (21 – 33)
180	15	48 (34 – 48)
200	10	58 (49 – 58)
250	3	61 (59 – 61)
	$\sum f = N = 61$	

Now Median, M = Size of  $\left(\frac{N+1}{2}\right)^{\text{th}}$  item

$$\begin{aligned} M &= \text{Size of } \left(\frac{6+1}{2}\right)^{\text{th}} \text{ item} \\ &= \text{Size of } 31^{\text{st}} \text{ item} \\ &= 150 \end{aligned}$$

⇒ Median, M = 150 cms.

### 7.2.3 Median in case of Continuous Series

Following are the steps in case of continuous series:

1. Arrange the data in ascending or descending order.
2. Find the cumulative frequency of the series.
3. Find the  $\left(\frac{N}{2}\right)^{\text{th}}$  term
4. Now look at this term in the cumulative frequency of the series. The value equal to or higher than term calculated in third step is the median class.
5. Find median using following formula.
6.  $M = L + \frac{\frac{N}{2} - C.f}{f} \times i$
7. Where M = Median
8. L = Lower Limit of Median Class
9. N = Number of Observations.
10. c.f. = Cumulative frequency of the Median Class.
11. f = Frequency of the class preceding Median Class.
12. i = Class interval of Median Class

**Example 4. Calculate Median**

Marks	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35
No. of Students	8	7	14	16	9	6

**Solution:**

C. I.	No. of Students f	Cumulative Frequency C · f
5 – 10	8	8 (1 – 8)
10 – 15	7	15 (9 – 15)
15 – 20	14	29 (16 – 29)
20 – 25	16	45 (30 – 45)
25 – 30	9	54 (46 – 54)
30 – 35	6	60 (55 – 60)
	$\sum f = N = 60$	

Median, M = Size of  $\left(\frac{N}{2}\right)^{\text{th}}$  item

$$M = \text{Size of } \left(\frac{60}{2}\right)^{\text{th}} \text{ item}$$

$$= \text{Size of } 30^{\text{th}} \text{ item}$$

⇒ Median lies in the class interval 20 – 25

As Median,  $M = L + \frac{\frac{N}{2} - C \cdot f}{f} \times i$

Here L = Lower limit of the median class = 20

$N = 60, \quad C \cdot f = 29, \quad f = 16$

i = Class – length of the median class = 5

∴  $M = 20 + \frac{(30-29)}{16} \times 5$   
 $= 20 + \frac{5}{16} = 20 + 9.312 = 29.312$

⇒ M = 29.312

**Inclusive Series** – It must be converted to Exclusive Series before calculation of the Median.

**Example 5. Find Median from the given data**

<b>X</b>	<b>10 – 19</b>	<b>20 – 29</b>	<b>30 – 39</b>	<b>40 – 49</b>	<b>50 – 59</b>	<b>60 – 69</b>	<b>70 – 79</b>	<b>80 – 89</b>
<b>f</b>	<b>6</b>	<b>53</b>	<b>85</b>	<b>56</b>	<b>21</b>	<b>16</b>	<b>4</b>	<b>4</b>

**Solution:** Converting the given data into exclusive form, we get

$$\left[ \text{Correction factor} = \frac{L_2 - U_1}{2} = \frac{20 - 19}{2} = \frac{1}{2} = 0.5 \right]$$

(0.5 is subtracted from all lower limits and added to all upper limits)

X	f	Cumulative frequency C · f
9.5 – 19.5	6	6 (1 – 6)
19.5 – 29.5	53	59 (7 – 59)
29.5 – 39.5	85	144 (60 – 144)
39.5 – 49.5	56	200 (145 – 200)
49.5 – 59.5	21	221 (201 – 221)
59.5 – 69.5	16	237 (222 – 237)
69.5 – 79.5	4	241 (238 – 241)
79.5 – 89.5	4	245 (242 – 245)
	$\sum f = N = 245$	

Median, M = Size of  $\left(\frac{N}{2}\right)^{\text{th}}$  item

$$M = \text{Size of } \left(\frac{245}{2}\right)^{\text{th}} \text{ item}$$

$$= \text{Size of } 122.5^{\text{th}} \text{ item}$$

∴ The real class limits of the median class = (29.5 – 39.5)

So  $M = L + \frac{\left(\frac{N}{2} - C.f\right)}{f} \times i$

$$\Rightarrow M = 29.5 + \left(\frac{122.5 - 59}{85}\right) \times 10$$

$$= 29.5 + \left(\frac{63.5}{85} \times 10\right)$$

$$= 29.5 + \left(\frac{635}{85}\right)$$

$$= 29.5 + 7.47 = 36.97$$

$$\Rightarrow M = 36.97$$

### Cumulative Series (More than and less than)

**Example 6. Find median, if the data is as given below:**

<b>Marks More than</b>	<b>20</b>	<b>35</b>	<b>50</b>	<b>65</b>	<b>80</b>	<b>95</b>
<b>No. of Students</b>	<b>100</b>	<b>94</b>	<b>74</b>	<b>30</b>	<b>4</b>	<b>1</b>

**Solution:** Converting the given data into class – interval form, we get

Marks C. I.	Frequency f	Cumulative Frequency C · f
20 – 35	100 – 94 = 6	6 (1 – 6)
35 – 50	94 – 74 = 20	26 (7 – 26)
50 – 65	74 – 30 = 44	70 (27 – 70)
65 – 80	30 – 4 = 26	96 (71 – 96)
80 – 95	4 – 1 = 3	99 (97 – 99)
95 – 110	1	100 (100)
	$\sum f = N = 100$	

Now Median,  $M = \text{Size of } \left(\frac{N}{2}\right)^{\text{th}} \text{ item}$

$$M = \text{Size of } \left(\frac{100}{2}\right)^{\text{th}} \text{ item}$$

$$= \text{Size of } 50^{\text{th}} \text{ item}$$

⇒ Median lies in the class interval = 50 – 65

So 
$$M = L + \frac{\left(\frac{N}{2} - C.f\right)}{f} \times i$$

⇒ 
$$M = 50 + \left(\frac{50-26}{44}\right) \times 15$$
$$= 50 + \left(\frac{24}{44} \times 15\right)$$
$$= 50 + 8.18 = 58.18$$

⇒ 
$$M = 58.18$$

**Example 7. Find median, if the data is as given below:**

Marks Less than	10	20	30	40	50	60	70	80
No. of Students	20	30	50	94	96	127	198	250

**Solution:** Converting the given data into class interval form, we get

Marks C. I.	No. of Students f	Cumulative Frequency C · f
0 – 10	20	20 (1 – 20)
10 – 20	30 – 20 = 10	30 (21 – 30)
20 – 30	50 – 30 = 20	50 (31 – 50)
30 – 40	94 – 50 = 44	94 (51 – 94)
40 – 50	96 – 94 = 2	96 (95 – 96)
50 – 60	127 – 96 = 31	127 (97 – 127)
60 – 70	198 – 127 = 71	198 (128 – 198)
70 – 80	250 – 198 = 52	250 (199 – 250)
	$\sum f = N = 250$	

Now Median,  $M =$  Size of  $\left(\frac{N}{2}\right)^{\text{th}}$  item

$$M = \text{Size of } \left(\frac{250}{2}\right)^{\text{th}} \text{ item}$$
$$= \text{Size of } 125^{\text{th}} \text{ item}$$

⇒ Median lies are the class – interval = 50 – 60

So 
$$M = L + \frac{\frac{N}{2} - C.f}{f} \times i$$

⇒ 
$$M = 50 + \left(\frac{125-96}{31}\right) \times 10$$
$$= 50 + \left(\frac{29}{31} \times 10\right)$$



$$= 50 + \frac{290}{31} = 50 + 9.35 = 59.35$$

$$\Rightarrow M = 59.35$$

### Mid – Value Series

**Example 8. Find the value of median for the following data:**

Mid Value	15	25	35	45	55	65	75	85	95
f	8	26	45	72	116	60	38	22	13

**Solution:** It is clear from the mid – value that the class size is 10. For finding the limits of different classes, apply the formula:

$$L = m - \frac{i}{2} \quad \text{and} \quad U = m + \frac{i}{2}$$

Where, L and U denote the lower and upper limits of different classes, ‘m’ denotes the mid – value of the corresponding class interval and ‘i’ denotes the difference between mid values.

