CHAPTER ONE

1. INTRODUCTION

Dear students:

- From your previous knowledge what do you mean by statistics?
- What are the two broad areas of statistics?
- Why study Statistics?
- What are some of the application areas of statistics?

Objectives

Dear learner at the end of this chapter, you will be able to:

- *Define Statistics*
- *Identify Classification of Statistics*
- *Understand areas of application of Statistics*

1.1. Definition of statistics

Dear student, you might think of statistics as something related to numerical measurements of some attributes of interest. In fact, in every day usage, the term statistics is often taken to mean numerical facts. Since different people have different understanding of statistics, we can say there are as many definitions as the number of people who have tried to define the term statistics. Some of these definitions are given below.

- Statistics is a branch of mathematics that consists of a set of analytical techniques that can be applied to data to help in making judgments and decisions in problems involving uncertainty.
- Statistics is a scientific discipline consists of procedures for collecting, describing, analyzing and interpreting numerical data.
- Statistics is a body of principles and methods concerned with extracting useful information from a set of numerical data.

In general, its meaning can be categories into two entirely different categories. These are statistics in plural sense and singular sense categories.

In the plural sense, statistics are the raw data themselves, like statistics of births, statistics of deaths, statistics of students, statistics of imports and exports, employment statistics, accident statistics, population statistics, economic statistics, agricultural statistics, statistics of business firms of imports and exports, index numbers that help us understand a variety of business and economic situations, etc.

Statistics are aggregates of numerically expressed facts or figures collected in a systematic manner for a pre-determined purpose. But every numerical facts or figures may not be statistics. Such a numerical data to be statistics it must fulfill the following.

- The numerical facts must be collected for a predetermined purpose.
- The numerical facts must be collected in a systematic manner.
- The numerical facts must be aggregate of facts, and
- The numerical facts must be affected by a multiplicity of cases.

Examples of statistics include:

- Daily exchange rates in a country
- Monthly unemployment rate in a country
- Scores of students for a test
- Number of students registered in a university for each year
- Initial capital of investment for investors
- Number of employees by sex and department in an organization

In singular sense, Statistics is defined as the art and science of collecting, analyzing, presenting, and interpreting data. Particularly in business and economics, the information provided by collecting, analyzing, presenting, and interpreting data gives managers and decision makers a better understanding of the business and economic environment and thus enables them to make more informed and better decisions.

In general Statistics teaches us how to summarize, analyze, and draw meaningful inferences from data that lead to improved decisions making. These decisions that we made help us to improve the running; for example, a department, a company, the entire economy, etc.

Dear learner you study statistics for several reasons:

1. Data are available everywhere.
2. You may be called on to conduct research in your field of study. To accomplish this, you must be able to design experiments; collect, organize, analyze, and summarize data; and possibly make reliable predictions or forecasts for future use. In such a case you use statistics since statistical procedures are basic to research. You must also be able to communicate the results of the study in your own words.
3. You can also use the knowledge gained from studying statistics to become better consumers and citizens. For example, you can make intelligent decisions about what products to purchase based on consumer studies, about government spending based on utilization studies, and so on.
4. No matter what your career, you will make professional decisions that involve data. Understanding of statistical methods will help you make these decisions effectively.
5. Statistical techniques are used to make many decisions that affect our lives
6. Like professional people, you must be able to read and understand the various statistical studies performed in your fields. To have this understanding, you must be knowledgeable about the vocabulary, symbols, concepts, and statistical procedures used in these studies.

1.2. Classifications of statistics

Depending on how data can be used statistics is divided in to two main areas or branches.

1. **Descriptive Statistics:** Methods of organizing, summarizing, and presenting data in an informative way.

In descriptive statistics the statistician tries to describe a situation. Consider the national census conducted by the Ethiopia government every 10 years. Results of this census give you the average age, income, and other characteristics of the Ethiopia population. To describe such a data you can use tables, charts, graphs or summary statistics like averages, percentages, or measures of dispersion.

Additional examples include

- Presenting the starting salaries of 1000 Management graduates in different organizations.
- From sample we have 40% of the employee suggest positive attitude toward the management of the organization.
- From 500 management graduates in this year 70% were females.
- In Ethiopia the minimum monthly salary is 600 Birr.

2. **Inferential Statistics:** The methods used to determine something about a population on the basis of a sample. For example, the average income of all families (the population) in Ethiopia can be estimated from figures obtained from a few hundred (the sample) families. Here, the statistician tries to make inferences from samples to populations. Inferential statistics uses probability, i.e., the chance of an event occurring. Probability theory is used in the insurance industry and other areas.

It is important to define terms like the following to understand inferential statistics

A **population** consists of all subjects (human or other) that are being studied or population is a group with common characteristics. For example; all households in a certain location, all

financial institutes in a country, all students in a university, all employees in an organization, and all account receivables in an organization are population.

Most of the time, due to the expense, time, size of population, medical concerns, etc., it is not possible to use the entire population for a statistical study; therefore, researchers use samples.

A **sample** is a group of subjects selected from a population systematically. Therefore, it is some part of the population.

If the subjects of a sample are properly selected, most of the time they should possess the same or similar characteristics as the subjects in the population.

Variable is characteristics of a population that assumes or takes different values. For example, if you want to study on employees in an organization then age, sex, service year in the organization, beginning monthly salary, contract type, education level, current monthly salary and field of study of the employees are variables.

Data is a record value of a variable and usually it is a raw fact.

Parameter is summary value of a variable in the entire study population. Usually, it is unknown but it can be estimated. It may be expressed in average or proportion or percentage or dispersion. For example; if you study about employees in an organization and if you are interested in the variable's current monthly salary and contract type of employees, then the mean monthly salary of employees and the proportion of full-time workers in the organization are example of parameter.

Statistic is summary value of the study variable in the sample not in the entire study population. Since the size of the sample is manageable compared to population size statistic is usually known and it is estimated value of a parameter value. It is expressed as sample average, proportion, percentage or numerical measures of spread.

1.3. Applications, uses and limitations of statistics

Applications of Statistics

In today's global business and economic environment, anyone can access vast amounts of statistical information. The most successful managers and decision makers understand the information and know how to use it effectively. In this section, we provide examples that illustrate some of the uses of statistics in business and economics.

In management

Modern business management is more a Science than an Art. Ever increasing global competition mandates business managers to address uncertainty by using scientific methods and be objective decision makers. Forecasting, planning, organizing and decision making; some of the key activities of a manager intend better future for the business. The only certainty about the future is its 'uncertainty'. Even though one cannot eliminate uncertainty, it is possible to measure uncertainty using Statistics: manager can make informed decisions by using Statistical methods and Statistical thinking. This calls for unravelling the power of Statistics for managers. Broadly, knowledge of statistics helps a manager to describe the problem, identify and evaluate alternative courses of action, estimate error, monitor processes and take appropriate corrective actions to achieve optimum results.

Both Descriptive and Inferential statistical methods find important place in business management. To quote a few of the many applications across functions,

- A marketing manager needs to gather and analyze a large amount of data pertaining to market dynamics and target customers. Ideally, marketing strategy depends up on the outcomes of a market research, which involves statistical methods for collecting and analyzing data, application of sampling techniques and evaluating the effect of various marketing strategies.

- A production manager would ideally use statistical process control techniques to improve productivity and quality. Knowledge and application of control charts, sampling techniques and probability distributions ensures better processes and products. This also leads to the reduction in production cost and higher profits.
- A human resource manager would be interested in identifying the best approach to train employees and evaluate the impact of training. There is a need to measure attrition and understand the underlying factors.
- For finance manager, crunching financial data and using financial techniques is an integral part of day-to-day job. Knowledge of statistics enhances competency and proficiency of a manager as a researcher and therefore provides an edge.

In Operations Management

A critical aspect of production is quality. The quality of a final product is a function of the quality of the product's components. If the components don't fit, the product will not function as planned and likely cease functioning before its customers expect it to. For example, if a car door is not made to its specifications, it will not fit. As a result, the door will leak both water and air.

Operations managers attempt to maintain and improve the quality of products by ensuring that all components are made so that there is as little variation as possible. As you have already seen, statisticians measure variation by computing the variance.

In Economics

- Economists frequently provide forecasts about the future of the economy or some aspect of it.
- They use a variety of statistical information in making such forecasts. For instance, in forecasting inflation rates,
- Economists use statistical information on such indicators as the Producer Price Index, the unemployment rate, and manufacturing capacity utilization. Often these statistical indicators are entered into computerized forecasting models that predict inflation rates.

In Production

- Today's emphasis on quality makes quality control an important application of statistics in production.
- A variety of statistical quality control charts are used to monitor the output of a production process.

In Finance

- Financial analysts use a variety of statistical information to guide their investment recommendations.
- In the case of stocks, the analysts review a variety of financial data including price/earnings ratios and dividend yields. By comparing the information for an individual stock with information about the stock market averages, a financial analyst can begin to draw a conclusion as to whether an individual stock is over- priced or underpriced.

In general, Statistics have the following applications

- Almost all human beings in their daily life are subjected to obtaining numerical facts e.g., about price.
- Applicable in some process e.g., invention of certain drugs, extent of environmental pollution.
- In industries especially in quality control area.
- In almost all fields of human endeavor.

Uses of Statistics

The main function of statistics is to enlarge our knowledge of complex phenomena.

The following are some uses of statistics:

- It presents facts in a definite and precise form.
- Data reduction.
- Measuring the magnitude of variations in data.
- Furnishes a technique of comparison
- Estimating unknown population characteristics.
- Testing and formulating of hypothesis.
- Studying the relationship between two or more variable.
- Forecasting future events.

Limitations of statistics

As a science statistics has its own limitations. The following are some of the limitations:

- It deals with only quantitative information.
- Deals with only aggregate of facts and not with individual data items.
- Statistical data are only approximately and not mathematical correct.
- Statistics can be easily misused and therefore should be used by experts.

Exercise1.1

Answer all questions briefly

1. Write True if the statement is correct or False if the statement is wrong.

    a) Summary value of a variable in the sample is parameter.
    b) All numerical facts are statistics.
    c) Statistics is used to make comparison among different groups.
    d) In inferential statistics the statistician tries to describe a situation.
    e) Sample is a subset of the study population.
    f) Variable is record value of a unit.

2. In each of these statements, tell whether descriptive or inferential statistics have been used.
    i) Expenditures for the cable industry were 6.5 billion birr in 1990
    ii) By 2035 at least 3.5 billion people will run shortage of water.
    iii) Nine out of ten on-the-job fatality rates are men.
    iv) The mean household income for people aged 25–34 is 25,000 birr.
    v) Last year in Ethiopia the inflation rate was 30%.
    vi) In a country for every 100 workable people 20 are unemployed.
    vii) The prediction of fish production in Ethiopia will increase by 30% in 2025.

3. The unemployment rate in a given country at a given period is believed to be 25%. The government embarked on a series of projects to reduce unemployment. It was of interest to determine whether the unemployment rate decreases as a result of the projects. A random sample of 500 workable people was chosen, and 100 of them were found to be unemployed. Based on the information identify:
    a) The study population in this research
    b) The variable of interest for the researcher.
    c) The parameter
    d) The statistic

CHAPTER TWO
2. DATA COLLECTION AND PRESENTATION

Dear learner what is a variable? What do you mean by data? How can you classify data? How can you collect data? What are the sources of data? How can you present data?

Introduction

In statistical investigation, investigators gain information about a particular situation by collecting data for variables. If the collected data are inaccurate and inadequate, the whole analysis may be faulty and the conclusions drawn are misled.

The collected data, also known as raw data are always in an unorganised form and need to be organised and presented in meaningful and readily comprehensible form in order to facilitate further statistical analysis.

This chapter explains data and its classification, methods of data collection, how to organize data by constructing frequency distributions and how to present the data by constructing diagrams and graphs.

Chapter Objectives:

Dear learner after successful completing of this chapter, you should be able to

- Identify types of data
- Identify different methods of data collection
- Organize data by constructing frequency distributions
- Present data by constructing diagrams and graphs

2.1 Data definition

Definition 2.1

Data are the facts and figures collected, analysed, and summarized for presentation and interpretation. The word 'data' is the plural of the Latin word 'datum', meaning 'fact'. All the data collected in a particular study are referred to as the **data set** for the study.

One of the main functions of statistics is to provide information which will help on making decisions. Statistics provides the type of information by providing a description of the present, a profile of the past and an estimate of the future.

Data collection is the first and the most important activity in the use of statistical methods for undertaking an investigation. Different types of data can be collected for different purposes. The data can be collected in connection with time or geographical location or in connection with time and location.

2.2 Classification of Data

Definition 2.2

Classification is the process of grouping into different classes or sub classes according to some characteristics.

The classification of data is generally done on geographical, chronological, qualitative or quantitative basis, or based on measurement scale.

i) **Chronological classification:** In chronological classification the collected data are arranged according to the order of time expressed in years, months, weeks, etc. The data are generally classified in ascending order of time. Example the data related with population census, sales of a firm, imports and exports of a country in a year.
ii) **Geographical classification:** In this type of classification the data are classified according to geographical region or place. Example: production of Goods in different countries.

iii) **Qualitative classification:** In this type of classification data are classified on the basis of same attributes or quality. Example: data classified according to sex, literacy, religion, employment etc.

iv) **Quantitative classification:** refers to the classification of data according to some characteristics that can be measured/ expressed in natural numeric. Example: height, weight, revenue, profit, cost, inflation rate, deflation rate, unemployment rate,etc.

Qualitative and Quantitative Data

According to **the nature of variables,** data can be classified as either qualitative or quantitative.

- **Qualitative data** (also called **categorical data**) are data that include labels or names used to identify an attribute of each element. They might be nonnumeric or numeric. When the qualitative data use a numeric code, arithmetic operations such as addition, subtraction, multiplication, and division do not provide meaningful results.
  A qualitative variable is a variable with qualitative data. Examples: Gender, product type, blocks number and geographic locations of residents, contract type of employees, customer ratings of products, rank of companies, education level of workers in a plant, economic system of a country and so forth.
- **Quantitative data** require numeric values that indicate how much or how many. They are numeric. Arithmetic operations often provide meaningful results for a quantitative variable. A quantitative variable is a variable with quantitative data. Examples: Number of students registered per year per program, monthly salary of employees, yearly cost, revenue, profit of companies, weekly wages of workers and so on.

Quantitative variable in turn can be further classified into two groups: discrete and continuous. **Discrete variables** can assume only certain values, and there are usually "gaps" between the values. Further such variables are formed or created by counting. Number of children in a family, number of full-time workers in an organization, number of products in a factory per day, number of daily labourers in a plant and so forth are illustrative examples of discrete quantitative variables. **Continuous Variables** can assume an infinite number of values in an interval between any two specific values. Price of gasoline per gallon, profit, revenue and cost of a plant in cash, area of investment land, and capital budget of organizations are some examples of continuous quantitative variables.

Cross-Sectional, Time Series and Panel Data

Data can be classified based on time reference as cross-sectional, time series and panel data.

Cross-sectional data are data collected at the same or approximately the same point in time. Example: number of unemployed persons in African countries this year.

Time series data are data collected over several time periods. Example: monthly electric consumption of a household from November 2009 through October 2010, government budget of Ethiopia from 1991 to 2019, number of students registered in Wollo University from 2007 to 2019 GC, and yearly profit in cash of Dashin Brewery from its establishment to this year.

Panel data are data collected over several time periods with several locations. For example, the budget allocated in our country from 1984 to 2010 E.C by regions, number of students registered in Wollo University from 2000 to 2011 E.C by department, and amount of wheat production in quintal in our country by regions from 1984 to 2011 E.C are panel data.