∴ Corresponding to mid – value ‘15’, we have

$$L = 15 - \frac{10}{2} \quad \text{and} \quad U = 15 + \frac{10}{2}$$

i. e. C. I. = 10 – 20

Similarly other class intervals can be located

Mid Value	f	C. I.	Cumulative Frequency C · f
15	8	10 – 20	8 (1 – 8)
25	26	20 – 30	34 (9 – 34)
35	45	30 – 40	79 (35 – 79)
45	72	40 – 50	151 (80 – 151)
55	116	50 – 60	267 (152 – 267)
65	60	60 – 70	327 (268 – 327)
75	38	70 – 80	365 (328 – 365)
85	22	80 – 90	387 (366 – 387)
95	13	90 – 100	400 (388 – 400)
	N = 100		

Now Median,  $M = \text{Size of } \left(\frac{N}{2}\right)^{\text{th}}$  item

$$M = \text{Size of } \left(\frac{400}{2}\right)^{\text{th}} \text{ item}$$

$$= \text{Size of } 200^{\text{th}} \text{ item}$$

⇒ Median lies in the class – interval = 50 – 60

$$\begin{aligned} \text{So } M &= L + \frac{\frac{N}{2} - C \cdot f}{f} \times i \\ \Rightarrow M &= 50 + \left( \frac{200 - 151}{116} \right) \times 10 \\ &= 50 + \left( \frac{49}{116} \times 10 \right) \\ &= 50 + \frac{490}{116} \\ &= 50 + 4.224 = 54.224 \\ \Rightarrow M &= 54.224 \end{aligned}$$

### Determination of Missing Frequency

**Example 9.** Find the missing frequency in the following distribution if  $N = 72$ ,  $Q_1 = 25$  and  $Q_3 = 50$

C.I.	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80
f	4	8	–	19	–	10	5	–

**Solution:** Let the missing frequencies be  $f_1$ ,  $f_2$  and  $f_3$  respectively.

C.I.	f	Cumulative Frequency C · f
0 – 10	4	4
10 – 20	8	12
20 – 30	$f_1$	$12 + f_1$
30 – 40	19	$31 + f_1$
40 – 50	$f_2$	$31 + f_1 + f_2$
50 – 60	10	$41 + f_1 + f_2$
60 – 70	5	$46 + f_1 + f_2$
70 – 80	$f_3$	$46 + f_1 + f_2 + f_3$
Now N =	$N = 72 = \sum f$	72
= $\sum f$	$\sum f = 46 + f_1 + f_2 + f_3$	

$$= 46 + f_1 + f_2 + f_3$$

$$\Rightarrow f_1 + f_2 + f_3 = 72 - 46 = 26$$

$$\Rightarrow f_1 + f_2 + f_3 = 26 \quad \dots(i)$$

Also,  $Q_1 = 25$  (Given)

$\Rightarrow Q_1$  lies in the class – interval 20 – 30

$$\Rightarrow Q_1 = L + \frac{\frac{N}{4} - C \cdot f}{f} \times i$$

$$25 = 20 + \frac{\frac{72}{4} - 12}{f_1} \times 10$$

$$25 = 20 + \frac{18-12}{f_1} \times 10$$

$$25 - 20 = \frac{6}{f_1} \times 10$$

$$5f_1 = 60$$

$$f_1 = \frac{60}{5}$$

$$\Rightarrow f_1 = 12 \quad \dots(\text{ii})$$

Similarly, we are given that

$$Q_3 = 50$$

$\Rightarrow Q_3$  lies in the class – interval 50 – 60

$$\Rightarrow Q_3 = L + \frac{\frac{3N}{4} - C \cdot f}{f} \times i$$

$$50 = 50 + \frac{\frac{3 \times 72}{4} - (31 + f_1 + f_2)}{10} \times 10$$

$$50 = 50 + \frac{54 - (31 + 12 + f_2)}{1}$$

( $\because f_1 = 12$  By (ii))

$$50 - 50 = 54 - (43 + f_2)$$

$$0 = 54 - (43 + f_2)$$

$$43 + f_2 = 54$$

$$f_2 = 54 - 43$$

$$\Rightarrow f_2 = 11 \quad \dots(\text{iii})$$

Putting (ii) and (iii) in (i), we get

$$f_1 + f_2 + f_3 = 26$$

$$12 + 11 + f_3 = 26$$

$$23 + f_3 = 26$$

$$f_3 = 26 - 23$$

$$\Rightarrow f_3 = 3$$

#### 7.2.4 Merits of Median

1. Median is easy to calculate.
2. It is capable of Graphic presentation.
3. It is possible even in case of open-ended series.
4. This is rigidly defined.
5. It is not affected by extreme values.
6. In case of qualitative data, it is very useful.

#### 7.11.5 Limitations of Median

1. It is not capable of further algebraic treatment.

2. It is positional average and is not based on all observation.
3. It is very much affected by fluctuation in sampling.
4. Median needs arrangement of data before calculation.
5. In case of continuous series, it assumes that values are equally distributed in a particular class.

### TEST YOUR PROGRESS (A)

1. Calculate Median

30, 45, 75, 65, 50, 52, 28, 40, 49, 35, 52,

2. Calculate Median

36, 32, 28, 22, 26, 20, 18, 40,

3. Find Median

Wages:	100	150	80	200	250	180
No. of workers	24	26	16	20	6	30

4. Calculate Median

X;	0-5	5-10	10-15	15-20	20-25	25-30	30-35
F:	4	6	10	16	12	8	4

5. Calculate Median:

X;	10-19	20-29	30-39	40-49	50-59	60-69
F:	4	8	12	16	10	6

6. Find Median:

Income	100-200	200-400	400-700	700-1200	1200-2000
Number of firms	40	100	260	80	20

7. Find missing frequency when median is 50 and number is 100.

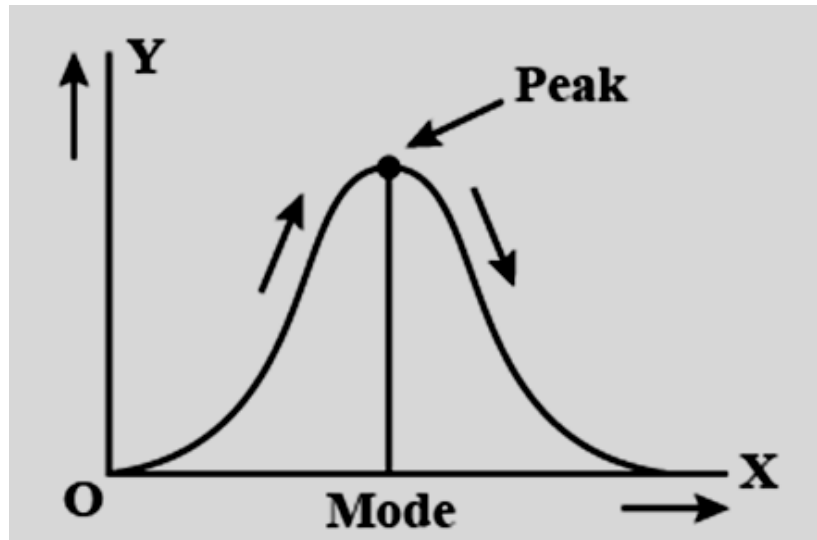
X;	0-20	20-40	40-60	60-80	80-100
F:	14	?	27	?	15

### Answers

1) 49	4) 18.125	7) 23,21
2) 27	5) 42	8) 41, 53, 47, 41, 51.08
3) 150	6) 526.9	9) $M = 46.4, Q_1 = 34.2, Q_3 = 57.5, D_6 = 50$

### 7.3 MODE

Mode is another positional measure of Central Tendency. Mode is the value that is repeated most number of time in the series. In other words, the value having highest frequency is called Mode. The term 'Mode' is taken from French word 'La Mode' which means the most fashionable item. So, Mode is the most popular item of the series.



#### For calculating Mode

1. Series should be in ascending or descending order.
2. Series should be exclusive, not inclusive.

Series should have equal class intervals.

#### 7.3.1 Mode in Individual Series

In case of Individual series, following are the steps of finding the Mode.

1. Arrange the series either in ascending order or descending order.
2. Find the most repeated item.
3. This item is Mode.

**Example 1. Calculate mode from the following data of marks obtained by 10 students**

<b>S. No.</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>	<b>10</b>
<b>Marks obtained</b>	<b>10</b>	<b>27</b>	<b>24</b>	<b>12</b>	<b>27</b>	<b>27</b>	<b>20</b>	<b>18</b>	<b>15</b>	<b>30</b>

**Solution:** By Inspection

It can be observed that 27 occur most frequently i. e. 3 times. Hence, mode = 27 marks

By converting into discrete series

Marks Obtained	Frequency
10	1

12	1
15	1
18	1
20	1
24	1
27	3
30	1
	N = 10

Since, the frequency of 27 is maximum i. e. 3

It implies the item 27 occurs the maximum number of times. Hence the modal marks are 27.

$$\text{Mode} = 27$$

### 7.3.2 Mode in discrete series

In case of discrete series, we can find mode by two methods that are Observation Method and Grouping Method.

**1. Observation Method:** Under this method value with highest frequency is taken as mode.

**2. Grouping Method:** Following are the steps of Grouping method:

- Prepare a table and put all the values in the table in ascending order.
- Put all the frequencies in first column. Mark the highest frequency.
- In second column put the total of frequencies taking two frequencies at a time like first two, then next two and so on. Mark the highest total.
- In third column put the total of frequencies taking two frequencies at a time but leaving the first frequency like second and third, third and fourth and so on. Mark the highest total.
- In fourth column put the total of frequencies taking three frequencies at a time like first three, then next three and so on. Mark the highest total.
- In fifth column put the total of frequencies taking three frequencies at a time but leaving the first frequency like second, third and fourth; than fifth, sixth and seventh and so on. Mark the highest total.
- In sixth column put the total of frequencies again taking three frequencies at a time but leaving the first two frequencies. Mark the highest total.
- Value that is marked highest number of times is the mode.

**Example 2. Find the modal value for the following distribution**

Age (in years)	8	9	10	11	12	13	14	15
No. of Persons	5	6	8	7	9	8	9	6

**Solution:** Here, as maximum frequency 9 belongs to two age values 12 and 14, so its not possible to determine mode by inspection. We will have to determine the modal value through grouping and analysis table.

Grouping Table						
Age (in years)	Frequency					
	G <sub>1</sub>	G <sub>2</sub>	G <sub>3</sub>	G <sub>4</sub>	G <sub>5</sub>	G <sub>6</sub>
8	5	11	14	19	21	24
9	6					
10	8	15	16	24	26	23
11	7					
12	9	17	17	24	26	23
13	8					
14	9	15	17	24	26	23
15	6					

Analysis Table								
Group No.	8	9	10	11	12	13	14	15
G <sub>1</sub>					×		×	
G <sub>2</sub>					×	×		
G <sub>3</sub>						×	×	
G <sub>4</sub>				×	×	×		
G <sub>5</sub>					×	×	×	
G <sub>6</sub>			×	×	×			
Total	×	×	1	2	5	4	3	×

Since, 12 occurs maximum number of times i. e. 5 times, the modal age is 12 years

$$\text{Mode} = 12$$

### 7.3.3 Mode in Continuous series

In case of continuous series, we can find mode by two methods that are Observation Method and Grouping Method.

- 1. Observation Method:** Under this method value with highest frequency is taken as mode class than the mode formula is applied which is given below.
- 2. Grouping Method:** Following are the steps of Grouping method:

- Prepare a table and put all the classes of data in the table in ascending order.
- Put all the frequencies in first column. Mark the highest frequency.
- In second column put the total of frequencies taking two frequencies at a time like first two, then next two and so on. Mark the highest total.
- In third column put the total of frequencies taking two frequencies at a time but leaving the first frequency like second and third, third and fourth and so on. Mark the highest total.
- In fourth column put the total of frequencies taking three frequencies at a time like first three, then next three and so on. Mark the highest total.
- In fifth column put the total of frequencies taking three frequencies at a time but leaving the first frequency like second, third and fourth; than fifth, sixth and seventh and so on. Mark the highest total.
- In sixth column put the total of frequencies again taking three frequencies at a time but leaving the first two frequencies. Mark the highest total.
- Class that is marked highest number of times is the mode class.
- Apply following formula for calculating the mode:

$$Z = L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times i$$

Where,

Z = Mode, L = Lower limit of the mode class

$f_m$  = Frequency of mode class,  $f_1$  = Frequency of class preceding mode class

$f_2$  = Frequency of class succeeding mode class,  $i$  = Class interval

**Example 3. Find the mode for the following frequency distribution**

Age (in years)	30 – 35	35 – 40	40 – 45	45 – 50	50 – 55	55 – 60
No. of Persons	3	8	12	20	15	2

**Solution:** Here, the maximum frequency is corresponding to the class – interval 45 – 50.

So, the modal class is 45 – 50.

Now, the mode is given by the formula

$$\text{Mode, } Z = L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times i$$

Here L = Lower limit of modal class = 45

$f_m$  = Frequency of modal class = 20

$f_1$  = Frequency of class preceding the modal class = 12

$f_2$  = Frequency of class succeeding the modal class = 15

$i$  = Class length of modal class = 5



$$\begin{aligned}
\therefore \text{Mode, } Z &= 45 + \frac{20-12}{(2 \times 20) - 12 - 15} \times 5 \\
&= 45 + \frac{8}{40-27} \times 5 \\
&= 45 + 3.07 \\
&= 48.1 \text{ years (approx.)} \\
\Rightarrow Z &= 48.1 \text{ year}
\end{aligned}$$

**Example 4. Calculate mode from the following data**

C. I.	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
f	2	9	10	13	11	6	13	7	4	1

**Solution:** Here as it is not possible to find modal class by inspection, so we have to determine it through grouping and analysis table.

Grouping Table						
C. I.	Frequency					
	G <sub>1</sub>	G <sub>2</sub>	G <sub>3</sub>	G <sub>4</sub>	G <sub>5</sub>	G <sub>6</sub>
0 – 10	2	11	19	21	32	34
10 – 20	9					
20 – 30	10	23	24	30	30	26
30 – 40	13					
40 – 50	11	17	19	24	12	26
50 – 60	6					
60 – 70	13	20	11	24	12	26
70 – 80	7					
80 – 90	4	5			12	26
90 – 100	1					

Analysis Table										
Group No.	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
G <sub>1</sub>				×			×			
G <sub>2</sub>			×	×						
G <sub>3</sub>				×	×					
G <sub>4</sub>				×	×	×				
G <sub>5</sub>		×	×	×						

$G_6$			×	×	×					
Total	×	1	3	6	3	1	1	×	×	×

Clearly the modal class is 30 – 40

Now the mode is given by the formula

$$\text{Mode, } Z = L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times i$$

Here  $L =$  Lower limit of modal class 30 – 40 = 30

$f_m =$  Frequency corresponding to modal class = 13

$f_1 =$  Frequency of interval preceding modal class

$f_2 =$  Frequency of interval succeeding and

$i =$  Class length of modal class

$$\begin{aligned} \therefore \text{Mode, } Z &= 30 + \frac{13-10}{(2 \times 13) - 10 - 11} \times 10 \\ &= 30 + \frac{3}{26-21} \times 10 \\ &= 30 + \frac{30}{5} \\ &= 30 + 6 \qquad \qquad \qquad = 36 \end{aligned}$$

$$\Rightarrow Z = 36$$

**Example 5. Determine the missing frequencies when it is given that  $N = 230$ , Median,  $M = 233.5$  and Mode,  $Z = 234$**