Data classification based on Scales of Measurement

Measurement (or observation) is the process of determining and recording which of the possible traits of a variable an individual/ object case exhibits or possesses. The variable 'sex' has two possible traits, female and male, and measurement involves deciding into which of

these two categories a given person falls. This set of categories that can possibly be assigned to individual cases makes up a scale of measurement.

A scale of measurement specifies a range of scores (also called points on the scale) that can be assigned to cases during the process of measurement. When we construct a scale of measurement, we need to follow two particular rules.

First, *the scale must capture sufficient variation to allow us to answer our research question(s).* Take the research questions that how old are you? I may use 'number of whole years elapsed since birth' as the operational definition for measuring the ages of students in my statistics class. This produces a scale with whole years as the points on the scale, and will yield a variety of scores for age, given the fact that my students were born in different years. Imagine, though, that I was teaching a class of university students; measuring age with this scale will be inadequate since most or all of the students will have been born in the same year. Using a measurement scale for age that only registers whole years will not pick up enough variation to help me meet my objectives; every student in the class will appear to be the same age. I might consider, instead, using 'number of whole *months* elapsed since birth' as the scale of measurement. Age *in months* will capture variation among students in a primary class that age *in years* will miss. This example of the age for a group of students highlights an intrinsic problem when we try to set up a scale to measure sufficient variation for a continuous variable that does not arise when we try to measure a discrete variable. A discrete variable has a finite number of values. A continuous variable can vary in quantity by infinitesimally small degrees.

Discrete variables often have a unit of measurement that cannot be subdivided, such as the number of children per household. Other examples of discrete variables are the number of prisoners per jail cell, the number of welfare agencies in a district, and the number of industrial accidents in a given year.

The age of students, on the other hand, is a continuous variable. Age can conceivably change in a *gradual* way from person to person or for the same person over time. Because of this, continuous variables are measured by units that can be infinitely subdivided. Age, for example, does not have a basic unit with which it is measured. We may begin by measuring age in terms of years. But a year can be divided into months, and months into weeks, weeks into days, and so on. The only limit is exactly how precise we want to be: years capture less variation than months, and months less than weeks. *Theoretically*, with a continuous variable we can move gradually and smoothly from one value of the variable to the next without having to jump. *Practically*, though, we will always have to 'round off' the measurement and treat a continuous variable *as if* it is discrete, and this causes the scale of measurement to 'jump' from one point on the scale to the next. The *scale* is by necessity discrete, even though the underlying variable is continuous.

The use of a discrete measurement scale to measure age, whether we do it in years or months, causes us to cluster cases together into groups. The points on the scale act like centers of gravity pulling in all the slight variations around them that we do not want to worry about. We may say that two people are each 18 years old, but they will in fact be different in terms of age, unless they are born precisely at the same moment. But the slight difference that exists between someone whose age is 18 years, 2 months, 5 days, 2 hours, 12 seconds... and someone whose age is 18 years, 3 months, 14 days, 7 hours, 1 second...might be irrelevant for the research problem we are investigating and we treat them the same in terms of the variable, even though they are truly different. No measurement scale can ever hope to capture the full variation expressed by a continuous variable. The practical problem we face is whether the scale captures *enough* variation to help us answer our research question.

The second rule of measurement is that *a scale must allow us to assign each case into one, and only one, of the points on the scale*. This statement actually embodies two separate principles of measurement. The first is **the principle of exclusiveness**, which ***states that no***

case should have more than one value for the same variable. For example, someone cannot be both 18 years of age and 64 years of age. Measurement must also follow **the principle of exhaust**iveness, which ***states that every case can be classified into a category***. A scale for health status that only had 'healthy' and 'very healthy' as the points on the scale is obviously insufficient; anyone who is less than healthy cannot be measured on this scale.

According to the scale of measurement, data can be classified as nominal, ordinal, interval and ratio. The scale of measurement determines the amount of information contained in the data and indicates the most appropriate data summarization and statistical analyses.

Nominal Scale of measurement: When the data for a variable consist of labels or names used to identify values of variable and there is no ranking or order can be placed on the data is considered a nominal scale of measurement. Examples: religion, blood type, eye colour, marital status, identification number, employee contract type, economic system of a government.

Ordinal Scale of measurement: Data measured at this level can be placed into categories, and these categories can be ordered, or ranked; however, precise differences between the ranks do not exist. Examples: service quality (excellent, good, poor). Here hence, excellent indicate the best service, followed by good and then poor. Letter grades of students on a course (A, B, C, D, and F). Living standard of a family: poor, medium, rich. Consumer ratings of products, education level of workers in a factory, and company rankings are additional examples of ordinal scale data.

Interval Scale of measurement: data show the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure. Examples include intelligent quotient of a person, temperature in degree Celsius and Fahrenheit scale. One property lacking in the interval scale is there is no true zero. For example, intelligent quotient tests do not measure people who have no intelligence. For temperature, 0^0F does not mean no any heat at all.

Ratio scale of measurement possesses all properties of interval data and the ratio of two values is meaningful. Examples include distance, height, weight, inflation rate, unemployment rate, number of labour force and time, costs, revenues, profits, etc. This scale requires that a zero value be included to indicate that nothing exists for the variable at the zero point. For example, consider the cost of an automobile. A zero value for the cost would indicate that the automobile has no cost and is free. In addition, if we compare the cost of 300,000 (in Birr) for one automobile to the cost of 150,000 (in Birr) for a second automobile, the ratio property shows that the first automobile is $(300{,}000\ Birr/150{,}000\ Birr = 2\ times)$, twice, the cost of the second automobile.

2.3 Methods of Data Collection

Any statistical data can be classified under two categories depending upon the sources utilized: primary and secondary data. **Primary data** are those which are collected a fresh and for the first time and thus happened to be original in character while **secondary data** are those which have been collected by someone else and which have already been passed through the statistical process. Primary data are thus the one which are collected by the investigator himself/ herself for the purpose of a specific investigation or study.

Secondary data are those data gathered or compiled from published and unpublished sources or files. The source of secondary data includes publication of central, state, and local government; technical and trade journals, books, magazines, newspapers; reports and publication of industry, bank and stock exchange; public records, and reports by research scholars, universities and economists. The following are some of the points that are to be considered in the use of secondary data

1. How the data has been collected and processed
2. The accuracy of the data

3. How far the data has been summarized
4. How to interpret the data, especially when figures collected for one purpose is used for another

☛ Dear learner you have to notice that data which are primary for one may be secondary for the other.

Various methods of primary data collection are employed. The most commonly used methods of primary data collection are:

2.3.1 Observation

Observation may be defined as systematic viewing, coupled with consideration of seen phenomenon. Observation method is a method under which data from the field is collected with the help of observation by the observer or by personally going to the field. It has the advantage of eliminating subjective bias (no bias information), researchers get current information and the information is independent of the respondents. Its disadvantages include it is expensiveness (it needs more time), the information is limited, unforeseen factors may interfere with observational task and respondent's opinion cannot be recorded on certain subjects.

2.3.2 Interview method

It is oral-verbal communication where interviewer asks questions (which are aimed to get information required for the study) to the respondent. This method of data collection involves presentation or oral-verbal stimuli and reply in terms of oral-verbal responses. There are different types of interviews. Some of them include:

2.3.2.1 Personal interviews

The interviewer asks questions generally in face-to-face contact to the other person or persons. Its advantage includes

- Greater depth of information
- Flexibility of restructuring the questionnaire
- Interviewer by his/ her skill can come over resistance.
- Non response is generally low
- Samples can be controlled more effectively
- Personal information can be obtained
- Interviewer can collect supplementary information about respondents' personal characteristics and environment which has value in interpreting results.

While its disadvantage includes the following:

- It is expensive method
- Respondent may give bias information
- Some executive people are not approachable so the data collected may not be adequate
- Take more time when samples are more
- Systematic errors may be occurred
- Supervisors have to do complex work of selecting, training, and supervising the field staff.

2.3.2.2 Telephone interview

It is the process of contacting samples (respondents) on telephone. It is uncommon method may be used in developed regions. It has the following merits:

- Flexible compared to mailing method
- Faster than other methods
- Cheaper than personal interview method
- Call-backs are simple and economical
- High response than mailing method

- When it is not possible to contact the respondent directly, then interview is conducted through telephone.
- Replies can be recorded without embarrassment to respondent
- Interviewer can explain requirements more easily
- No field staff is required
- Wider distribution of sample is possible

Its demerits include:

- Little time is given to respondent
- Survey is restricted to respondents who have telephones
- Not suitable for intensive survey where comprehensive answers are required
- Bias information may be more
- Very difficult to make questionnaire because it should short and to the point

2.3.3 Questionnaire method

This method of data collection is quite popular, particularly in case of big inquiries. The questionnaire is sent or mailed to the respondents who are expected to read and understand the questions and write down the reply in the space meant for the purpose in the questionnaire itself. The respondents have to answer the questions on their own. This questionnaire method has merits like low cost even the geographical area is large to cover, answers are in respondents' word so free from bias, it gives adequate time for respondents to think for answers, non-approachable respondents may be conveniently contacted, and large samples can be used so results are more reliable. While it has the demerits like low rate of return of dully filled questionnaires, can be used when respondent is educated and cooperative, it is inflexible, omission of some questions, difficult to know the expected respondent have filled the form or it is filled by someone else, and it is the slowest method of data collection.

2.3.4 Schedule method

In the wording of Thomas Carson Marcormic the schedule method is nothing more than a list of questions which it seems necessary to test the hypothesis. It is one of the important methods for the study of social problems. It is like questionnaire method but it is filled by enumerator. Enumerators are specially appointed for filling questionnaires; they explain the aim and objective to respondent and fill the answers in the provided space.

Now let us see some differences between questionnaire and schedule method.

Questionnaire Method	Schedule Method
• **Questionnaires are generally sent through mail and no further assistant from sender**	• Schedule is filled by the enumerator or research worker.
• **Is cheaper method**	• Costly requires field workers
• **Non response is high**	• Non response is low
• **In questionnaire method it is not confirmed that the expected respondent have filled the answers**	• In schedule method identity of person is known
• **Very slow method**	• Information is collected well in time
• **No personal contact**	• Direct personal contact
• **It is used only when respondent is educated or cooperative**	• Information can be collected from illiterate also
• **Wider distribution of sample is possible**	• Difficult for wider area
• **Incomplete and wrong information is more**	• Relatively more correct and complete answer is obtained
• **Depend on quality of questionnaire**	• Depend on honesty and competence of enumerator

• **Physical appearance of the questionnaire should be attractive**	• The attractiveness of the questionnaire is not necessary in schedule method
• **Observation method cannot use**	• It is possible to use observation at the time of filling schedule by enumerator

2.3.5 Warranty cards

Post card size cards sent to customers and feedback collected by asking questions on that card.

2.3.6 Pantry Audit

Pantry audit is surveys of consumer goods have at a given time. Pantry audit technique is used to estimate consumption of the basket of goods at the consumer level. Further it is used to know purchase habit of the people (which product, of what brand, etc. the customer uses?). In this type of audit, the investigator collects an inventory of types, quantities and prices of commodities consumed. Thus, in pantry audit data are collected from the examination of customer's pantry, and questions may be asked at the time of audit.

2.3.7 Distributor or Store Audits

Audits are done by distributor or manufacturer's salesperson. It is the process of doing observation or copying information about inventory in retail shops. It is useful method for knowing market share, market size, and effect of advertisement in store promotion.

2.3.8 Panel Method

In this method, data is collected from the same sample respondents at some interval either by mail or personal interview. This is used for studies on:

- Expenditure pattern
- Consumer behaviour
- Effectiveness of advertising
- Voting behaviour and so on.

2.3.9 Experimental method

The experimental method involves manipulating one variable if a change in one variable causes change in another variable. Experiments are generally conducted in order to test the strength of the relationship between variables. We also saw that the researcher is testing the influence of one variable on another; the variable doing the influencing is called the independent variable while the other being influenced is called the dependent variable. For example, in studying the effect of two different methods for teaching this course, the teaching method would be the independent variable and the students' performance on exams of this course knowledge would be the dependent variable. Furthermore, in studying the effect of advertising cost of products on companies' profit, advertising costs of products is independent variable while companies profit is dependent variable.

2.3.10 Focus group discussion

A focus group discussion (FGD) is a good way to gather together people from similar backgrounds or experiences to discuss a specific topic of interest. The group of participants is guided by a moderator (or group facilitator) who introduces topics for discussion and helps the group to participate in a lively and natural discussion amongst themselves.

The strength of FGD relies on allowing the participants to agree or disagree with each other so that it provides an insight into how a group thinks about an issue, about the range of opinion and ideas, and the inconsistencies and variation that exists in a particular community in terms of beliefs and their experiences and practices.

FGDs can be used to explore the meanings of survey findings that cannot be explained statistically, the range of opinions/views on a topic of interest and to collect a wide variety of local terms. In bridging research and policy, FGD can be useful in providing an insight into different opinions among different parties involved in the change process, thus enabling the process to be managed more smoothly. It is also a good method to employ prior to designing questionnaires.

FGD sessions need to be prepared carefully through identifying the main objective(s) of the meeting, developing key questions, developing an agenda, and planning how to record the session. The next step is to identify and invite suitable discussion participants; the ideal number is between six and eight.

The crucial element of FGD is the facilitation. Some important points to bear in mind in facilitating FGDs are to ensure even participation, careful wording of the key questions, maintaining a neutral attitude and appearance, and summarizing the session to reflect the opinions evenly and fairly. A detailed report should be prepared after the session is finished. Any observations during the session should be noted and included in the report.

FGDs can be also done online. This is particularly useful for overcoming the barrier of distance. While discussion is constrained, the written format can help with reporting on the discussion.

2.4. Methods of data presentation

To describe situations, draw conclusions, or make inferences about events, an investigator must organize the data in some meaningful way. That is to present it in a readily comprehensible condensed form that aids in order to draw inferences from it.

There are two types of methods of data presentation

1. Tabular presentation
2. Diagrammatical or graphical presentation

2.4.1 Tabular presentation

Tabular presentation, also called data tables, is a table facilitates representation of even large amount of data in attractive, easy to read and organized manner. The data is organized in row and columns. Usually, the variables are written in columns while the records are written in rows. A table has contained table number and title.

Definition 2.3
Recorded information in its original collected form, whether it is counts or measurements, is referred to as ***raw data or data set***.

Definition 2.4
Frequency is the number of times a specific value occurs in a frequency distribution.

One way of organizing data in table, to make it easier to understand, is to put it into a frequency distribution. A **frequency distribution** is the organization of raw data in table form by writing the variable values in one column along with the corresponding frequencies in another column. Depending on the nature of variable we do have two types of frequency distribution: categorical (qualitative) frequency distribution and quantitative frequency distribution.

2.4.1.1 Frequency distribution for categorical data

Categorical frequency distribution is used for data that can be placed in specific categories, such as nominal- or ordinal-level data. For example, data such as political affiliation, religious affiliation, economic status, type of industry, type of economic system of government, consumer ratings of products or services, education level, major field of study, employment type, etc. are presented in qualitative frequency distribution.

Example: In a company there are 50 workers. The academic rank of each worker is listed below (PhD = Doctor of Philosophy, MA= Master of art, BA= Bachelor of art, D= Diploma)

PhD	**MA**	**BA**	**BA**	**MA**	**BA**	**D**	**MA**	**D**	**MA**
MA	MA	D	MA	PhD	MA	PhD	BA	MA	PhD
PhD	D	BA	PhD	D	BA	D	D	PhD	MA
MA	BA	MA	BA	BA	MA	D	MA	BA	BA
BA	D	PhD	PhD	MA	D	BA	D	MA	MA

To construct a frequency distribution for the above data set, since the data are categorical, discrete classes can be used. From the raw data there are four academic ranks: PhD, MA, BA, and D. these will be used as the classes for the distribution. The procedure for constructing a frequency distribution for categorical data is given next.