C.I	200-210	210-220	220-230	230-240	240-250	250-260	260-270
f	4	16	–	–	–	6	4

**Solution:** Let the missing frequencies be  $f_1$ ,  $f_2$  and  $f_3$  respectively.

C. I	f	C · f
200 – 210	4	4
210 – 220	16	20
220 – 230	$f_1$	$20 + f_1$
230 – 240	$f_2$	$20 + f_1 + f_2$
240 – 250	$f_3$	$20 + f_1 + f_2 + f_3$
250 – 260	6	$26 + f_1 + f_2 + f_3$
260 – 270	4	$30 + f_1 + f_2 + f_3$
	$N = 230 = \sum f$	
	$\sum f = 30 + f_1 + f_2 + f_3$	

Now  $N = 230 = \sum f$  (Given)  
 $= 30 + f_1 + f_2 + f_3$

$$\Rightarrow f_1 + f_2 + f_3 = 230 - 30 = 200$$

$$\Rightarrow f_1 + f_2 + f_3 = 200 \quad \dots(i)$$

Also, Median = 233.5 (Given)

$\Rightarrow$  Median class is 230 – 240

$$\Rightarrow M = L + \frac{\frac{N}{2} - C.f}{f} \times i$$

$$233.5 = 230 + \frac{\frac{230}{2} - (20 + f_1)}{f_2} \times 10$$

$$3.5 = \frac{115 - 20 - f_1}{f_2} \times 10$$

$$3.5f_2 = 950 - 10f_1$$

$$\Rightarrow 10f_1 + 3.5f_2 = 950 \quad \dots(ii)$$

Now Mode = 234 lies in 230 – 240

$$\therefore Z = L + \frac{f_2 - f_1}{2f_2 - f_1 - f_3} \times i$$

$$\Rightarrow 234 = 230 + \frac{f_2 - f_1}{2f_2 - f_1 - f_3} \times 10$$

$$\Rightarrow 4 = \frac{f_2 - f_1}{2f_2 - f_1 - (200 - f_1 - f_2)} \times 10 \quad \text{[Using (i)]}$$

$$\Rightarrow 4 = \frac{f_2 - f_1}{2f_2 - f_1 - 200 - f_1 - f_2} \times 10$$

$$\Rightarrow 4 = \frac{(f_2 - f_1) \times 10}{3f_2 - 200}$$

$$\Rightarrow 12f_2 - 800 = 10f_2 - 10f_1$$

$$\Rightarrow 2f_2 - 800 + 10f_1 = 0$$

$$\Rightarrow 10f_1 + 2f_2 = 800 \quad \dots(iii)$$

Solving (ii) and (iii), we get

$$10f_1 + 3.5f_2 = 950$$

$$10f_1 + 2f_2 = 800$$

$$(-) \quad (-) \quad (-)$$

---


$$1.5f_2 = 150$$

$$\Rightarrow f_2 = \frac{150}{1.5} = 100$$

$$f_2 = 100 \quad \dots(iv)$$

Put (iv) in (iii)

$$10f_1 + 2(100) = 800$$

$$\Rightarrow 10f_1 = 800 - 200 = 600$$

$$\Rightarrow 10f_1 = 600$$

$$\Rightarrow f_1 = 60 \quad \dots(v)$$

Put (iv) and (v) in (i)

$$60 + 100 + f_3 = 200$$

$$\Rightarrow f_3 = 40$$

∴ The missing frequencies are 60, 100 and 40.

### 7.3.4 Merits of Median

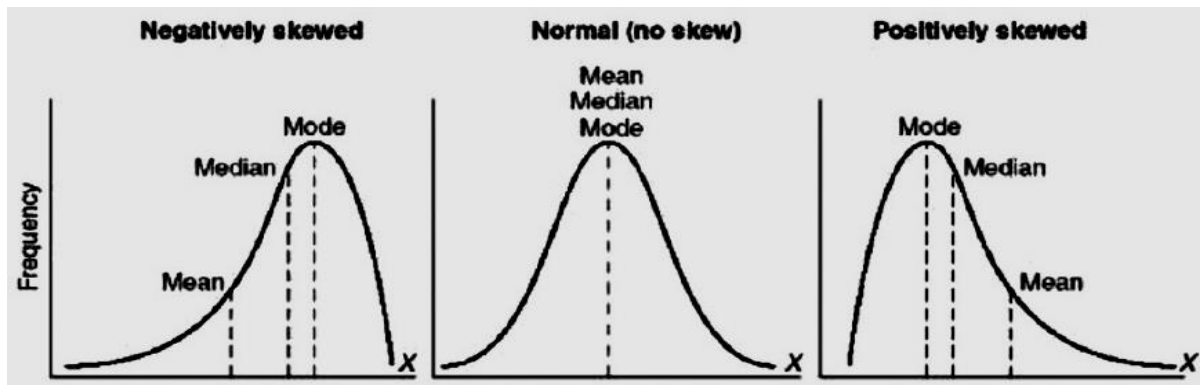
1. Mode is easy to calculate.
2. People can understand this in routine life.
3. It is capable of Graphic presentation.
4. It is possible even in case of open-end series.
5. This is rigidly defined.
6. It is not affected by extreme values.
7. In case of qualitative data, it is very useful.

### 7.3.5 Limitations of Median

1. It is not always determinable as series may be Bi-modal or Tri-modal.
2. It is not capable of further algebraic treatment.
3. It is positional average and is not based on all observation.
4. It is very much affected by fluctuation in sampling.
5. Mode needs arrangement of data before calculation.

## 7.4 RELATION BETWEEN MEAN, MEDIAN AND MODE

In a normal series the value of Mean, Median and Mode is always same. However, Karl Pearson studied the empirical relation between the Mean, Median and Mode and found that in moderately skewed series the Median always lies between the Mean and the Mode. Normally it is two third distance from Mode and one third distance from Mean.



On the basis of this relation following formula emerged

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$$

or  $Z = 3M - 2\bar{X}$

**Example 6. Calculate M when  $\bar{X}$  and Z of a distribution are given to be 35.4 and 32.1 respectively.**

**Solution:** We are given that

$$\text{Mean, } \bar{X} = 35.4$$

$$\text{Mode, } Z = 32.1$$

As we know the empirical relation between Mean, Median and Mode.

i. e.  $\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$

$$\Rightarrow Z = 3M - 2\bar{X}$$

$$\Rightarrow M = \frac{1}{3} (Z + 2\bar{X})$$

$$\begin{aligned} \Rightarrow M &= \frac{1}{3} (32.1 + 2(35.4)) \\ &= \frac{1}{3} (32.1 + 70.8) \\ &= \frac{1}{3} (102.9) = 34.3 \end{aligned}$$

$$\Rightarrow \text{Median, } M = 34.3$$

### CHECK YOUR PROGRESS - B

1. Find Mode:

X: 22, 24, 17, 18, 19, 18, 21, 20, 21, 20, 23, 22, 22, 22

2. Find Mode by inspection method

X	6	12	18	24	30	36	42	48
f	9	11	25	16	9	10	6	3

3. Find Mode by Grouping Method

X	21	22	25	26	27	28	29	30
F	7	10	15	18	13	7	3	2

4. Find Mode by Grouping Method and inspection method

X:	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
F:	2	18	30	45	35	20	6	4

5. Calculate mode using grouping and analysis methods.

X	100-110	110-120	120-130	130-140	140-150	150-160	160-170	170-180
f	4	6	20	32	33	17	8	2

6. Find Mode

X	0-100	100-200	200-400	400-500	500-700
F:	5	15	40	32	28

### Answers

1) 22	3) 26	5) 56.46
2) 18	4) 36	6) 440

### 7.5 LET US SUM UP

- There are mainly five types of average Arithmetic Mean, Geometric Mean, Harmonic Mean, Median, Mode.
- Median divide the series in two equal parts.
- Mode is value repeated most number of time.
- There is a existence of relationship between mean medium and mode.