Step 1: Make a table as shown below

Category (1)	Tally (2)	Frequency (3)	Relative frequency (4)	Percent (5)
PhD				
MA				
BA				
D				

Step 2: Tally the data and place the result in column (2).

Step 3: Count the tally and place the result in column (3).

Step 4: Find the relative frequency of each category by using:

$$Relative\ frequency = \frac{category\ frequency}{Total} = \frac{f_i}{\sum_{i=1}^{k} f_i},$$

$$where\ f_i\ \ is\ frequency\ of\ category\ or\ class\ i$$

Relative frequency tells us the weight of each class from the total. Dear learner please notice that the sum of all relative frequencies is always unity and relative frequencies are always between 0 and one inclusively.

Step 5: Find the percentages of values in each class by using;

$$Percentage = \frac{category\ frequency}{Total} * 100\% = \frac{f_i}{\sum_{i=1}^{k} f_i} * 100\%$$

Percentages are not normally part of a frequency distribution, but they can be added since they are used in certain types of graphs such as pie graphs. In addition, a percentage tells us the weight of each class from 100 per cent, and the sum of all percentages is always 100%.

Step 5: Find the totals for columns 3 (frequency), relative frequency (4) and percent (5). The completed table is shown below.

Table 2.1 Academic rank of 50 workers in a company

Education level (1)	Tally (2)	Frequency (3)	Relative frequency (4)	Percentage (%) (5)
PhD	𝍸 ////	9	0.18	18
MA	𝍸 𝍸 𝍸 //	17	0.34	34
BA	𝍸 𝍸 ///	13	0.26	26
D	𝍸 𝍸 /	11	0.22	22
Total		50	1	100

In the company, more workers have master of art rank than any other ranks. This frequency distribution provides the summary of the academic rank of 50 workers in a company.

2.4.1.2 Frequency distribution for Quantitative data

2.4.1.2.1 Ungrouped frequency distribution

Ungrouped frequency distribution is a table of all potential raw scored values that could possibly occur in the data along with their corresponding frequencies. Ungrouped frequency distribution is often constructed for small set of data or a discrete variable.

Constructing an ungrouped frequency distribution

To construct an ungrouped frequency distribution, first find the smallest and the largest raw scores in the collected data. Then make a columnar table of all potential raw scored values arranged in order of magnitude with the number of times a particular value is repeated, i.e., the frequency of that value. To facilitate counting method, tallies can be used.

Example: The following data are the ages in years of 20 women who attend entrepreneur training last year: 30, 41, 39, 41, 32, 29, 35, 31, 30, 36, 33, 36, 32, 42, 30, 35, 37, 32, 30, and 41.

To construct a frequency distribution for these data; arrange the data in ascending order, construct a table, tally the data and complete the frequency column. To further modify the table, you can determine the relative frequency, percentage and cumulative frequencies. For the above data the frequency distribution becomes as follows.

Table 2.2 Frequency distribution of age of women who get entrepreneur training last year

Age	Tally	Frequency	Relative frequency	Percentage (%)	Cumulative frequency			
					Less than cumulative		More than cumulative	
					age	frequency	age	frequency
29	/	1	0.05	5	<= 29	1	>=29	20
30	////	4	0.20	20	<=30	5	>=30	19
31	/	1	0.05	5	<=31	6	>=31	15
32	///	3	0.15	15	<=32	9	>=32	14
33	/	1	0.05	5	<=33	10	>=33	11
35	//	2	0.10	10	<=35	12	>=35	10
36	//	2	0.10	10	<=36	14	>=36	8
37	/	1	0.05	5	<=37	15	>=37	6
39	/	1	0.05	5	<=39	16	>=39	5
41	///	3	0.15	15	<=41	19	>=41	4
42	/	1	0.05	5	<=42	20	>=42	1
Total		20	1	100				

Cumulative Frequency

The total frequency of all values less than or equal/ greater than or equal to a given value is called the cumulative frequency of that value. A table showing the cumulative frequencies is called a cumulative frequency distribution. There are two types of cumulative frequency distributions. **Less than cumulative frequency distribution** is obtained by adding successively the frequencies of all the previous values including the frequency of the value against which it is written while **more than cumulative frequency distribution** is obtained by finding the cumulate total of frequencies starting from the highest to the lowest value.

2.4.1.2.2 Grouped frequency distribution

Grouped frequency distribution is a tabular summary of data showing the number (frequency) of items in each of several non-overlapping classes. Before we see the procedure of

constructing a grouped frequency distribution for quantitative data please notice the following terms.

- Raw data

It is data collected in original form.

- Array

Data arranged, in ascending or descending order.

- Class

It is different, non-overlapping groups of data.

- **Class limits**

Class limits are the values of a variable which typically serve to identify the classes of a frequency distribution. The smaller and the larger values of a class are known as the lower and the upper-class limits, respectively. Class limits must be chosen so that each data item belongs to one and only one class. It may be exclusive or inclusive class limit.

Exclusive class limit

When the upper limit of one class interval is the lower-class limit of the next class interval then the class limit is exclusive class limit. For Example; if the class limit is given in the form 0-10, 10-20, 20-30, 30-40, then the class limit is exclusive class limit form. In this kind of interval if the value of the variable value is 10, 20, or 30, their frequency must be counted respectively in the intervals 10-20, 20-30, 30-40.

Inclusive class limit

When the upper limit of one class interval is included in the class interval itself the form of the class limit inclusive class limit. For example; if the class limit is given in the form 0-10, 11-20, 21-30, 31-40, then the class limit is inclusive class limit form. In this kind of interval if the value of the variable value is 10, 20, or 30 presents in the raw data set their frequencies must be counted respectively in the intervals 0-10, 11-20, and 21-30.

- **Class boundaries**

Class boundaries are used to separate the classes so that there are no gaps in the frequency distribution. The class limits should have the same decimal place value as the data, but the class boundaries should have one additional place value and end in a 5. Class boundaries are class limits in the form of exclusive class limit.

- **Class mark/Mid-point**

The class mark is the point which divides each class into two equal parts. This can be determined by dividing the sum of the two limits or the sum of the two boundaries of each class by 2.

- **Class width**

The class width is the difference between the upper- and lower-class boundaries of any class. It is also the difference between the lower limits of any two consecutive classes or the difference between any two consecutive class marks.

- **Cumulative frequency**

Cumulative frequency is the number of observations less than or equal /more than or equal to a specific value. It may be less than type or more than type. **Greater than type Cumulative frequency** is the total frequency of all values greater than or equal to the lower-class limit or lower-class boundary of a given class while less than type c**umulative frequency** is the total frequency of all values less than or equal to the upper-class boundary of a given class.

- **Relative frequency (rf)**

It is the frequency divided by the total frequency.

- **Relative cumulative frequency (rcf)**

It is the cumulative frequency divided by the total frequency.

Steps for constructing frequency Distribution for quantitative data

1. Find the largest and the smallest value
2. Compute the Range (R) = Largest –Smallest

3. Decided the number of classes (K), usually between 5 and 20 or with the help of Sturges' Rule. According to him, the number of classes can be determined by the formula
   $K = 1 + 3.32l \log_{10} n$ where n is number of observations
4. Find the class width (W) by dividing the range by the number of classes and rounding up to get an integer value. $W = {}^{R}/_{K}$
5. Select a starting point for the lowest class limit. This can be the smallest data value or any convenient number less than the smallest data value. Add consecutively the width to the first lower class limit to get the lower limits
6. Subtract one unit from the lower limit of the second class to get the upper limit of the first class. Then add consecutively the width to each upper limit to get all the upper limits.
7. Find the class boundaries by subtracting 0.5 from each lower-class limit and adding 0.5 to each upper-class limit.
8. Find the numerical frequencies from the tallies.
9. Find the cumulative frequencies (greater than and less than types)
10. (If necessary) Find relative frequencies and cumulative relative frequency (greater than and less than types).

Example: The food services division of an Amusement Park Inc. is studying the amount families who visit the amusement park spend per day on food and drink. A sample of 40 families who visited the park yesterday revealed they spent the following amounts in dollar.

77	**18**	**63**	**84**	**38**	**54**	**50**	**59**	**54**	**56**
36	26	50	34	44	41	58	58	53	51
62	43	52	53	63	62	62	65	61	52
60	60	45	66	83	71	63	58	61	71

To construct a grouped frequency distribution for these sample data we will follow the following steps:

Step 1. Find largest and smallest value in the raw data set

$$\text{Largest value}(L) = 84, and \text{ Smalest value } (S) = 18$$

Step 2. Find range (R)

$$R = L - S = 84 - 18 = 66$$

Step 3. Find Number of class(K)

$$K = 1 + 3.32 \log_{10} 40 = 6.32 = 7 \text{ (rounding up)}$$

Step 4. Find the class width(W)

$W = {}^{R}/_{K} = {}^{66}/_{7} = 9.43 = 10$ (Rounding up)

Step 5. Take the lowest value as starting point (Lower class limit (LCL) of first class) and by adding the class width successively we get the lower-class limits of the other lower-class limit: 18, 28, 38, 48, 58, 68, 78

Step 6. Find the first-class upper-class limit (UCL). To find it subtract one unit from the lower-class limit of the second class.

$$UCL\ of\ the\ first\ class\ interval = 2^{nd}\ LCL - 1 = 28 - 1 = 27$$

And to get the upper-class limits of the other classes we add class width successively: 27, 37, 47, 57, 67, 77, 87

By combining step 5 and 6 we have

Class limits
18 – 27
28 – 37
38 – 47
48 – 57
58 – 67
68 – 77
78 – 87

Step 7. Find the class boundaries.
To find the class boundaries (CB)
$LCB = LCL - 0.5$, $UCB = UCL + 0.5$

Class boundaries
17.5 – 27.5
27.5 – 37.5
37.5 – 47.5
47.5 – 57.5
57.5 – 67.5
67.5 – 77.5
77.5 – 87.5

Step 8. Tally the data.
Step 9. Count the tallies to get frequency (f)
Step 10. Find cumulative frequency (cf) less than and greater than type
Step 11. Calculate relative frequency $(rf) = {}^{f}/_{n}$ and find the cumulative relative frequency (Crf) less than and greater than type
Combining the last eleven steps the complete grouped frequency distribution is given in the accompanying table.

Table 2.3 Money spent of families visited the Amusement Park Inc. per day

Class limits	Class boundaries	Class Mark	Tally	Frequency (f)	Cf less than type	Cf greater than type	rf	Crf less than type	Crf greater than type
18 – 27	17.5 – 27.5	22.5	//	2	2	40	0.05	0.05	1
28 – 37	27.5 – 37.5	32.5	//	2	4	38	0.05	0.1	0.95
38 – 47	37.5 – 47.5	42.5	/𝍸/	5	9	36	0.125	0.225	0.9
48 – 57	47.5 – 57.5	52.5	/𝍸/ 𝍸//	10	19	31	0.25	0.475	0.775
58 – 67	57.5 – 67.5	62.5	/𝍸/ 𝍸// /	16	35	21	0.4	0.875	0.525
68 – 77	67.5 – 77.5	72.5	//	2	37	5	0.05	0.925	0.125
78 – 87	77.5 – 87.5	82.5	///	3	40	3	0.075	1	0.075

2.4.2 Diagrammatical and Graphical Presentation of data

Diagrammatic and graphic presentations are techniques for presenting data in visual displays of statistical data. If we draw diagrams and graphs on the basis of the data collected, they will easily be understood and appreciated by all. Diagrams and graphs are extremely useful because of the following reasons.

- They are attractive
- They make data simple and intelligible.
- They facilitate comparison possible
- They have universal utility.
- They have a great memorizing effect.

2.4.2.1 Diagrammatical presentation of data

Diagrams are appropriate for presenting discrete and categorical data. The two most commonly used diagrammatic presentation for discrete as well as qualitative data are: Bar chart and Pie chart.

a) Bar chart

A bar chart represents the data by using vertical or horizontal bars whose heights or lengths represent the frequencies of the data.

There are different types of bar charts. The most common are:

- Simple bar chart
- Component bar chart.
- Multiple bar charts.
- Deviation bar chart

Simple bar chart

Simple bar chart can be drawn either on horizontal or vertical base, but bars on horizontal base more common. Bars must be uniform width and space between bars must be equal. It is used to display data on one variable.

Example: The following table represent sales volume in million dollars by product type and year of a given company for three products A, B, C.

Table 2.4 Sales volume by product type and sales year

Product	Sales ($) in 1957	Sales ($) in 1958	Sales ($) in 1959
A	12	14	18
B	24	21	18
C	24	35	54

You can draw a simple bar chart either for sales of single product or sales in each year.

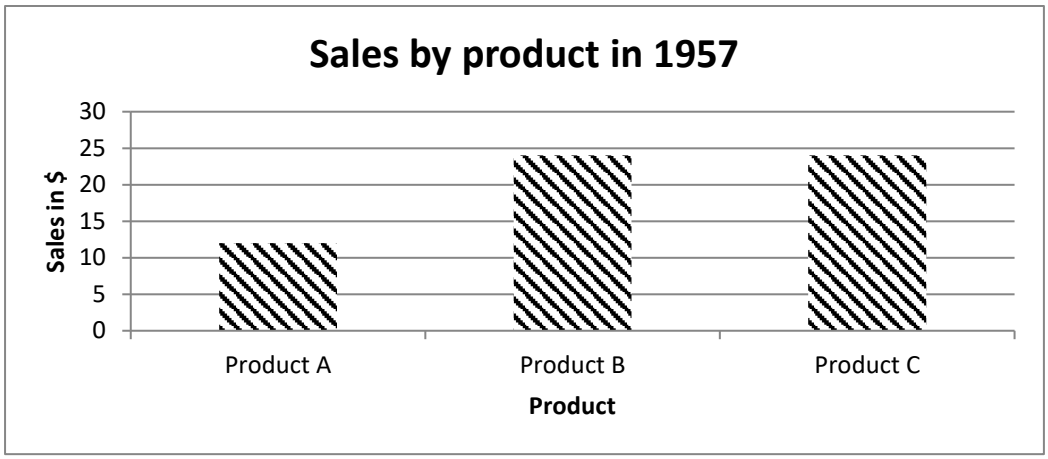


Figure 2.1 Simple bar chart of company sales volume in million dollars in 1957

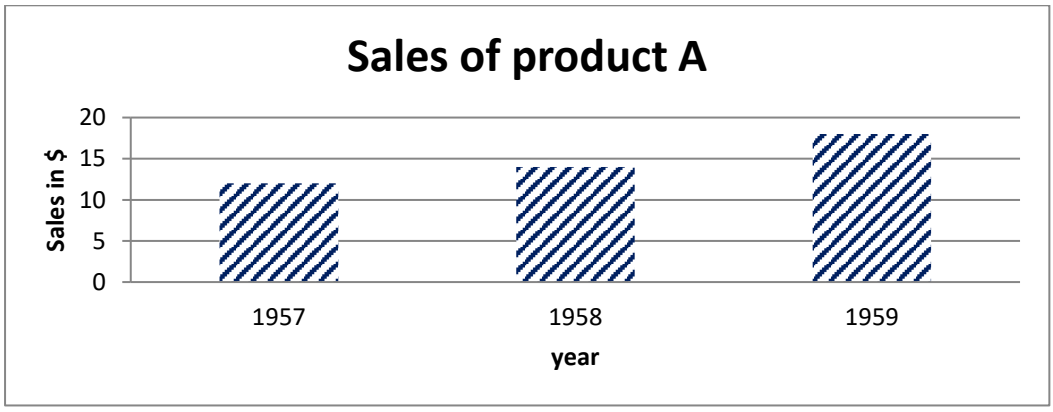


Figure 2.2 Simple bar chart of company sales volume in million dollars for product A by sales year

Deviation bar-diagrams

When the data take both positive and negative values (for instance data on profit, net export, percent change, etc.) deviation bar-diagrams are appropriate.

Example

Table 2.5 represents net profit (in million Birr) of a company in oil sales for five years. Present the data using a suitable bar-diagram.

Table 2.5 Net profit (in million Birr) of a company in oil sales for five years

Year	Profit (in thousands)
1997	12
1998	-5
1999	14
2000	9
2001	-6

The deviation bar-diagram for the data looks like the following.

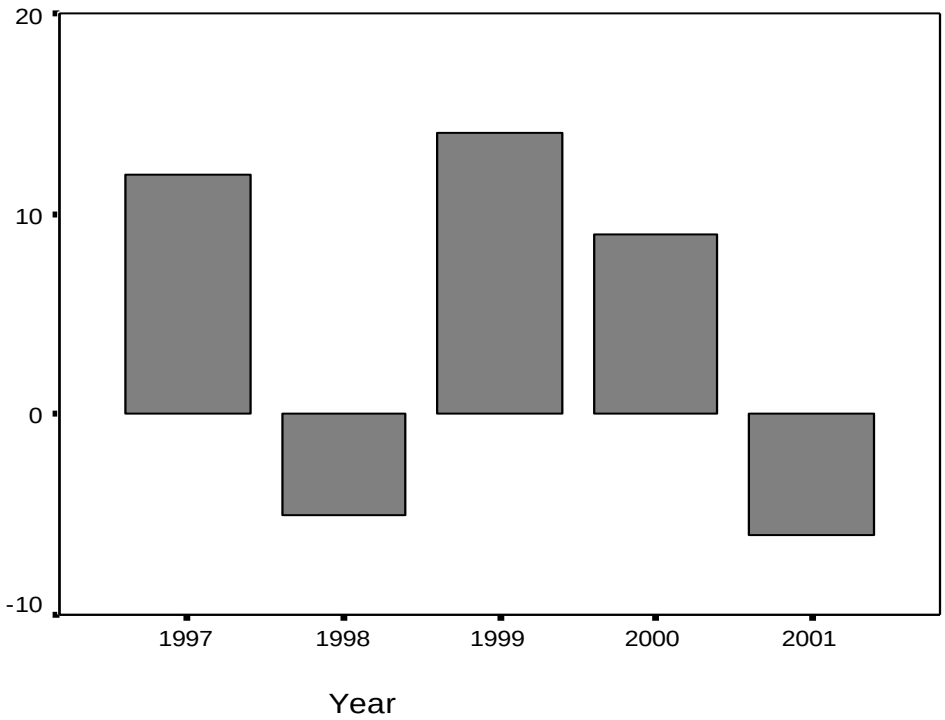


Figure 2.3 Deviation bar diagram of net profit (in million Birr) of a company in oil sales for five years

Component or sub divided Bar chart

In a sub-divided bar diagram, the bar is sub-divided into various parts in proportion to the values given in the data and the whole bar represent the total. The main defect of such a diagram is that all the parts do not have a common base to enable one to compare accurately the various components of the data.

Example: For the data given in Table 2.4 draw component bar chart to represent the sales by product from 1957 to 1959.

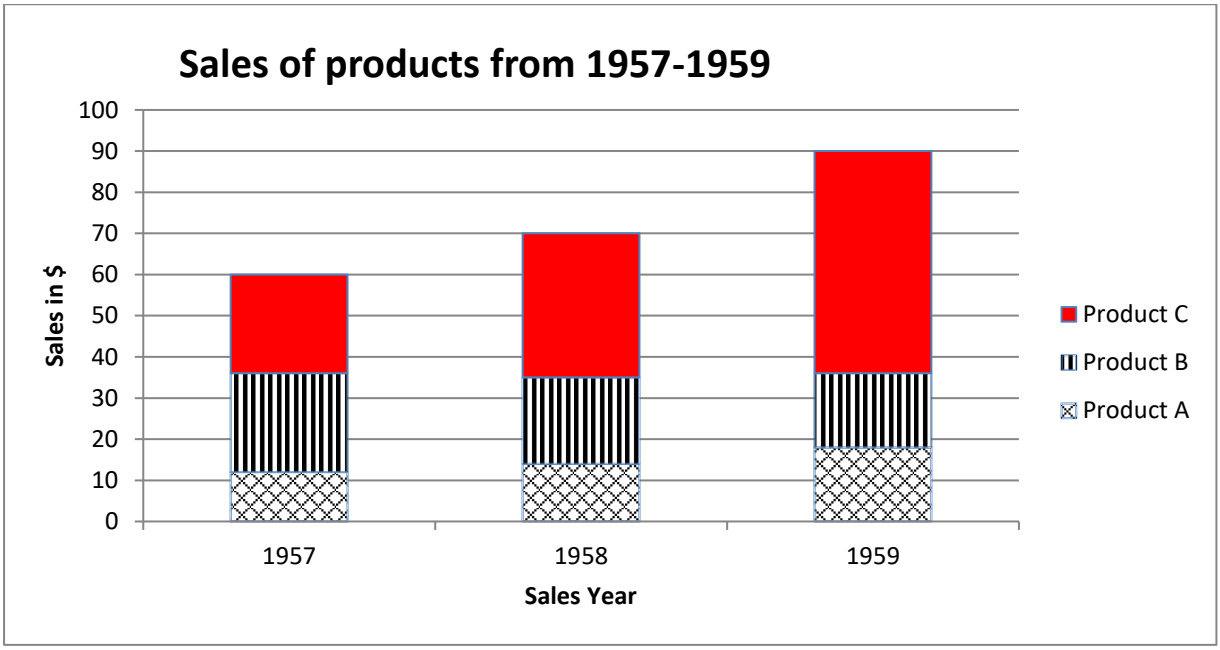


Figure 2.4 Component bar chart of sales volume in million dollars of three products for three years

Multiple bar charts

Multiple bar diagram is used for comparing two or more sets of statistical data. Bars are constructed side by side to represent the set of values for comparison. In order to distinguish bars, they may be either differently coloured or there should be different types of crossings or dotting, etc. An index is also prepared to identify the meaning of different colours or dotting.

Example

From the data given in Table 2.4 draw a multiple bar chart to represent the sales by product from 1957 to 1959.

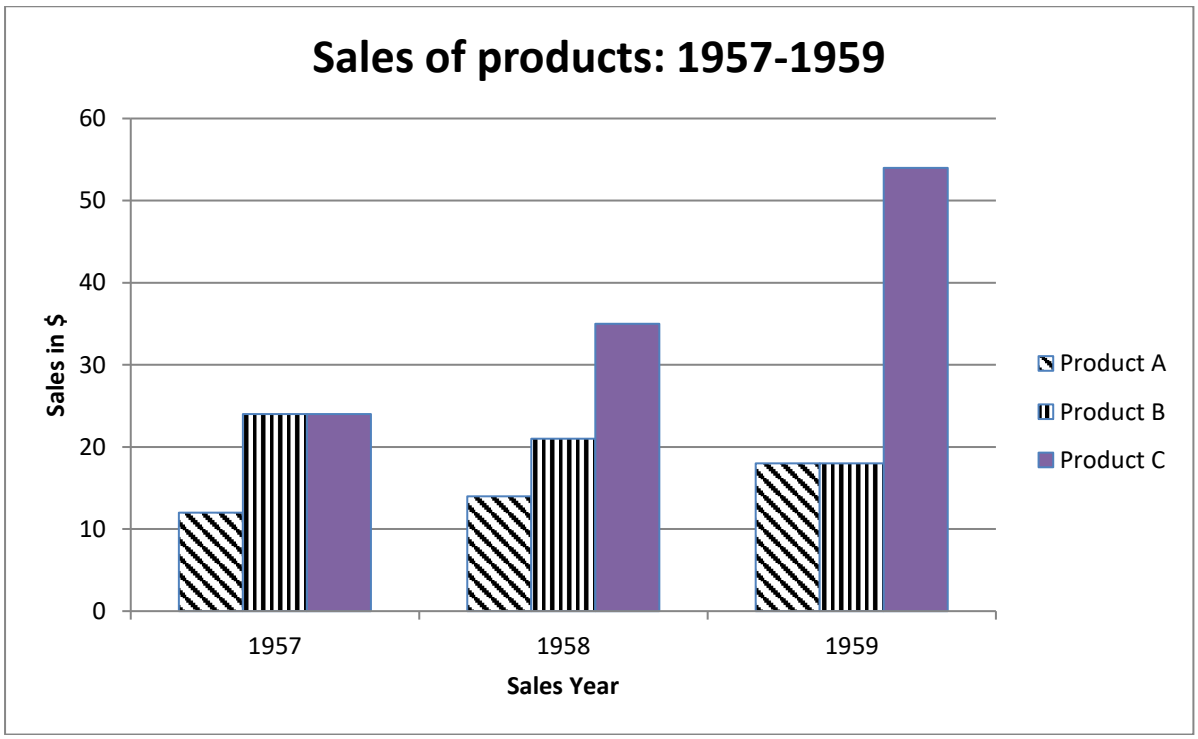


Figure 2.5 Multiple bar chart of sales volume in million dollars of three products

b) Pie Chart

A pie chart is a circle that is divided into sections or wedges according to the percentage of frequencies in each category of the distribution. Since there are 360in a circle, the frequency for each class must be converted into a proportional part of the circle. This conversion is done by using the formula

$Degrees = \frac{frequeny}{Total} * 360^o = \frac{f_i}{n}$, where $f_i$frequency of each category$i$ and $n$is sum of the frequencies of all categories. Each frequency must be converted to percentages by using the formula

$$p_i = \frac{f_i}{n} * 100\% \ \ where\ p_i\ is\ percentage\ of\ category\ i.$$

Example: A small business consultant is investigating the performance of several companies. The sales in 2008 (in thousands of dollars) for the selected companies were given in the accompanying table.

Table 2.6 Sales volume in thousand dollars of five companies in 2008

Company	Sales (in thousands $)
A	255
B	520
C	750
D	420
E	330

To draw a pie chart we follow the following procedures

Step 1: Find the number of degrees for each class.

Step 2: Find the percentage.

Step 3: Draw the pie chart.

Corporation	Sales (in thousands $)	Degree	Percent
A	255	40.4	11.2
B	520	82.3	22.9
C	750	118.7	33.0
D	420	66.5	18.5
E	330	52.2	14.5
Total	2275	360	100

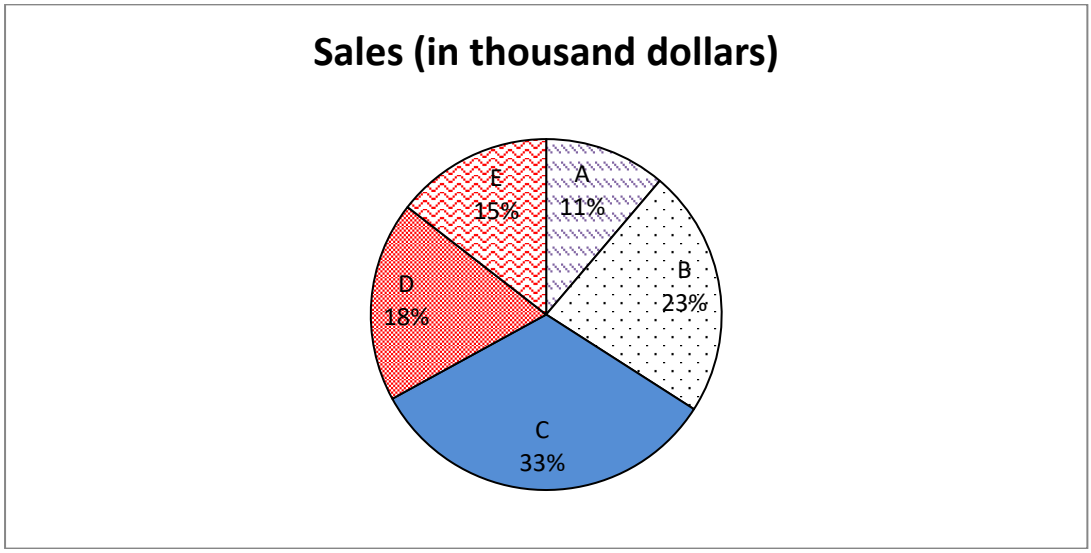


Figure 2.6 Pie chart of Sales volume in thousand dollars of five companies in 2008

2.4.2.2 Graphical Presentation of data

Graphs are appropriate for presenting continuous data. Some of the most commonly used graphical presentations are:

1. Histogram
2. Frequency polygon
3. Cumulative frequency polygon (Ogive)
4. Line graphs
5. Scatter plot

1. Histogram

A histogram is another way of data presentation, which is more suitable for frequency distributions with continuous classes. Histogram is a graph that displays the data by using continuous vertical bars of various heights to represent the frequencies of the classes. In drawing a histogram; we put the class boundaries of each class on the horizontal axis and its respective frequency on the vertical axis. The height of each rectangle represents the frequency of the class while the width of each rectangle represents the width of each class interval. Histograms are sometimes confused with bar charts. A histogram is used for continuous data, where the bins represent ranges of data, while a bar chart is a plot of categorical variables. Some authors recommend that bar charts have gaps between the rectangles to clarify the distinction.

Example: The following table shows the amount of money spent in thousand dollars by a merchandize shop to purchase products for sale. Construct a histogram for the data given in Table 2.7.

Table 2.7 Amount of money spent in thousand dollars by a merchandize shop

Money Spent (Class Limit)	Number of shops	Class Boundary
18-27	2	17.5-27.5
28-37	2	27.5-37.5
38-47	5	37.5-47.5
48-57	10	47.5-57.5
58-67	16	57.5-67.5
68-77	3	67.5-77.5
78-87	2	77.5-87.5

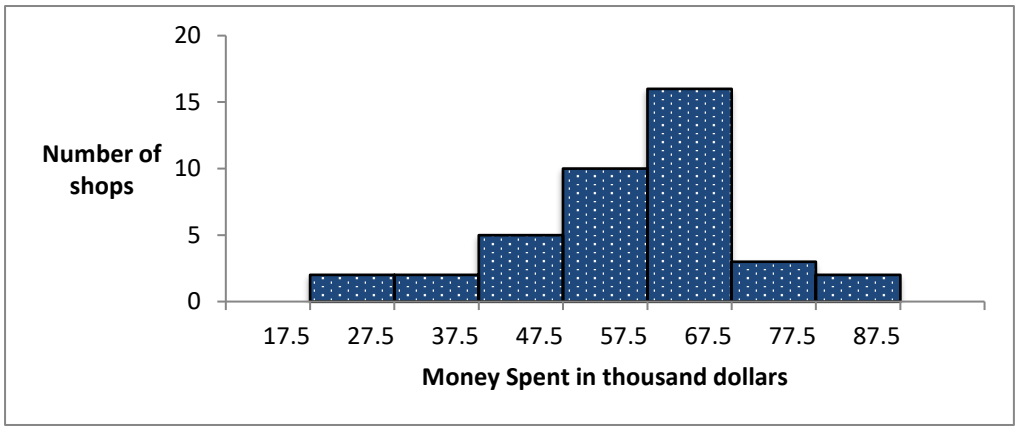


Figure 2.7 Histogram of money spent by a merchandize shop in thousand dollars

2. Frequency polygon

The relevance of presentation of data in the pictorial or graphical form is immense. A frequency polygon is a graphical form of representation of data. It is used to depict the shape of the data and to depict trends. It is usually drawn with the help of a histogram but can be drawn without it as well. A histogram is a series of rectangular bars with no space between them and is used to represent frequency distributions.