### 7.6 QUESTIONS FOR PRACTICE

- Q1. What is median? How it is calculated?
- Q2. Give merits and limitations of Median.
- Q3. What is mode? How it is calculated. Give its merits and limitations.
- Q4. Explain grouping method of calculating Mode.
- Q5. Give relation between Mean, Median and Mode.
- Q6. What is positional average. Give various positional average.

### 7.7 SUGGESTED READINGS

- J. K. Sharma, Business Statistics, Pearson Education.
- S.C. Gupta, Fundamentals of Statistics, Himalaya Publishing House.
- S.P. Gupta and Archana Gupta, Elementary Statistics, Sultan Chand and Sons, New Delhi.
- Richard Levin and David S. Rubin, Statistics for Management, Prentice Hall of India, New Delhi.

**COURSE NAME: STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY**

**SARM 1: INTRODUCTION TO STATISTICS**

**SEMESTER I**

**UNIT 8: GEOMETRIC MEAN, HARMONIC MEAN AND OTHER POSITIONAL AVERAGES**

**STRUCTURE**

**8.0 Learning Objectives**

**8.1 Introduction: Geometric Mean**

**8.1.1 Geometric Mean in individual series**

**8.1.2 Geometric Mean in discrete series**

**8.1.3 Geometric Mean in continuous series**

**8.1.4 Merits of Geometric Mean**

**8.1.5 Limitations of Geometric Mean**

**8.2 Harmonic Mean**

**8.2.1 Harmonic Mean in individual series**

**8.2.2 Harmonic Mean in discrete series**

**8.2.3 Harmonic Mean in continuous series**

**8.2.4 Merits of Harmonic Mean**

**8.2.5 Limitations of Harmonic Mean**

**8.3 Other Positional Measures**

**8.3.1 Quartiles**

**8.3.2 Percentiles**

**8.3.3 Deciles**

**8.4 Sum Up**

**8.5 Key Terms**

**8.6 Questions for Practice**

**8.7 Suggested Readings**

**8.0 OBJECTIVES**

After studying the unit, the learner will be able to know:

- Meaning of geometric mean
- Calculation of geometric mean for different series
- Meaning of harmonic mean
- Calculation of harmonic mean for different series
- Merits/ demerits of geometric mean and harmonic mean
- Meaning of Quartile, Percentile and Decile

## 8.1 INTRODUCTION: GEOMETRIC MEAN

Besides the arithmetic mean, median, and mode there are other averages that are relatively unimportant but may be appropriate in particular situations. These are Geometric Mean and Harmonic Mean. Often, we see that all the observations do not have equal importance. In such cases, we need to give differential importance to different items. Here we use weighted means - arithmetic, geometric, or harmonic - instead of simple means. Sometime we deal with such quantities or items that change over a period of time. In that case we are interested in finding the rate of change in the item over the period of time. In other words, we can say that we are interested in finding the rate of growth or rate of decline in the item. For example, we want to know average rate of growth in the population, growth in national income of the country or annual decline rate in the value of machinery etc. In that case, the most appropriate measure of average is the geometric mean. The geometric mean is represented by G.M. we can define the Geometric mean as:

“Geometric mean of N items is root nth of the product of item”

Symbolically we can write Geometric mean as:

$$G. M. = \sqrt[N]{X_1 \times X_2 \times X_3 \times \dots \times X_N}$$

Where G. M. = Geometric Mean

N = Number of items

X<sub>1</sub>, X<sub>2</sub>, X<sub>3</sub> ..... = Various items or observations.

### 8.1.1 Geometric Mean in Individual Series

Following are the steps for calculating Geometric mean in the individual series

1. Take the logarithm of all the values.
2. Find the sum of the values after taking the logarithm.



3. Divide the sum by number of items.
4. Find out antilogarithm of the resultant figure.

$$\text{the Geometric Mean or G. M.} = \text{Antilog} \left( \frac{\sum \log X}{N} \right)$$

**Example 1. Calculation geometric mean for the data given below**

**Solution:**

X	log X	
60	log 60	1.7782
75	log 75	1.8751
90	log 90	1.9542
90	log 90	1.9542
90	log 90	1.9542
N = 5	$\sum \log X = 9.5159$	

As Geometric Mean,  $G = \text{Antilog} \left( \frac{\sum \log X}{N} \right)$

$$G = \text{Antilog} \left( \frac{9.5159}{5} \right)$$

$$= \text{Antilog}(1.9032) = 80$$

⇒ G = 80

### 8.1.2 Geometric Mean in Discrete Series

Following are the steps for calculating Geometric mean in the individual series

1. Take the logarithm of all the values.
2. Multiply the logarithm with the corresponding frequency of the items.
3. Find the sum of the product.
4. Divide the sum with number of items.
5. Find out antilogarithm of the resultant figure.

$$\text{Geometric Mean or G. M} = \text{Antilog} \left( \frac{\sum f \log X}{N} \right)$$

**Example 2. Find G, the geometric mean, for the following data**

<b>X</b>	<b>15</b>	<b>25</b>	<b>35</b>	<b>45</b>	<b>55</b>
<b>f</b>	<b>5</b>	<b>10</b>	<b>15</b>	<b>7</b>	<b>4</b>

**Solution:**

X	f	log X	f log X
---	---	-------	---------

15	5	log 15	1.1761	5.8805
25	10	log 25	1.3979	13.9790
35	15	log 35	1.5441	23.1615
45	7	log 45	1.6532	11.5724
55	4	log 55	1.7404	6.9616
	$\sum f = N = 41$			$\sum f \log X = 61.5550$

As Geometric Mean,  $G = \text{Antilog} \left( \frac{\sum f \log X}{N} \right)$

$$\begin{aligned} \therefore G &= \text{Antilog} \left( \frac{61.5550}{41} \right) \\ &= \text{Antilog}(1.5013) = 31.72 \\ G &= 31.72 \end{aligned}$$

### 8.1.3 Geometric Mean in Discrete Series

Finding G.M. in the continuous series is same as in case of discrete series except that we have to find the mid values of the class intervals. Rest all the steps are same. Following are the steps for calculating Geometric mean in the individual series

1. Find mid values of each class interval.
2. Take the logarithm of all the mid values.
3. Multiply the logarithm with the corresponding frequency of the items.
4. Find the sum of the product.
5. Divide the sum by number of items.
6. Find out antilogarithm of the resultant figure.

$$\text{Geometric Mean or G.M} = \text{Antilog} \left( \frac{\sum f \log X}{N} \right)$$

#### Example 3. Find GM

C.I.	1.5 – 2.5	2.5 – 3.5	3.5 – 4.5	4.5 – 5.5	5.5 – 6.5
f	10	15	7	18	12

**Solution:**

C.I.	f	Mid Value x	log x	f log x
1.5 – 2.5	10	2	0.3010	3.0100
2.5 – 3.5	15	3	0.4771	7.1565
3.5 – 4.5	7	4	0.6021	4.2147
4.5 – 5.5	18	5	0.6990	12.5820

5.5 – 6.5	12	6	0.7782	9.3384
	$\sum f = N = 62$			$\sum f \log x = 36.3016$

As  $G = \text{Antilog} \left( \frac{\sum f \log X}{\sum f} \right)$   
 $G = \text{Antilog} \left( \frac{36.3016}{62} \right)$   
 $= \text{Antilog}(0.5855) = 3.850$   
 $G = 3.850$

#### 8.1.4 Merits of Geometric Mean

1. Geometric mean is rigidly defined.
2. It is very suitable for calculating growth or decline rate.
3. Its calculation is based on all the items under observation.
4. Further mathematical treatment can be applied to it.
5. Like Arithmetic mean shows biases for higher values, Geometric mean shows business for lower values which is useful in many situations like price analysis.
6. It is comparatively less affected by Extreme value.
7. It does not change much with the change in sample.

#### 8.1.5 Demerits of Geometric Mean

1. It is comparatively difficult to calculate.
2. It is also difficult to understand and interpret.
3. It cannot be calculated if negative values are present in the series.
4. Even if a single observation is zero in the series the geometric mean becomes zero.