Steps to Draw a Frequency Polygon

- Mark the class intervals for each class on the horizontal axis. We will plot the frequency on the vertical axis.
- Calculate the class mark for each class interval. The formula for class mark is:

  Class mark = (Upper limit + Lower limit) / 2

- Mark all the class marks on the horizontal axis. It is also known as the mid-value of every class.
- Corresponding to each class mark, plot the frequency as given to you. The height always depicts the frequency. Make sure that the frequency is plotted against the class mark and not the upper or lower limit of any class.
- Join all the plotted points using a line segment. The curve obtained will be kinked. This resulting curve is called the frequency polygon.

Dear learners note that the above method is used to draw a frequency polygon without drawing a histogram. You can also draw a histogram first by drawing rectangular bars against the given class intervals. After this, you must join the midpoints of the bars to obtain the frequency polygon.

Example: Draw a frequency polygon for the data given in Table 2.7.

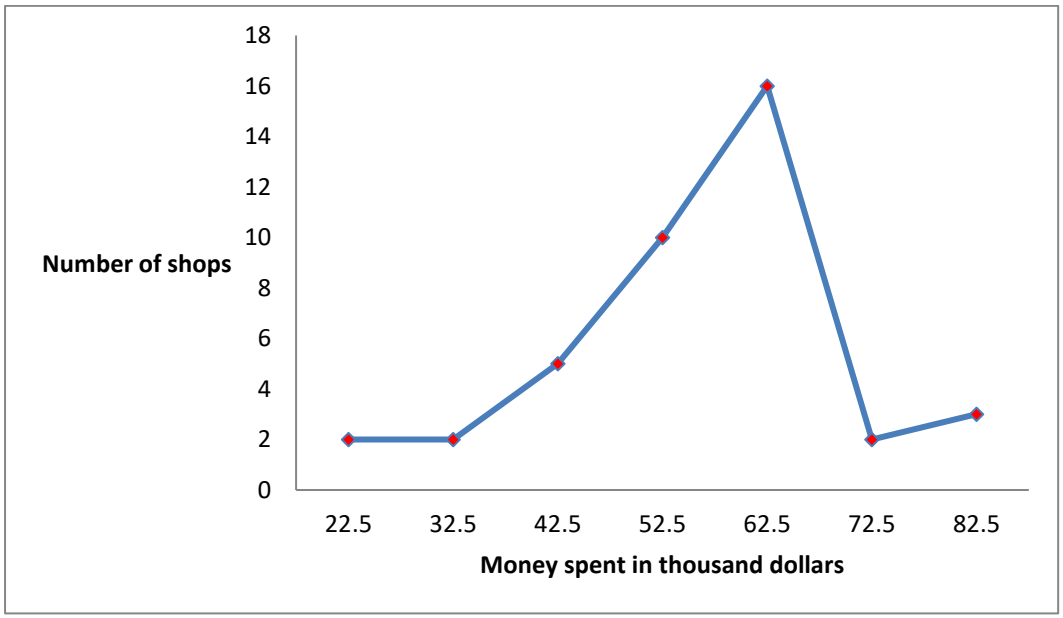


Figure 2.8 Frequency polygon of money spent by a merchandize shop in thousand dollars

3. Cumulative frequency polygon (Ogive)

A curve that represents the cumulative frequency distribution of grouped data on a graph is called a cumulative frequency curve or an ogive. Representing cumulative frequency data on a graph is the most efficient way to understand the data and derive results. There are two types of cumulative frequency curves (or ogives):

- More than type Cumulative Frequency Curve
- Less than type Cumulative Frequency Curve

Steps to plot a More than type Ogive:

- In the graph, put the lower limit on the x-axis
- Mark the cumulative frequency on the y-axis.
- Plot the points (x, y) using lower limits (x) and their corresponding Cumulative frequency (y)
- Join the points by a smooth freehand curve. It looks like an upside-down *S*.

For example, if you want to draw the more than ogive of the data given in Table 2.7, the plot looks like the figure given in Figure 2.9.

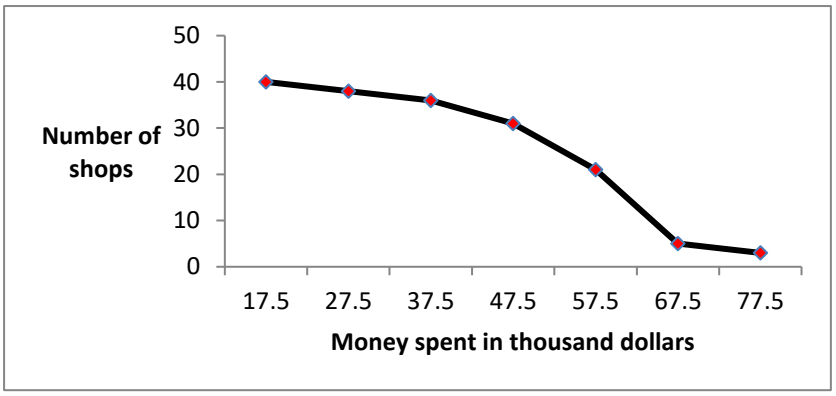


Figure 2.9 More than ogive of money spent by a merchandize shop in thousand dollars

Steps to plot a Less than type Ogive:

- In the graph, put the upper limit on the x-axis
- Mark the cumulative frequency on the y-axis.
- Plot the points (x, y) using upper limits (x) and their corresponding Cumulative frequency (y)
- Join the points by a smooth freehand curve. It looks like an elongated *S*.

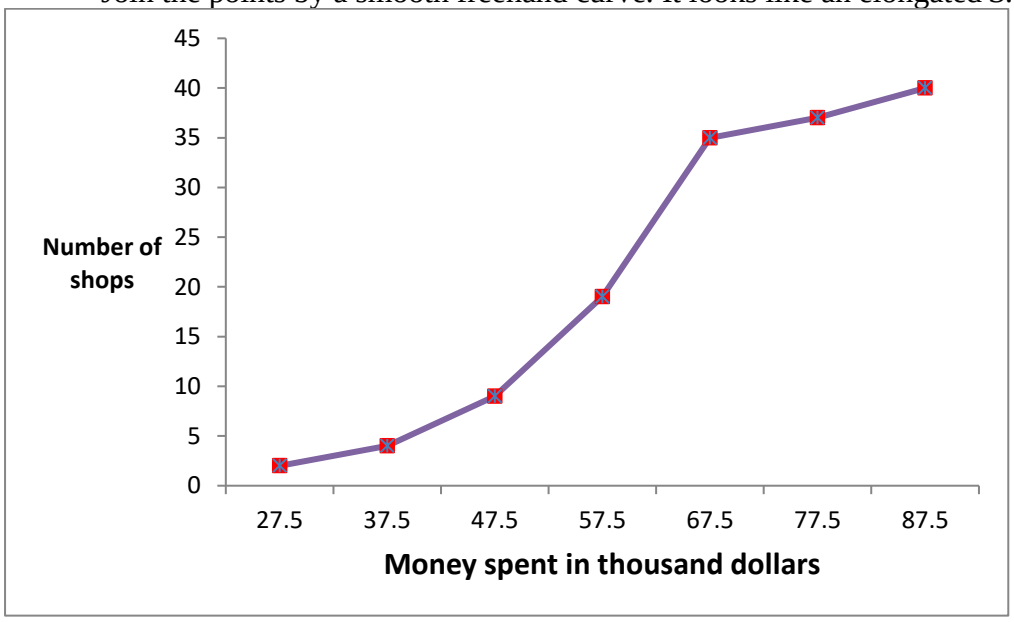


Figure 2.10 Less than ogive of money spent by a merchandize shop in thousand dollars

Further if you want to plot the less than cumulative curve of the data given in Table 2.7, the graph looks like the plot given in Figure 2.10.

Cumulative Graphs can also be used to calculate the Median of given data. If you draw both the curves on the same graph, the point at which they intersect, the corresponding value on the x-axis, represents the Median of the given data set.

4. Line Graph

Line graph is a graph with points connected by lines to show how something changes in value as time goes by, or as something else changes. A line chart or line graph is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments. It is a basic type of chart common in many fields. It is similar to a scatter plot except that the measurement points are ordered (typically by their x-axis value) and joined with straight line segments. A line chart is often used to visualize a trend in data over intervals of time – a time series – thus the line is often drawn chronologically. In these cases, they are known as run charts.

Example: A country spends more money on health care than any other country. To gauge how fast costs are increasing Table 2.8 was produced, which lists the total health care expenditures in the country annually for 1981 to 2009 (expenditures are in $billions).

Table 2.8 Total health care expenditures (in $billions) in the country annually for 1981 to 2009

Year	Health expenditure	Year	Health expenditure
1981	294	1996	1069
1982	331	1997	1125
1983	365	1998	1191
1984	402	1999	1265
1985	440	2000	1353
1986	472	2001	1470
1987	513	2002	1603
1988	574	2003	1732
1989	639	2004	1895
1990	714	2005	2021
1991	782	2006	2152
1992	849	2007	2284
1993	913	2008	2391
1994	962	2009	2486
1995	1017		

To present the data given in Table 2.8 we use line graphs as shown in Figure 2.11.

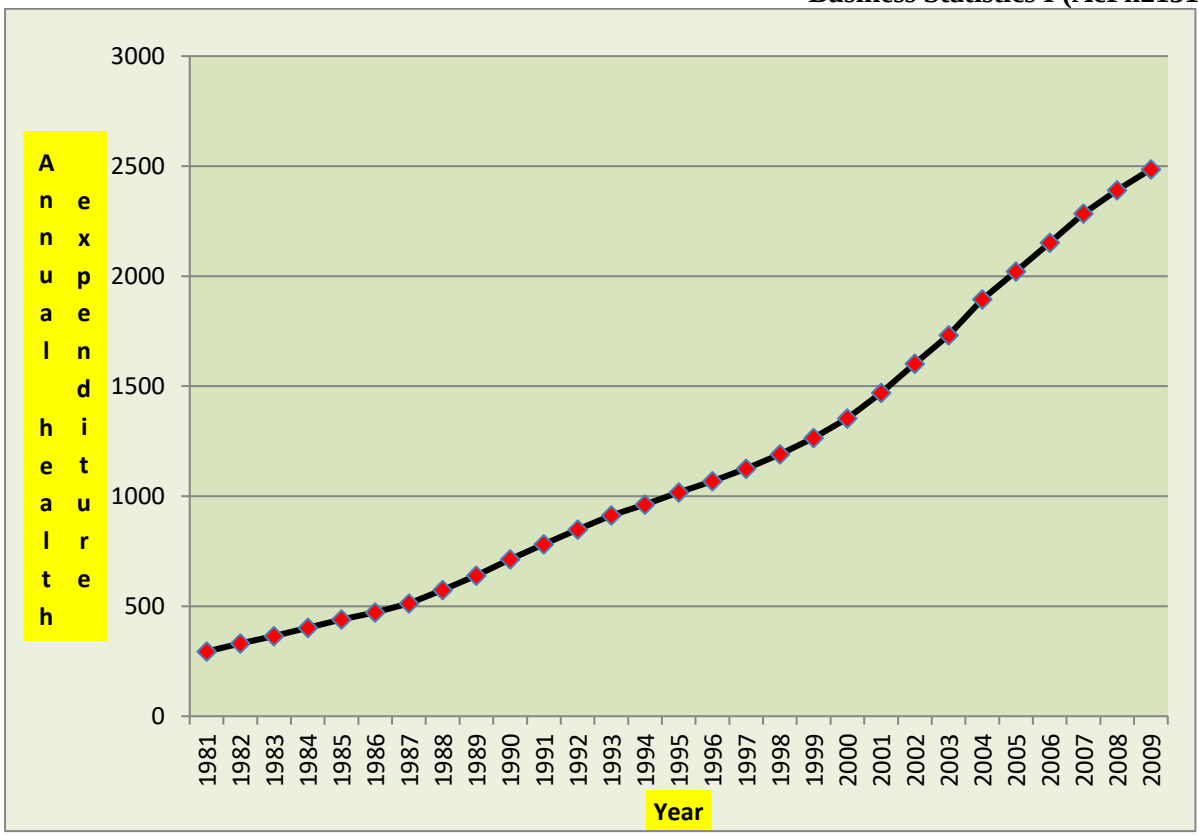


Figure 2.11 Line graph of Total health care expenditures (in $billions) in the country annually for 1981 to 2009

5. Scatter plot

A scatter plot (also called a scatter graph, scatter chart, scatter gram, or scatter diagram) is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. If the points are color-coded, one additional variable can be displayed. The data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis.

A scatter plot can be used either when one continuous variable that is under the control of the experimenter and the other depends on it or when both continuous variables are independent. If a parameter exists that is systematically incremented and/or decremented by the other, it is called the *control parameter* or independent variable and is customarily plotted along the horizontal axis. The measured or dependent variable is customarily plotted along the vertical axis. If no dependent variable exists, either type of variable can be plotted on either axis, or a scatter plot will illustrate only the degree of correlation (not causation) between two variables.

Example1

Because inflation reduces the purchasing power of the dollar, investors seek investments that will provide higher returns when inflation is higher. It is frequently stated that common stocks provide just such a hedge against inflation. The annual percentage rates of return on common stock and annual inflation rates for a recent 10-year period are listed here. Use a scatter graph to depict the relationship between the two variables.

Annual return rate	25	8	6	11	21	-15	12	-1	33	0
Annual inflation rate	4.4	4.2	4.1	4	5.2	5	3.8	2.1	1.7	0.2

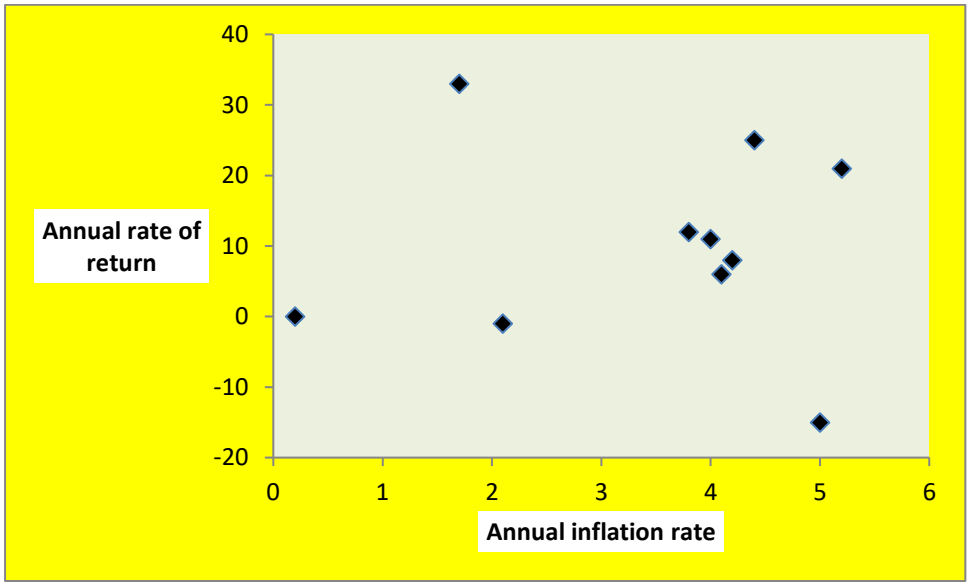


Figure 2.12 Scatter plot of annual rate of return vs. annual inflation rate of a common stock for recent ten years

Example 2

A real estate agent wanted to know to what extent the selling price of a home is related to its size. To acquire this information, he took a sample of 12 homes that had recently sold, recording the price in thousands of dollars and the size in square feet. These data are listed in the accompanying table. Use a graphical technique to describe the relationship between size and price.