### 8.2 HARMONIC MEAN

The harmonic mean is an average that is used for finding the average rate we are interested in finding the average speed of the vehicle or we know that three persons take 10, 12 and 14 hours to complete a work individually and we are interested in finding average time. In this case there is reciprocal relation between the time taken and speed of the work, more is the time taken by the person less is the speed and less is the time taken by the person more is the speed. In these situations, we can use harmonic mean. In harmonic mean we give more weightage to smaller items and less weightage to large items. it is most useful measure of Central tendency for calculating the ratio. Harmonic mean can be defined as

“It is reciprocal of Arithmetic mean of reciprocal of the observations.”

Mathematically we can write Harmonic Mean as

$$\text{Harmonic Mean or H. M.} = \frac{N}{\Sigma\left(\frac{1}{X}\right)}$$

Where, H. M. = Harmonic Mean

N = Number of items

$\Sigma X$  = Sum of observation.

### 8.2.1 Harmonic Mean in Individual Series

Following are the steps for calculating Geometric mean in the individual series

1. Take the reciprocal of all the values.
2. Find the sum of the reciprocal of the values.
3. Find the arithmetic mean of sum of reciprocal.
4. Reciprocal to the arithmetic mean so calculated is Harmonic Mean to the data.

$$\text{Harmonic Mean or H. M.} = \frac{N}{\Sigma\left(\frac{1}{X}\right)}$$

**Example 4. Find the H. M. for the data given below:**

<b>X</b>	<b>35</b>	<b>45</b>	<b>89</b>	<b>87</b>	<b>66</b>	<b>76</b>	<b>110</b>	<b>135</b>
----------	-----------	-----------	-----------	-----------	-----------	-----------	------------	------------

**Solution:**

X	$\frac{1}{X}$
35	0.0286
45	0.0222
89	0.0112
87	0.0115
66	0.0151
76	0.0131
110	0.0091
135	0.0074
N = 8	$\Sigma\left(\frac{1}{X}\right) = 0.1184$

As

$$\begin{aligned}\text{Harmonic Mean, H. M.} &= \frac{N}{\Sigma\left(\frac{1}{X}\right)} \\ &= \frac{8}{0.1184} = 67.57\end{aligned}$$

$$\therefore \text{H. M.} = 67.57$$

### 8.2.2 Harmonic Mean in Discrete Series

Following are the steps for calculating Geometric mean in the individual series

1. Take the reciprocal of all the values.
2. Multiply the reciprocal with corresponding frequencies to find product.
3. Find the sum of the product of reciprocals and frequencies.
4. Find the arithmetic mean of sum of reciprocal.
5. Reciprocal to the arithmetic mean so calculated is Harmonic Mean to the data.

$$\text{Harmonic Mean or H. M.} = \frac{N}{\sum \left( f \times \frac{1}{X} \right)}$$

**Example 5. Find H. M.**

<b>X</b>	<b>20</b>	<b>50</b>	<b>55</b>	<b>65</b>
<b>f</b>	<b>10</b>	<b>20</b>	<b>15</b>	<b>10</b>

**Solution:**

X	f	$\frac{1}{X}$	$\frac{f}{X}$
20	10	$\frac{1}{20}$	$\frac{10}{20} = 0.5$
50	20	$\frac{1}{50}$	$\frac{20}{50} = 0.4$
55	15	$\frac{1}{55}$	$\frac{15}{55} = 0.2727$
65	15	$\frac{1}{65}$	$\frac{15}{65} = 0.2308$
	$\sum f = N = 60$		$\sum \left( \frac{f}{X} \right) = 1.4035$

As Harmonic Mean, H. M. =  $\frac{N}{\sum \left( \frac{1}{X} \right)}$   
 $= \frac{60}{1.4035} = 42.75$

$$\therefore \text{H. M.} = 42.75$$

### 8.2.3 Harmonic Mean in Continuous Series

Calculation of Harmonic mean in continuous and discrete series is almost same except that in continuous series we take mean value of the class intervals. Following are the steps for calculating Geometric mean in the individual series

1. Find mid value of each class.
2. Take the reciprocal of all the mid values.
3. Multiply the reciprocal with corresponding frequencies to find product.
4. Find the sum of the product of reciprocals and frequencies.
5. Find the arithmetic mean of sum of reciprocal.
6. Reciprocal to the arithmetic mean so calculated is Harmonic Mean to the data.

$$\text{Harmonic Mean or H. M.} = \frac{N}{\sum \left( f \times \frac{1}{X} \right)}$$

**Example 6. Find Harmonic mean, if the data is given as**

<b>C. I.</b>	<b>0 – 100</b>	<b>100 – 200</b>	<b>200 – 300</b>	<b>300 – 400</b>	<b>400 – 500</b>
<b>f</b>	<b>2</b>	<b>7</b>	<b>13</b>	<b>5</b>	<b>3</b>

**Solution:**

C. I.	f	Mid Value x	$\frac{f}{x}$
0 – 100	2	50	0.040
100 – 200	7	150	0.0467
200 – 300	13	250	0.052
300 – 400	5	350	0.0143
400 – 500	3	450	0.0067
	$\sum f = N = 30$		$\sum \left( \frac{f}{x} \right) = 0.1597$

As Harmonic Mean, H. M. =  $\frac{N}{\sum \left( \frac{f}{x} \right)}$

$$= \frac{30}{0.1597} = 180.79$$

$\therefore$  H. M. = 180.79

### 8.2.4 Merits of Harmonic Mean

1. Harmonic mean is rigidly defined.
2. Its calculation is based on all observations.
3. Further algebraic treatment can be applied on it.
4. It is not affected by fluctuation in the sampling.

5. In the problem related to time and work, time and speed etc, this is the best average to measure central tendency.

### 8.2.5 Limitations of Harmonic Mean

1. This is least understood average.
2. Calculation of reciprocal values is not easy task.
3. More weightage is given to small items and big items get lesser weightage.
4. If observation has zero or negative value, it cannot be calculated.

### CHECK YOUR PROGRESS (A)

1. Find Geometric Mean

X	10	110	35	120	50	59	60	7
---	----	-----	----	-----	----	----	----	---

2. Its arithmetic mean of data is 12.5 and G.M. is 10 find the difference between the items.  
3. Calculate G.M.

X	10	15	18	25
f	2	3	5	4

4. Find G.M from the following data

X	3834	382	63	9	.4	.009	.0005
---	------	-----	----	---	----	------	-------

5. Find the G.M of 2, 4 and 8 and prove it is less than A.M

6. Find G.M

C.I.	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
f	4	8	10	6	7

7. Find H.M.

X	10	20	40	60	120
---	----	----	----	----	-----

8. Calculate H.M.

X	10	20	25	40	50
f	20	30	50	15	5

9. Find H.M from the following data

X	3834	382	63	.8	.4	.03	.009	.0005
---	------	-----	----	----	----	-----	------	-------

10. Find H.M

C.I.	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60
f	4	6	10	7	3

## Answers

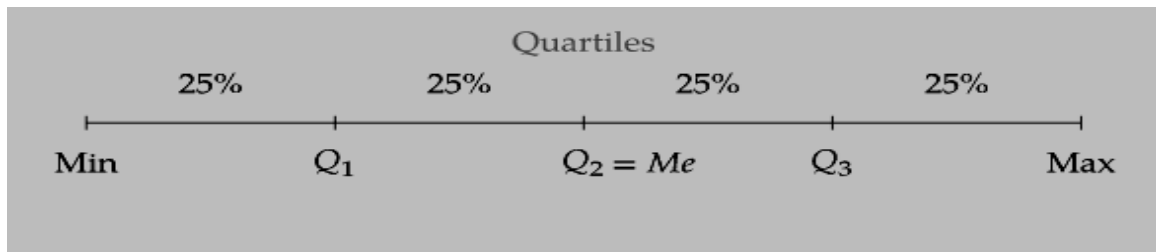
1) 46.56	2) 15	3) 18.2	4) 1.609	
6) 22.06	7) 25	8) 20.08	9) 0.00373	10) 29.88