Size (ft^2)	price ($1000)
2354	315
1807	229
2637	355
2024	261
2241	234
1489	216
3377	308
2825	306
2302	289
2068	204
2715	265
1833	195

The graphical techniques that describe the relationship between size and price of houses will be shown in the accompanying table.

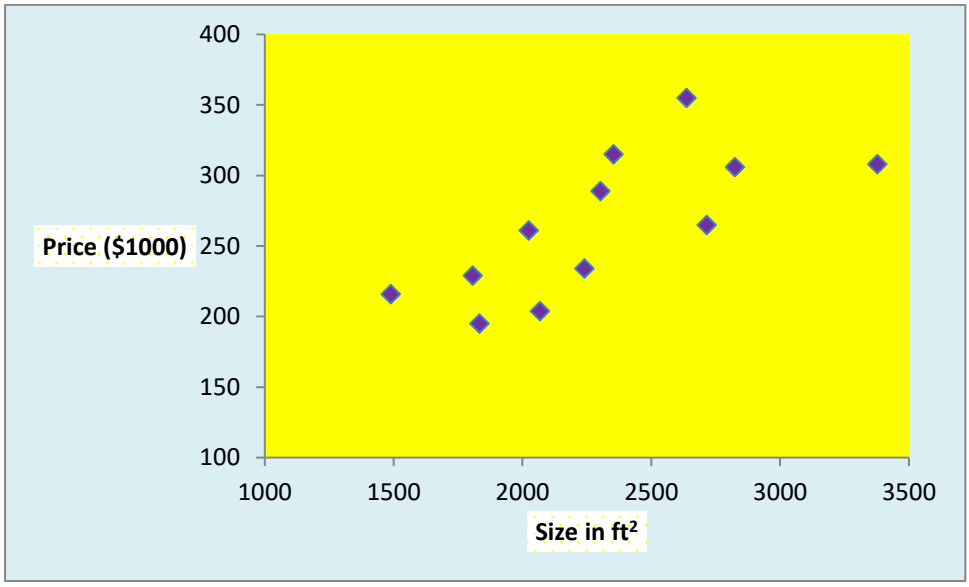


Figure 2.13 Scatter plot of selling price of a home in thousand dollars vs. its size in ft^2

Exercise 2.1

1. Say True if the statement is correct or False if the statement is wrong.
    a) Relative frequency of a variable value is always between 0 and 1.
    b) In schedule method of primary data collection technique, the response rate is high.
    c) For ordinal variable we can compute its average.
    d) Histogram is used to display qualitative data graphically.
    e) Less than ogive is increasing function while more than ogive is decreasing function.
    f) Focus group discussion is one the method for secondary data collection.
2. Classify the following first as qualitative or quantitative data and then classify based on the measurement scale.
    a) Monthly sales volume in cash of a supermarket
    b) Product type sold in a supermarket
    c) Customer ratings of a product
    d) Rank of companies
    e) Number of defective raw materials in a factory
    f) Intelligent quotient of a person
    g) Fund type of a project
    h) Company type
    i) Yearly budget of a company
    j) Size of workers in industry
3. Classify the following as cross-sectional, time series or panel data.
    a) Number of workers in a company in the 12 months of a year.
    b) The unemployment rate in Ethiopia from 1984-2011 E.C by regions.
    c) The number of students registered in this year in Wollo University by the different programs.
    d) The cost of an industry in cash from 2000-2019 G.C.
    e) The profit of Kombolcha textile factory from its establishment till now.
    f) The amount of wheat produced in Ethiopia from 1966 to 2010 E.C.
    g) The number of displaced people by region in 2011 E.C in Ethiopia.
4. In a survey respondents were asked the following. “Last week were you working full-time, part-time, going to school, keeping house, or what”? The responses were

1. Working full-time
2. Working part-time
3. Temporarily not working
4. Unemployed, laid off
5. Retired
6. School
7. Keeping house
8. Other

The responses were recorded using the codes 1, 2, 3, 4, 5, 6, 7, and 8, respectively. The first 150 observations are listed here.

2	**8**	**4**	**5**	**1**	**1**	**5**	**1**	**5**	**8**	**5**	**1**	**5**	**4**	**2**
2	2	1	1	5	1	5	5	1	1	5	1	1	1	2
1	7	1	5	1	7	1	1	1	2	7	1	1	5	5
8	8	1	1	1	1	1	1	1	4	1	2	1	1	2
5	2	5	1	1	1	7	1	1	7	4	1	4	1	5
8	1	1	1	7	6	6	5	1	2	1	1	1	1	5
7	1	1	1	7	1	1	1	7	1	2	7	1	4	7
7	1	7	1	1	5	1	7	1	6	4	1	1	5	1
7	1	5	2	5	5	1	1	5	1	4	5	1	1	1
1	6	2	1	5	2	1	2	1	2	1	5	2	1	1

Construct a frequency, relative frequency and percentage distribution for these data and graphically summarize the data by producing a bar chart and a pie chart.

5. The following table displays the energy consumption patterns of country A and Country B. The figures measure the heat content in metric tons (1,000 kilograms) of oil equivalent. Use a graphical technique to display the differences between the sources of energy for the two countries.

Energy source	Country A	Country B
Nonrenewable		
• coal and coal product	55385	1281
• Oil	33185	6275
• Natural gas	20350	5324
Renewable		
• Hydroelectric	1388	1848
• solid biomass	4741	805
• Other (liquid biomass, geothermal, solar and wind)	347	2761
Total	115396	18294

6. The number of items rejected daily by a manufacturer because of defects was recorded for the past 30 days. The results are as follows.

4	**9**	**13**	**7**	**5**	**8**	**12**	**15**	**5**	**7**	**3**	**8**	**15**	**17**	**19**
6	**4**	**10**	**8**	**22**	**16**	**9**	**5**	**3**	**9**	**19**	**14**	**13**	**18**	**7**

Construct a frequency, relative frequency, percentage and cumulative frequency distribution for these data.

7. A large investment firm wants to review the distribution of ages of its stockbrokers. The firm believes that this information can be useful in developing plans to recruit new brokers. The ages of a sample of 40 brokers are shown here.

46	**28**	**51**	**34**	**29**	**40**	**38**	**33**	**41**	**52**
53	**40**	**50**	**33**	**36**	**41**	**25**	**38**	**37**	**41**
36	**50**	**46**	**33**	**61**	**48**	**32**	**28**	**30**	**49**
41	**37**	**26**	**39**	**35**	**39**	**46**	**26**	**31**	**35**

Construct a grouped frequency, relative frequency, percentage and cumulative frequency distribution for these data. Draw a histogram, frequency polygon, less than and more than ogive curves for these data.

8. The average daily oil consumption and production (thousands of barrels) of Country A is shown for the years 1973 to 2011. Use a line graph technique to describe these figures. What does the graph tell you?

Year	**1973**	**1974**	**1975**	**1976**	**1977**	**1978**	**1979**	**1980**	**1981**	**1982**
Consumption	17,318	16,655	16,323	17,460	18,443	18,857	18,527	17,060	16,061	15,301
Production	9209	8776	8376	8132	8245	8706	8551	8597	8572	8649
Year	**1983**	**1984**	**1985**	**1986**	**1987**	**1988**	**1989**	**1990**	**1991**	**1992**
Consumption	15,228	15,722	15,726	16,277	16,666	17,284	17,327	16,988	16,710	17,031
Production	8689	8879	8972	8683	8349	8140	7615	7356	7418	7172
Year	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**
Consumption	17,328	17,721	17,730	18,308	18,618	18,913	19,515	19,699	19,647	19,758
Production	6847	6662	6561	6465	6452	6253	5882	5822	5801	5746
Year	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	
Consumption	20,034	20,731	20,799	20,686	20,684	19496	18771	19178	18949	
Production	5645	5435	5147	5088	5077	4999	5352	5479	5656	

CHAPTER THREE
3. MEASURES OF CENTRAL TENDENCY

Dear learner:

- ☛ What are measures of central tendency? Why measure them? How can you measure them?
- ☛ What is relative standing of data? How can you locate the relative standings of data? Why you need to know the relative standings of data?
- ☛ What is a measure of variability of data? How can you measure variability of data? Why you measure the variability of data?

3.1 Measures of Central Tendency

Dear learner what do you mean by measures of central tendency? For what purpose do you use them? How can you measure them? For what kind of data, you can apply them? What are the positional measures of data? Any way after careful study of this chapter, you should be able to provide brief solutions for the above questions.

In Chapter 2 we discussed how you can gain useful information from raw data by organizing them into a frequency distribution and then presenting the data by using various graphs. In this chapter, we introduce numerical descriptive techniques that allow the statistics practitioner to be more precise in describing various characteristics of a sample or population. These techniques are critical to the development of statistical inference.

Further as we pointed out in chapter 2, arithmetic calculations can be applied to interval and ratio scale data only. Consequently, most of the techniques introduced here may be used only to numerically describe interval and ratio scale data. However, some of the techniques can be used for ordinal data, and one of the techniques can be employed for nominal data. This sub chapter shows the statistical methods that can be used to summarize data. A single value that describes the characteristics of the entire mass of data is called measures of central tendency or average.

Objectives of measuring central tendency

Dear learner after careful study of this chapter, you should be able to do the following:

- get a single value that represent(describe) characteristics of the entire data
- summarizing/reducing the volume of the data
- facilitating comparison within one group or between groups of data
- enable further statistical analysis

Desirable properties of measure of central tendency

Since there are different measures of central tendency there are certain pre-requisites for a good measure of center generally fulfills. We say a measure of central tendency is best if it possesses most of the following.

1. It should be rigidly defined.
2. It should be easy to understand and compute.
3. It should be based on all items in the data.
4. It should be capable of further algebraic treatment.
5. It should have sampling stability/ little as affected by fluctuations of sampling.
6. It should be capable of being used in further statistical computations or processing.

3.1.1 The Use of Summation Notation (∑)

Let a data set consists of a number of observations, represents by $x_1, x_2, \ldots, x_n$ where *n* (the last subscript) denotes the number of observations in the data and x_i is the i^{th} observation. Then the sum of these *n* observations is given as:

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

For instance a data set consisting of six measurements 21, 13, 54, 46, 32 and 37 is represented by x_1, x_2, x_3, x_4, x_5 and x_6 where $x_1 = 21$, $x_2 = 13$, $x_3 = 54$, $x_4 = 46$, $x_5 = 32$ and $x_6 = 37$.

Their sum becomes $\sum_{i=1}^{6} x_i = 21+13+59+46+32+37=208$.

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

Some Properties of the Summation Notation

1. $\sum_{i=1}^{n} c = n.c$ where c is a constant number.

2. $\sum_{i=1}^{n} b.x_i = b\sum_{i=1}^{n} x_i$ where *b* is a nonzero constant number

3. $\sum_{i=1}^{n} (a + bx_i) = n.a + b\sum_{i=1}^{n} x_i$ where a and b are nonzero constant numbers

4. $\sum_{i=1}^{n} (x_i \pm y_i) = \sum_{i=1}^{n} x_i \pm \sum_{i=1}^{n} y_i$

The sum of the product of the two variables could be written:

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

Example 1: In a company monthly salary for 8 employees are given below

3450 3550 3480 2510 2008 5025 3550 2885

Find

a. The total amount of salaries the company paid in a month for the 8 employees.
b. The total amount of salaries the company paid for five months if it adds 500 allowance for each employee in the fifth month.

Solution: let x_i be the ith employee monthly salary

a. $\sum_{i=1}^{8} x_i = 3450 + 3550 + 3480 + 2510 + 2008 + 5025 + 3550 + 2885 = 26{,}458 birr$

b. $\sum_{i=1}^{8} (500 + 5x_i) = 8(500) + 5(\sum_{i=1}^{8} x_i) = 4000 + 5(26458) = 136{,}290 birr$

Example 2:

Let $\sum_{i=1}^{12} x_i = 26, \quad \sum_{i=1}^{12} y_i = 17, \quad \sum_{i=1}^{12} x_i^2 = 484, \quad \sum_{i=1}^{12} y_i^2 = 362$

Find *I*) $\sum_{i=1}^{12}(4x_i + 3y_i)$, *II*) $\sum_{i=1}^{12} 2x_i(x_i - 7)$

Solution: *I*) $\sum_{i=1}^{12}(4x_i + 3y_i) = 4\sum_{i=1}^{12} x_i + \sum_{i=1}^{12} Y_i = 4(26) + 3(17) = 105$

II) $\sum_{i=1}^{12} 2x_i(x_i - 7) = 2\sum_{i=1}^{12} x_i^2 - 14\sum_{i=1}^{12} x_i = 2(484) - 14(26) = 604$

3.1.2 Types of Measures of Central Tendency

Several types of averages or measures of central tendency can be defined, the most commons are

- the mean
- the mode
- the median

3.1.2.1 The Mean and its Properties

Perhaps the most important measure of location is the mean or average value, for a variable. The mean provides a measure of central location for the data. If the data are for a sample, the mean is denoted by $\bar{x}$; if the data are for a population, the mean is denoted by the Greek letter μ.

There are three types of mean

1. Arithmetic mean
2. Harmonic mean
3. Geometric mean.

3.1.2.1.1 Arithmetic mean

Arithmetic mean is computed by summing the observations and dividing by the number of observations. We label the observations in a sample $x_1, x_2, \ldots, x_n$, where x_1 is the first observation, x_2 is the second, and so on until n, where *n* is the sample size. As a result, the sample mean is denoted $\bar{x}$. In a population, the number of observations is labeled *N*, the population mean is denoted by μ (Greek letter *mu*).

Arithmetic mean

- Population mean: $\mu = \frac{X_1 + X_2 + \cdots + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N}$
- Sample mean: $\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$

Example: In a company monthly salary for 12 employees are given below

3450 3550 3480 2510 2008 5025 3550 2885

Find the arithmetic mean?

Solution $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{26{,}458}{8} = 3307.25$. This means on average the company paid monthly salary of 3307.25 per employee.

Arithmetic Mean for Ungrouped Frequency Distribution

When the data are arranged or given on the form of ungrouped frequency distribution, then the formula for the mean is

$$\bar{X} = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_k x_k}{f_1 + f_2 + \cdots + f_k} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} \text{ Note that } \sum_{i=1}^{k} f_i = n$$

Example

Monthly incomes of firms in thousand Birr are given in the following frequency distribution.

Monthly income	54	64	50	65	95	104	100
Number of firms	3	2	4	4	5	2	3

Compute the mean monthly income for these data.

Solution:

$$\bar{X} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{54 * 3 + 64 * 2 + \cdots + 100 * 3}{3 + 2 + \cdots + 3} = \frac{1733}{23} = 75.35$$

There for $\bar{X} = 75.35 * 1000\ Birr = 75350\ Birr$

Arithmetic Mean for Grouped Frequency Distribution

If data are given in the form of continuous frequency distribution, the sample mean can be computed as

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$$

Where x_i is the class mark of the i^{th} class; $i = 1, 2, \ldots, k$

f_i= the frequency of the i^{th} class and k = the number of classes

Note that $\sum_{i=1}^{k} f_i = n$ = the total number of observations.

Example: The following table gives the daily wages of laborers in ten Birr. Calculate the average daily wages paid to a laborer.

Wages	11-13	13-15	15-17	17-19	19-21	21-23	23-25
Number of laborers (f_i)	3	4	5	6	6	4	3

Solution:

Class mark (x_i)	12	14	16	18	20	22	24

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{560}{31} = 18.06$$

Therefore, $\bar{X} = 18.06 * 10\ Birr = 180.6\ Birr$

Properties of the Arithmetic Mean

1. The sum of the deviations of the items from their arithmetic mean is zero. i.e.

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

2. The sum of the squares of the deviations of a set of observations from any number, say A, is minimum when A=$\bar{X}$. That is, $\sum(x_i - \bar{x})^2 \leq \sum(x_i - A)^2$
3. When a set of observations is divided into *k* groups and $\bar{x}_1$ is the mean of n_1 observations of group 1, $\bar{x}_2$ is the mean of n_2 observations of group2, …, $\bar{x}_k$ is the mean of n_k observations of group *k* , then the combined mean ,denoted by $\bar{x}_c$, of all observations taken together is given by

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

Example

A company has three branches. The average monthly salary of employees was 8000, 8300 and 7600 respectively for the 1st, 2nd, and 3rd branch. There are 28, 32 and 35 employees in each branch, respectively. Find the mean salary for all employees in the company.

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3}{n_1 + n_2 + n_3} = \frac{28(8000) + 32(8300) + 35(7600)}{28 + 32 + 35} = \frac{755600}{95} = 7954$$

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

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

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

Example: An average net income of 10 firms was calculated to be 65000 Birr per month, but latter, it was discovered that one measurement was misread as 40000 Birr instead of 80000 Birr. Calculate the corrected average net income.

Solution: $\bar{X}_{cr} = \frac{n\bar{X}_{wr} + X_c - X_{wr}}{n} = \frac{10(65000) + 80000 - 40000}{10} = 69000\ Birr$

5. If the mean of $x_1, x_2, ..., x_n$ is $\bar{x}$, then
   a) the mean of $x_1 \pm k, x_2 \pm k, ..., x_n \pm k$ will be $\bar{x} \pm k$
   b) The mean of $kx_1, kx_2, ..., kx_n$ will be $k\bar{x}$.

Weighted Arithmetic Mean

When a proper importance is desired to be given to different data values a weighted mean is appropriate. Weights are assigned to each item in proportion to its relative importance. The weighted mean is a special type of arithmetic mean.

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

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

Example1

A student's final mark in Mathematics, Economics, Statistics and Management are respectively 82, 80, 90 and 70.If the respective credits received for these courses are 3, 5, 3 and 1, determine the approximate average mark the student has got for the courses.

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

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

Average mark of the student for four courses is approximately 82.

Example 2

In a vacancy for a position advertised in an organization, the criteria of selection were work experience, entrance exam, and interview results. The relative importance of these criteria was regarded to be different. The weights of these criteria and scores obtained by 3 candidates (out of 100 in each criterion) are given in the following table. In addition, the selection of a candidate is based on the average result on these criteria.

Criterion	Weight	Candidates		
		Dechasa	Gutema	Engdawork
Work experience	4	70	89	85
Entrance exam	3	78	83	89
Interview result	2	90	92	90

Who is the appropriate candidate for this position advertised based on the criteria?

Solution
We use the weighted mean since the relative importance's of the criterion are different:

Criterion	Weight (W_i)	Candidates					
		Dechasa		Gutema		Engdawork	
		X_i	X_i W_i	X_i	X_iW_i	X_i	X_iW_i
Work experience	4	70	280	89	356	85	340
Entrance exam	3	78	234	83	249	89	267
Interview	2	90	180	92	184	90	180
Total	9	238	694	264	789	264	787

The weighted mean and the simple arithmetic mean for the applicants are as follows:

Applicant	Weighted mean	Simple arithmetic mean
Dechasa	694/9 =77.1	238/3 = 79.33
Gutema	789/9 = 87.67	264/3 = 88
Engdawork	787/9 = 87.44	264/3 = 88

If we use the simple arithmetic mean of the scores, both Gutema and Engdawork have got equal chances to be recruited. However, the relative importance of the criteria is different. So, we have to use the weighted mean for discriminating among the candidates. The weighted mean of the scores obtained by Gutema is larger than the others. So Gutema should be recruited for the job.

Merits of Arithmetic Mean

- Arithmetic mean has a rigidly defined value.
- It is calculated based on all observations.
- Arithmetic mean is simple to calculate and easy to understand.
- It doesn't need arrangement of data in increasing or decreasing order.
- Arithmetic mean is also capable of further algebraic treatment.
- It affords a good standard of comparison.

Demerits of Arithmetic Mean

- It is highly affected by extreme value.
- It can be a number which does not exist in the series.
- Sometimes it gives such results which appear almost absurd. For example, it is likely that we can get an average of '3.6 children' per family.
- It can't be calculated for open-ended classes.

3.1.2.1.2 Geometric Mean

The geometric mean is a measure of central tendency used when data is changing over time. Examples might be the growth of investments, the unemployment rate, the inflation rate, or the change of the gross national product. The geometric mean is given by $\bar{X}_g = GM = \sqrt[n]{x_1 * x_2 * \ldots * x_n}$, and the average rate of growth $r$is given by

$$r = \left(\sqrt[n]{\frac{x_n}{x_1}} - 1 \right) * 100$$

Where x_1 and x_n are the first and the last observation of the n original series and r is the percentage growth rate.

Example1: Consider the growth of an initial investment of $1,000 in a savings account that is deposited for a period of 5 years. The interest rate, which is accumulated annually, is different for each year. The following table gives the interest and the growth of the investment

Year	Interest rate (%)	Growth factor	Value year-end
1	6.0	1.060	1060.00
2	7.5	1.075	1139.50
3	8.2	1.082	1232.94
4	7.9	1.079	1330.34
5	5.1	1.051	1398.19

The geometric mean is

$$\sqrt[5]{1.060 * 1.075 * 1.082 * 1.079 * 1.051} = 1.0693$$

This is an average growth rate (r) of 6.93% per year ($1.0693 - 1 = 0.0693\ or\ 6.93\%$). Thus, the value of the \$1,000 at the end of 5 years will become,

$$X_5 = X_1 * (1 + r)^n = \$1000 * 1.0693^5 = \$1398.19$$

Example 2

The total numbers of registered as unemployed (in thousands) in country Z each year between 2014 and 2017 was:

Year	2014	2015	2016	2017
Unemployment	1513	2395	2770	2984

Suppose we want to determine the annual average rate of growth in unemployment over this four-year period. The actual percentage changes from year to year can be calculated as:

	Percentage change from year to year
2014-2015	58.29%
2015-2016	15.66%
2016-2017	7.72%

The arithmetic mean of these three values (27.22%) is not a suitable measure of mean annual growth rate of unemployment because the above percentages were each calculated using different bases. A more appropriate measure is geometric mean. Therefore, for the above data the geometric mean level of unemployment is

$$\bar{X}_g = \sqrt[4]{X_1 * X_2 * X_3 * X_4} = \sqrt[4]{1513 * 2395 * 2770 * 2984} = 2339.41$$
$$\approx 2340\ per\ annum\ between\ 2014\ and\ 2017$$

And the annual average rate of growth of unemployment over the period in question for country Z is

$$r = \left(\sqrt[n]{\frac{X_{2017}}{X_{2014}}} - 1\right) * 100 = \left(\sqrt[4]{\frac{2984}{1513}} - 1\right) * 100 = 18.51\%\ per\ annum$$

Dear learner if the observations have frequencies you can compute the geometric mean by using the accompanying formula

$$GM = \overline{X}_g = \sqrt[n]{X_1^{f_1} \cdot X_2^{f_2} \ldots X_k^{f_k}}$$

3.1.2.1.3 Harmonic Mean

The harmonic mean is a very specific type of average. It's generally used when dealing with averages of units, like speed or other rates and ratios.

Harmonic mean for

- Population : $\mu_h = \frac{N}{\frac{1}{X_1}+\frac{1}{X_2}+\cdots+\frac{1}{X_N}} = \frac{N}{\sum_{i=1}^{N}\left(\frac{1}{X_i}\right)}$ where N is population size
- Sample : $\bar{x}_h = \frac{n}{\frac{1}{X_1}+\frac{1}{X_2}+\cdots+\frac{1}{X_n}} = \frac{n}{\sum_{i=1}^{n}\left(\frac{1}{X_i}\right)}$ where n is sample size

If the data arranged in the form of frequency distribution its harmonic mean is given by

Harmonic mean for

- Population : $\mu_h = \frac{\sum_{i=1}^{K} f_i}{\frac{f_1}{X_1}+\frac{f_2}{X_2}+\cdots+\frac{f_K}{X_K}} = \frac{N}{\sum_{i=1}^{K}\left(\frac{f_i}{X_i}\right)}$ where N is population size
- Sample : $\bar{x}_h = \frac{\sum_{i=1}^{k} f_i}{\frac{f_1}{X_1}+\frac{f_2}{X_2}+\cdots+\frac{f_k}{X_k}} = \frac{n}{\sum_{i=1}^{k}\left(\frac{f_i}{X_i}\right)}$ where n is sample size

Example: A motorist travels 480km in 3 days. She travels for 10 hours at rate of 48km/hr on 1st day, for 12 hours at rate of 40km/hr on the 2nd day and for 15 hours at rate of 32km/hr on the 3rd day. What is her average speed?

$$\overline{X}_h = \frac{3}{\frac{1}{48}+\frac{1}{40}+\frac{1}{32}} = 39.92$$

Relations among different means

1. $\bar{x} \geq \bar{x}_g \geq \bar{x}_h$
2. For two observations $\sqrt{\bar{x} * \bar{x}_h} = \bar{x}_g$
3. $\bar{x} = \bar{x}_g = \bar{x}_h$ if and only if all observations have equal magnitude

3.1.2.2 The Median and Its Properties

Suppose you are seeking to buy a condominium. Your real estate agent says that the typical price of the units currently available is $110000. Would you still want to look if you have budgeted your maximum purchase price at $75000? You might think they are out of your price range. However, checking the individual units' prices might change your mind. They are $60000, $65000, $70000, $80000 and $275000. The arithmetic mean price is $110000 as the real estate agent reported but one price ($275000) is pulling the arithmetic mean upward, causing it to be an unrepresentative average. It does seem that a price around $70000 (median price) is a more typical or representative average, and it is. In the case of such data median provides a more valid measures of location.

The **median** is calculated by placing all the observations in order (ascending or descending). The observation that falls in the middle is the median. The median of a set of items (numbers) arranged in order of magnitude (i.e., in an array form) is the middle value or the arithmetic mean of the two middle values. The median value divides the arrays in to two equal parts: 50% of them have values above it and the rest half percent have values below it. We shall denote the median by $\tilde{x}$ for the sample median and by $\tilde{\mu}$ for the population median. For ungrouped data the median is obtained by

Median

- Population: $\tilde{\mu} = \begin{cases} X_{\frac{N+1}{2}}, & if\ N\ is\ odd \\ \frac{1}{2}\left(X_{\frac{N}{2}} + X_{\frac{N+2}{2}}\right), & if\ N\ is\ even \end{cases}$ $where\ N\ is\ the\ population\ size$
- Sample: $\tilde{X} = \begin{cases} X_{\frac{n+1}{2}}, & if\ n\ is\ odd \\ \frac{1}{2}\left(X_{\frac{n}{2}} + X_{\frac{n+2}{2}}\right), & if\ n\ is\ even \end{cases}$ $where\ n\ is\ the\ sample\ size$

Example: In a company monthly salary for 8 employees are given below

3450 3550 3480 2510 2008 5025 3550 2885

Find the median salary?
Solution: Arrange the data in ascending order

2008 2510 2885 3450 3480 3550 3550 5025

Since n (number of observation)$= 8$ is even, the median is $\frac{3450+3480}{2} = 3465$

Median for grouped data
When the data set given with class intervals in the form

Class interval	Frequency
$X_1 - X_2$	f_1
$X_2 - X_3$	f_2
•	•
•	•
•	•
$X_p - X_{p+1}$	f_p
•	•
•	•
•	•
$X_k - X_{k+1}$	f_k

To find the sample median

i. Arrange the distribution in ascending order
ii. Construct the less than cumulative frequency

Class interval	Frequency	Less than cumulative frequency
$x_1 - x_2$	f_1	F_1
$x_2 - x_3$	f_2	F_2
•	•	•
•	•	•
•	•	•
$x_p - x_{p+1}$	f_p	F_p
•	•	•
•	•	•
•	•	•
$x_k - x_{k+1}$	f_k	$F_k = n$

Let n be number of observations in the distribution

iii. Identify the median class interval

To identify the median class interval first determine the constant $n/2$. Secondly search out the minimum less than cumulative frequency which is greater or equal to the constant $n/2$. Then the class interval corresponding to the minimum less than cumulative frequency which is greater or equal to the constant $n/2$ is the median class interval.

iv. Find the unique median value by adopting the formula

$\tilde{x} = L_{med} + \left(n/2 - C\right)w/f$, where

L_o - be lower class boundary of the median class interval

n - be total number of observations in the distribution

C - be less than cumulative frequency of a class above the median class interval

W- be class width of the median class interval

f - be frequency of the median class interval

Examples: The following table gives the distribution of the weekly wages of laborers of a small firm.

Wages in Birr	No. of workers
126 and below	3
127 – 135	5
136 – 144	9
145 – 153	12
154 – 162	5
163 – 171	5
172 and above	2

a) Find the median weekly wage.

b) Why is the median a more suitable measure of central tendency than the mean in this case?

Solution:

a.

Cf (less than type)
3
8
17
29
34
39
41

The median class is $\frac{n}{2} = \frac{41}{2} = 20.5$ position in Cf (less than type) which is the *4th* class 145-153

$L_{med} = 145.5, W = 9, f_{med} = 12, C = 17$

$$\tilde{X} = L_{med} + \frac{W}{f_{med}}\left(\frac{n}{2} - c\right) = \tilde{X} = 145.5 + \frac{9}{12}(20.5 - 17) = 145.5 + 2.625 = 148.125$$

b. The median a more suitable measure of central tendency than the mean since there is open ended class, it is impossible to calculate the mean.

Merits of median

- It is not influenced by extreme values.
- Rigidly defined
- Median can be calculated even in case of open-ended intervals.

Demerits of median

- It is not capable of further algebraic treatment.
- It is not a good representative of the data if the number of items (data) is small.
- The arrangement of items in order of magnitude is sometimes very tedious process if the number of items is very large.
- It is susceptible to sampling fluctuations.

3.1.2.3 Mode and its properties

The mode or the modal value is the most frequently occurring score/observation in a series and denoted by $\hat{x}$. Note that the mode may not exist in the series or, even if it does exist, it may not be unique. For qualitative variable and ungrouped frequency distribution it is found by inspection.

The following are monthly sales of an Industry in millions Birr for the last year.

Month	Jan	Feb	Mar	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec
Sales	10	12	11	14	12	14	12	16	9	19	10	13

Find the modal monthly sales?

Solution: since 12 occurs 3 times which is the greatest frequency the modal monthly sales is 12 million.

The mode is of interest because that value that occurs most frequently is probably a response that deserves further investigation. For the above example the mode is 12. Thus in forecasting future sales we might conclude that there is a higher probability that sales will be 12 million in any given month.