### 8.5 OTHER POSITIONAL MEASURES (QUARTILES, DECILES AND PERCENTILES)

As median divide the series into two equal parts, there are many other positional measures also. These Positional measures are also known as partition values. Following are some of the positional measure

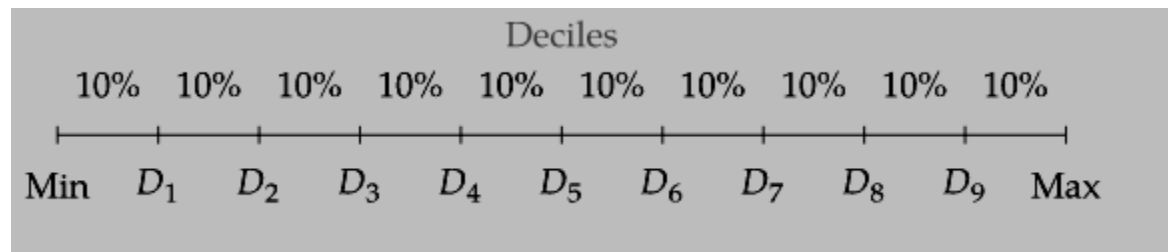
#### 8.5.1 Quartiles

Quartile are the values that divide the series in four equal parts. There is total three quarter in number denoted by  $Q_1$ ,  $Q_2$  and  $Q_3$ . First quartile is placed at 25% of the items, second quartile at 50% of the items, third quartile at 75% of the items. The value of second quartile is always equal to Median.



#### 8.5.2 Deciles

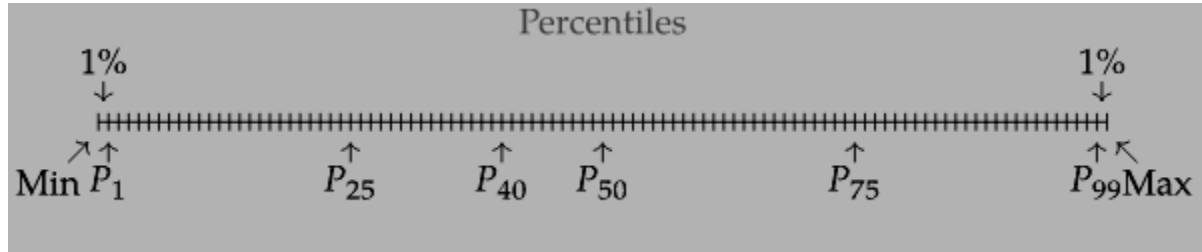
Deciles are the values that divide the series in ten equal parts. There is total nine Deciles in number denoted by  $D_1$ ,  $D_2$ ,  $D_3$  and so on up to  $D_9$ . The first decile is placed at 10% of the items, second quartile at 20% of the items, similarly last at 90% of the items. The value of fifth Decile is always equal to Median.



#### 8.5.3 Percentile



Percentiles are the values that divide the series in hundred equal parts. There is total ninety-nine Percentiles in number denoted by P1, P2, P3 and so on up to P99. The first Percentile is placed at 1% of the items, second quartile at 2% of the items, similarly last at 99% of the items. The value of fifteenth Percentile is always equal to Median.



The methods of finding positional measures are same as in case of median. However, following are the formulas that can be used for finding positional measures.

Partition Value	Individual Series	Discrete Series	Continuous Series	Continuous Series
Q1	Value of $\left(\frac{N+1}{4}\right)^{\text{th}}$ item	Value of $\left(\frac{N+1}{4}\right)^{\text{th}}$ item	Value of $\left(\frac{N}{4}\right)^{\text{th}}$ item	$L + \frac{\frac{N}{4}-C.f}{f} \times i$
Q3	Value of $3\left(\frac{N+1}{4}\right)^{\text{th}}$ item	Value of $3\left(\frac{N+1}{4}\right)^{\text{th}}$ item	Value of $3\left(\frac{N}{4}\right)^{\text{th}}$ item	$L + \frac{3\left(\frac{N}{4}\right)-C.f}{f} \times i$
D6	Value of $6\left(\frac{N+1}{10}\right)^{\text{th}}$ item	Value of $6\left(\frac{N+1}{10}\right)^{\text{th}}$ item	Value of $6\left(\frac{N}{10}\right)^{\text{th}}$ item	$L + \frac{6\left(\frac{N}{10}\right)-C.f}{f} \times i$
P40	Value of $40\left(\frac{N+1}{100}\right)^{\text{th}}$ item	Value of $40\left(\frac{N+1}{100}\right)^{\text{th}}$ item	Value of $40\left(\frac{N}{100}\right)^{\text{th}}$ item	$L + \frac{40\left(\frac{N}{100}\right)-C.f}{f} \times i$

Similarly, all the values can be calculated.

**(A) Individual Series**

**Example 7.** From the data given below, determine Q<sub>1</sub>, Q<sub>3</sub>, D<sub>5</sub>, P<sub>40</sub>.

Marks in Economics	18	20	25	24	32	50	55	45	55	40	60
--------------------	----	----	----	----	----	----	----	----	----	----	----

**Solution:** Arranging the given figures in ascending order, we get

S. No.	1	2	3	4	5	6	7	8	9	10	11	N = 11
--------	---	---	---	---	---	---	---	---	---	----	----	--------

Marks	18	20	24	25	32	40	45	50	52	55	66	
-------	----	----	----	----	----	----	----	----	----	----	----	--

Now  $Q_1 = \text{Value of } \left(\frac{N+1}{4}\right)^{\text{th}} \text{ item}$   
 $= \text{Value of } \left(\frac{11+1}{4}\right)^{\text{th}} \text{ item}$   
 $= \text{Value of } 3^{\text{rd}} \text{ item} = 24$

$\therefore Q_1 = 24$

$Q_3 = \text{Value of } 3 \left(\frac{N+1}{4}\right)^{\text{th}} \text{ item}$   
 $= \text{Value of } 3 \left(\frac{11+1}{4}\right)^{\text{th}} \text{ item}$   
 $= \text{Value of } 9^{\text{th}} \text{ item} = 52$

$\therefore Q_3 = 52$

$D_5 = \text{Value of } 5 \left(\frac{N+1}{10}\right)^{\text{th}} \text{ item}$   
 $= \text{Value of } 5 \left(\frac{11+1}{10}\right)^{\text{th}} \text{ item}$   
 $= \text{Value of } 6^{\text{th}} \text{ item} = 40$

$\therefore D_5 = 40$

$P_{40} = \text{Value of } 40 \left(\frac{N+1}{100}\right)^{\text{th}} \text{ item}$   
 $= \text{Value of } 40 \left(\frac{12}{100}\right)^{\text{th}} \text{ item}$   
 $= \text{Value of } 4.8^{\text{th}} \text{ item} = 24$

$\Rightarrow P_{40} = 4^{\text{th}} \text{ item} + 0.8 (5^{\text{th}} \text{ item} - 4^{\text{th}} \text{ item})$   
 $= 25 + 0.8 (32 - 25)$   
 $= 25 + 0.8 (7)$   
 $= 25 + 5.6 = 30.6$

$\therefore P_{40} = 30.6$

**(B) Discrete Series**

**Example 8.** Form the following data, compute  $Q_1$ ,  $Q_3$ ,  $D_8$  and  $P_{70}$ .

<b>X</b>	<b>110</b>	<b>120</b>	<b>130</b>	<b>140</b>	<b>150</b>	<b>160</b>	<b>170</b>
<b>f</b>	<b>2</b>	<b>3</b>	<b>5</b>	<b>10</b>	<b>5</b>	<b>3</b>	<b>2</b>

**Solution:**

X	f	C.I.
110	2	2 (1 - 2)
120	3	5 (3 - 5)
130	5	10 (6 - 10)

140	10	20 (11 – 20)
150	5	25 (21 – 25)
160	3	28 (26 – 28)
170	2	30 (29 – 30)
	$\sum f = N = 30$	

Now  $Q_1 = \text{Value of } \left(\frac{N+1}{4}\right)^{\text{th}} \text{ item}$   
 $= \text{Value of } \left(\frac{30+1}{4}\right)^{\text{th}} \text{ item}$   
 $= \text{Value of } 23.25^{\text{th}} \text{ item} = 150$

$\therefore Q_1 = 150$

$Q_3 = \text{Value of } 3 \left(\frac{N+1}{4}\right)^{\text{th}} \text{ item}$   
 $= \text{Value of } 3 \left(\frac{30+1}{4}\right)^{\text{th}} \text{ item}$   
 $= \text{Value of } 9.3^{\text{rd}} \text{ item} = 130$

$\therefore Q_3 = 130$

$D_8 = \text{Value of } 8 \left(\frac{N+1}{10}\right)^{\text{th}} \text{ item}$   
 $= \text{Value of } 8 \left(\frac{31+1}{10}\right)^{\text{th}} \text{ item}$   
 $= \text{Value of } 24.8^{\text{th}} \text{ item} = 150$

$\therefore D_8 = 150$

$P_{70} = \text{Value of } 70 \left(\frac{N+1}{100}\right)^{\text{th}} \text{ item}$   
 $= \text{Value of } 70 \left(\frac{30+1}{100}\right)^{\text{th}} \text{ item}$   
 $= \text{Value of } 21.7^{\text{th}} \text{ item} = 150$

$\therefore P_{70} = 150$

**(B) Continuous Series**

**Example 9. Calculate Quartiles,  $D_6$  and  $P_{20}$**

C. I.	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80
f	8	22	40	70	90	40	20	10

**Solution:**

C. I.	f	Cumulative Frequency (C · f)
0 – 10	8	8 (1 – 8)
10 – 20	22	30 (9 – 30)
20 – 30	40	70 (31 – 70)
30 – 40	70	140 (71 – 140)

40 – 50	90	230 (141 – 230)
50 – 60	40	270 (231 – 270)
60 – 70	20	290 (271 – 290)
70 – 80	10	300 (291 – 300)
	$\sum f = N = 300$	

Now  $Q_1 = \text{Size of } \left(\frac{N}{4}\right)^{\text{th}} \text{ item}$   
 $= \text{Size of } \left(\frac{300}{4}\right)^{\text{th}} \text{ item}$   
 $= \text{Size of } 75^{\text{th}} \text{ item}$

$\Rightarrow Q_1$  lies in the class – interval 30 – 40

$\Rightarrow Q_1 = L + \frac{\frac{N}{4} - C \cdot f}{f} \times i$   
 $= 30 + \frac{75 - 70}{70} \times 10$   
 $= 30 + \frac{50}{70} = 30.71$

$\therefore Q_1 = 30.71$

$Q_2 = \text{Size of } 2 \left(\frac{N}{4}\right)^{\text{th}} \text{ item}$   
 $= \text{Size of } 2 \left(\frac{300}{4}\right)^{\text{th}} \text{ item}$   
 $= \text{Size of } 150^{\text{th}} \text{ item}$

$\Rightarrow Q_2$  lies in the class – interval 40 – 50

$\Rightarrow Q_2 = L + \frac{2\left(\frac{N}{4}\right) - C \cdot f}{f} \times i$   
 $= 40 + \frac{150 - 140}{90} \times 10$   
 $= 40 + \frac{100}{90} = 41.11$

$\therefore Q_2 = 41.11$

$Q_3 = \text{Size of } 3 \left(\frac{N}{4}\right)^{\text{th}} \text{ item}$   
 $= \text{Size of } 3 \left(\frac{300}{4}\right)^{\text{th}} \text{ item}$   
 $= \text{Size of } 225^{\text{th}} \text{ item}$

$\Rightarrow Q_3$  lies in the class – interval 40 – 50

$\Rightarrow Q_3 = L + \frac{3\left(\frac{N}{4}\right) - C \cdot f}{f} \times i$   
 $= 40 + \frac{225 - 140}{90} \times 10$   
 $= 40 + \frac{850}{90} = 49 + 9.44 = 49.44$

$\therefore Q_3 = 49.44$

$$D_6 = \text{Size of } 6 \left(\frac{N}{10}\right)^{\text{th}} \text{ item}$$

$$= \text{Size of } 6 \left(\frac{300}{10}\right)^{\text{th}} \text{ item}$$

$$= \text{Size of } 180^{\text{th}} \text{ item}$$

⇒  $D_6$  lies in the class – interval 40 – 50

$$\Rightarrow D_6 = L + \frac{6\left(\frac{N}{10}\right) - C.f}{f} \times i$$

$$= 40 + \frac{180 - 140}{90} \times 10$$

$$= 40 + \frac{40}{9} = 40 + 4.44 = 44.44$$

∴  $D_6 = 44.44$

$$P_{20} = \text{Size of } 20 \left(\frac{N}{100}\right)^{\text{th}} \text{ item}$$

$$= \text{Size of } 20 \left(\frac{300}{100}\right)^{\text{th}} \text{ item}$$

$$= \text{Size of } 60^{\text{th}} \text{ item}$$

⇒  $P_{20}$  lies in the class – interval 20 – 30

$$\Rightarrow P_{20} = L + \frac{20\left(\frac{N}{100}\right) - C.f}{f} \times i$$

$$= 20 + \frac{60 - 30}{40} \times 10$$

$$= 20 + \frac{30}{4} = 20 + 7.5 = 27.5$$

∴  $P_{20} = 27.5$

### CHECK YOUR PROGRESS (B)

1. Find  $Q_1$ ,  $Q_3$ ,  $D_5$ ,  $P_{25}$  and  $P_{67}$

X: 37, 39, 45, 53, 41, 57, 43, 47, 51, 49, 55

2. Calculate Median,  $Q_1$ ,  $Q_3$  and  $D_6$

Marks Less Than	80	70	60	50	40	30	20	10
No. of Students	100	90	80	60	32	20	13	5

3. Calculate Medium,  $Q_1$  and  $P_{85}$

Income ('000)	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
No. of Families	75	250	350	192	68	35	24	6

### Answers

1) $Q_1 = 41$ , $Q_3 = 53$ , $D_5 = 47$ , $P_{25} = 41$ , $P_{67} = 51.08$	2) $M = 46.4$ , $Q_1 = 34.2$ , $Q_3 = 57.5$ , $D_6 = 50$	3) $M = \text{Rs. } 1250$ , $Q_1 = \text{Rs. } 850$ , $P_{85} = \text{Rs. } 1956$
---	---	--

## 8.4 SUM UP

- Geometric Mean is useful for calculating growth and decline rate.
- Harmonic mean is useful for speed and work etc.
- Median divide the series in two equal parts.
- Mode is value repeated most number of time.
- There are other positional measures like Quartile, Decile and Percentile

## 8.5 KEY TERMS

- Geometric Mean: Geometric mean of N items is root nth of the product of item
- Harmonic Mean: It is reciprocal of Arithmetic mean of reciprocal of the observations.
- Median: It is a value that divide the series in two equal parts.
- Mode: It is the most repeated value of the series.
- Quartile: It is a value that divide the series in four equal parts.
- Decile: It is a value that divide the series in ten equal parts.
- Percentile: It is a value that divide the series in hundred equal parts.

## 8.6 QUESTIONS FOR PRACTICE

- Q1. What is Geometric Mean. Give process of calculating Geometric mean in different series.
- Q2. What are merits and limitations of geometric mean.
- Q3. What is Harmonic Mean. How it is calculated.
- Q4. Give relation between Mean, Median and Mode.
- Q5. Explain the concept of Quartile, Percentile and Deciles.
- Q6. What is positional average? Give various positional average.
- Q7. According to you which measure of average is best.

## 8.7 SUGGESTED READINGS

- J. K. Sharma, Business Statistics, Pearson Education.
- S.C. Gupta, Fundamentals of Statistics, Himalaya Publishing House.
- S.P. Gupta and Archana Gupta, Elementary Statistics, Sultan Chand and Sons, New Delhi.
- Richard Levin and David S. Rubin, Statistics for Management, Prentice Hall of India, New Delhi.