Mode for grouped data

To find the mode for continuous distributions first you have to check that the distribution must be in ascending order and the class intervals for all classes are equal. When the data set are given with class intervals in the form

Class interval	Frequency
$x_1 - x_2$	f_1
$x_2 - x_3$	f_2
•	•
•	•
•	•
$x_{p-1} - x_p$	f_{p-1}
$x_p - x_{p+1}$	f_p
$x_{p+1} - x_{p+2}$	f_{p+1}
•	•
•	•
•	•
$x_k - x_{k+1}$	f_k

To find the unique modal value

- Find the modal class interval by inspection (the most frequent class interval)
  Let the maximum frequency be f_p then the modal class interval is $x_p - x_{p+1}$

- Compute the unique modal value by interpolation

$$\hat{x} = L_o + \frac{\Delta_1}{\Delta_1 + \Delta_2} * w$$

$$\Delta_1 = f_p - f_{p-1}$$

$$\Delta_2 = f_p - f_{p+1}$$

$$w = x_{p+1} - x_p$$

L_o - be lower class boundary of the modal class interval

f_p - Frequency of the modal class interval

f_{p-1} - Frequency of the class interval above the modal class interval

f_{p+1} - Frequency of the class interval below the modal class interval

w - be class width of the modal class interval

Example: Consider the frequency distributions of an approximation of the annual adjusted gross income in the internal revenue service reports. Find the modal income?

Adjusted gross income	Individual tax returns	Cumulative frequency (less than type)
0-24	15	15
25-49	27	42
50-74	35	77
75-99	12	89
100-124	6	95
125-149	8	103
150-174	4	107
175-199	3	110

Solution:50-74 is the modal class since it has the highest frequency

$\Delta_1 = 35 - 27 = 8$ *and* $\Delta_2 = 35 - 12 = 23$ $L_{\text{mod}} = 49.5$

$$\hat{x} = 49.5 + \left(\frac{8}{8+23}\right) * 25 = 49.5 + 6.45 = 55.95$$

Merits of mode

- Mode is not affected by extreme values.
- Mode can be calculated even in the case of open-end intervals.

- It can be computed for all level of data i.e., ratio, interval, ordinal or nominal.
- We can use it to classify whether the distribution as unimodal, bimodal and multimodal distribution.
- Can **be used for** qualitative **data**

Demerits of mode

- It is not rigidly defined.
- It is not based on all observations
- It is not suitable for further mathematical treatment.
- It is not stable average, i.e., it is affected by fluctuations of sampling.

Dear learner which measures of central tendency do you think is best? With three measures from which to choose, which one should you use?

There are several factors to consider when making our choice of measure of central location. The mean is generally our first selection. However, there are several circumstances when the median is better. The mode is seldom the best measure of central location. One advantage the median holds is that it is not as sensitive to extreme values as is the mean.

To see another advantage of the median over the mean, suppose you and your classmates have written a statistics test and the instructor is returning the graded tests. What piece of information is most important to you? The answer, of course, is *your* mark. What is the next important bit of information? The answer is how well you performed relative to the class. Most students ask their instructor for the class mean. This is the wrong statistic to request. You want the *median* because it divides the class into two halves. This information allows you to identify which half of the class your mark falls into. The median provides this information; the mean does not. Nevertheless, the mean can also be useful in this scenario. If there are several sections of the course, the section means can be compared to determine whose class performed best (or worst).

When the data are interval or ratio, we can use any of the three measures of central location. However, for ordinal and nominal data, the calculation of the mean is not valid. Because the calculation of the median begins by placing the data in order, this statistic is appropriate for ordinal data (in case of odd number of categories). The mode, which is determined by counting the frequency of each observation, is appropriate for nominal and ordinal data. However, nominal data do not have a "center," so we cannot interpret the mode of nominal data in that way. It is generally pointless to compute the mode of nominal data.

CHAPTER 4. MEASURES OF RELATIVE STANDINGS

4.1 Quantiles (Quartiles, Deciles, Percentiles)

Measures of position or relative standings /Quantiles are used to describe the position a specific data value possess in relation to the rest of the data. These are quartiles, deciles and percentiles.

- Quartiles (Q_i)

Quartiles are values of the variable that divides the set of rank data into quarters; each set of the data has three quartiles. The first quartile, Q_1 is a number such that at most 25% of the data smaller in value than Q_1 and at most 75%are larger. The second quartile, Q_2 is a number such that at most 50% of the data smaller in value than Q_2 and at most 50% are larger. The third quartile, Q_3 is a number such that at most 75% of the data smaller in value than Q_3 and at most 25% are larger.

- Deciles (D_i)

Deciles are values of the variable that divides the set of rank data into ten equal parts; each set of the data has nine deciles ($D_i, i=1,2,\cdots,9$).

- Percentiles (P_i , $i=1,2,\cdots,99$)

Percentiles are values of the variable that divides the set of rank data into hundred equal parts; each set of the data has ninety-nine percentiles. The i *th* percentile, P_i is a value such that at most i % of the data are smaller in value than P_i and at most (100- i) % of the data are larger.

☞ **Quantiles for ungrouped data**

To find quantiles for ungrouped frequency distributions

Step1. Let we have n observations, rank the n data, lowest to highest

Step2. Calculate the constant

$$\left\{ \begin{array}{l} \dfrac{ni}{4}, \text{ for } i=1,2,3, \text{ to find } Q_i \\ \dfrac{ni}{10}, \text{ for } i=1,2,\cdots,9, \text{ to find } D_i \\ \dfrac{ni}{100}, \text{ for } i=1,2,\cdots,99, \text{ to find } P_i \end{array} \right\}$$

Step3. Determine the depth, or position of the quantiles ($d(Q_i), d(D_i), d(P_i)$)

If the value calculated in *step2* is an integer A results, then $d(Q_i), d(D_i), d(P_i) = A.5$

If the value calculated in *step2* is a number with a fraction results, then $d(Q_i), d(D_i), d(P_i) = B$, where B is the next largest integer.

Step4. Determine the value of the quantiles

If the value calculated in *step2* is an integer A results, then the value of Q_i, D_i and P_i is half way between the value of the data in the A *th* position and the value of the next data

If the value calculated in *step2* is a number with a fraction results, then the value of Q_i, D_i and P_i is the value of the data in the B *th* position.

Example

The unemployment rate in Region A for the 12 months of 2017 is given in the table below.

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
7.8	8.2	7.6	7.2	6.6	6.6	6.1	5.8	5.9	5.7	6.3	6.9

Find

i. All quartiles
ii. The 5th and 9th deciles
iii. The 50th and 75th percentile

Solution: The data in ascending order are

Month	unemployment rate
Oct	5.7
Aug	5.8
Sep	5.9
Jul	6.1
Nov	6.3
May	6.6
Jun	6.6
Dec	6.8
Apr	7.2
Mar	7.6
Jan	7.8
Feb	8.2

i) Required to find first, second and third quartile

- To find the first quartile, Q_1

$$Constant = \frac{n * i}{4} = \frac{12 * 1}{4} = 3$$

Since 3 is integer depth of Q_1 is

$$d(Q_1) = 3.5$$

The value of Q_1 is half way between 3rd and 4th ordered value

$$Q_1 = \frac{1}{2}\left(3^{rd}\ ordered\ value + 4^{th}\ ordered\ value\right) = \frac{1}{2}(5.9 + 6.1) = 6$$

- To find the second quartile, Q_2

$$Constant = \frac{n * i}{4} = \frac{12 * 2}{4} = 6$$

Since 6 is integer depth of Q_2 is

$$d(Q_2) = 6.5$$

The value of Q_2 is half way between 6th and 7th ordered value

$$Q_2 = \frac{1}{2}\left(6^{th}\ ordered\ value + 7^{th}\ ordered\ value\right) = \frac{1}{2}(6.6 + 6.6) = 6.6$$

- To find the third quartile, Q_3

$$Constant = \frac{n * i}{4} = \frac{12 * 3}{4} = 9$$

Since 9 is integer depth of Q_3 is

$$d(Q_3) = 9.5$$

The value of Q_3 is half way between 9th and 10th ordered value

$$Q_3 = \frac{1}{2}\left(9^{th}\ ordered\ value + 10^{th}\ ordered\ value\right) = \frac{1}{2}(7.2 + 7.6) = 7.4$$

ii) Required to find fifth and ninth decile

- To find the fifth decile, D_5

$$Constant = \frac{n * i}{10} = \frac{12 * 5}{10} = 6$$

Since 6 is integer depth of D_5 is

$$d(D_5) = 6.5$$

The value of D_5 is half way between 6th and 7th ordered value

$$D_5 = \frac{1}{2}\left(6^{th}\ ordered\ value + 7^{th}\ ordered\ value\right) = \frac{1}{2}(6.6 + 6.6) = 6.6$$

- To find the ninth decile, D_9

$$Constant = \frac{n * i}{10} = \frac{12 * 9}{10} = 10.8$$

Since 10.8 is integer depth of D_9 is the next largest integer

$$d(D_9) = 11$$

The value of D_9 is the 11th ordered value

$$D_5 = 11^{th}\ ordered\ value = 7.8$$

iii) Require to find 50th and 75th percentile

- To find the 50th percentile, P_{50}

$$Constant = \frac{n * i}{100} = \frac{12 * 50}{100} = 6$$

Since 6 is integer depth of P_{50} is $d(P_{50}) = 6.5$
The value of P_{50} is half way between 6th and 7th ordered value

$$P_{50} = \frac{1}{2}\left(6^{th}\ ordered\ value + 7^{th}\ ordered\ value\right) = \frac{1}{2}(6.6 + 6.6) = 6.6$$

- To find the 75th percentile, P_{75}

$$Constant = \frac{n * i}{100} = \frac{12 * 75}{100} = 9$$

Since 9 is integer depth of P_{75} is $d(P_{75}) = 9.5$
The value of P_{75} is half way between 9th and 10th ordered value

$$P_{75} = \frac{1}{2}\left(9^{th}\ ordered\ value + 10^{th}\ ordered\ value\right) = \frac{1}{2}(7.2 + 7.6) = 7.4$$

Quantiles for grouped frequency distributions

When the data set given with class intervals in the form

Class interval	Frequency
$X_1 - X_2$	f_1
$X_2 - X_3$	f_2
•	•
•	•
•	•
$X_k - X_{k+1}$	f_k

To find the quantiles

1. Arrange the distribution in ascending order
2. Construct the less than cumulative frequency assuming $x_1 < x_2 < \cdots < x_{k-1} < x_k$

Class interval	Frequency	Less than cumulative frequency
$x_1 - x_2$	f_1	F_1
$x_2 - x_3$	f_2	F_2
•	•	•
•	•	•
•	•	•
$x_k - x_{k+1}$	f_k	$F_k = n$

3. Let n be number of observations in the distribution determine the constant

$$\left\{ \begin{array}{l} \frac{n*i}{4}, \text{ for } i=1,2,3, \text{ to find } Q_i \\ \frac{n*i}{10}, \text{ for } i=1,2,\cdots,9, \text{ to find } D_i \\ \frac{n*i}{100}, \text{ for } i=1,2,\cdots,99, \text{ to find } P_i \end{array} \right\}$$

4. Identify the quantile class interval

To do this search out the minimum less than cumulative frequency which is greater or equal to the constant $\frac{ni}{4}$, $i=1,2,3$ to find the i *th* quartile class interval , $\frac{ni}{10}, i=1,2,3,\cdots 9$ to find the i *th* decile class interval ,and $\frac{ni}{100}$, $i=1,2,3\cdots 99$ to find the i *th* percentile class interval.

Then the class interval corresponding to the minimum less than cumulative frequency which is greater or equal to the constant $\frac{ni}{4}$ is the i *th* quartile class interval, $\frac{ni}{10}$ is the i *th* decile class interval, and $\frac{ni}{100}$ is the i *th* percentile class interval.

5. Use interpolation to find the unique value of i *th* quantile class interval

$$Q_i = L_o + \left(in/4 - C \right) I/f \quad , \quad i = 1,2,3 \text{ for quartile}$$

Where

L_o - be lower class boundary of the i *th* quartile class interval

n - be total number of observations in the distribution

C - be less than cumulative frequency of a class above the i *th* quartile class interval

I - be class width of the i *th* quartile class interval

f - be frequency of the i *th* quartile class interval

$$D_i = L_o + \left({}^{in}/_{10} - C \right) {}^{I}/_{f} \quad , \quad i = 1,2,3 \cdots ,9$$

Where

L_o - be lower class boundary of the i *th* decile class interval

C - be less than cumulative frequency of a class above the i *th* decile class interval

I - be class width of the i *th* decile class interval

f - be frequency of the i *th* decile class interval

$$P_i = L_o + \left({}^{in}/_{100} - C \right) {}^{I}/_{f} \quad , \quad i = 1,2,3 \cdots ,99$$

Where

L_o - be lower class boundary of the i *th* percentile class interval

C - be less than cumulative frequency of a class above the i *th* percentile class interval

I - be class width of the i *th* percentile class interval

f - be frequency of the i *th* percentile class interval

Example: The internal revenue service reports the adjusted gross income (in thousand dollars) on individual tax returns (payers) in millions.

Adjusted gross income	Individual tax returns	Cumulative frequency (less than type)
0-24	15	15
25-49	27	42
50-74	35	77
75-99	12	89
100-124	6	95
125-149	8	103
150-174	4	107
175-199	3	110

Find

a. all quartiles
b. The 5^{th} and 9^{th} deciles
c. 65^{th} and 75^{th} percentiles

Solution:

a. The 1st quartile class is $\frac{n}{4} = \frac{110}{4} = 27.5$ position in Cf (less than type) which is the *2nd* class 25-49

$L_{Q_1} = 24.5, W = 25, f_{Q_1} = 27, C = 15$

$$Q_1 = L_{Q_1} + \frac{W}{f_{Q_1}}\left(\frac{n}{4} - c\right) = 24.5 + \frac{25}{27}(27.5 - 15) = 24.5 + 11.57 = 36.07 = P_{25}$$

The 2nd quartile class is $\frac{2n}{4} = \frac{110}{2} = 55$ position in Cf (less than type) which is the 3^{rd} class 50-74

$L_{Q_2} = 49.5, W = 25, f_{Q_2} = 35, C = 42$

$$Median = Q_2 = L_{Q_2} + \frac{W}{f_{Q_2}}\left(\frac{2n}{4} - c\right) = 49.5 + \frac{25}{35}(55 - 42) = 49.5 + 9.29 = 58.79 = D_5 = P_{50}$$

The 3rd quartile class is $\frac{3n}{4} = \frac{3(110)}{4} = 82.5$ position in Cf (less than type) which is the 3^{rd} class 75-99

$L_{Q_3} = 74.5, W = 25, f_{Q_3} = 12, C = 77$

$$Q_3 = L_{Q_3} + \frac{W}{f_{Q_3}}\left(\frac{3n}{4} - c\right) = 74.5 + \frac{25}{12}(82.5 - 77) = 74.5 + 11.46 = 85.96 = P_{75}$$

b. The 5th decile ($D_5 = Q_2 = Median = 58.79$

The 9th decile class is $\frac{9n}{10} = \frac{9(110)}{10} = 99$ position in Cf (less than type) which is the 6^{th} class 125-149

$L_{D_9} = 124.5, W = 25, f_{D_9} = 8, C = 95$

$$D_9 = L_{D_9} + \frac{W}{f_{D_9}}\left(\frac{9n}{10} - c\right) = 124.5 + \frac{25}{8}(99 - 95) = 124.5 + 12.5 = 137 = P_{90}$$

c. The 65th percentile class is $\frac{65n}{100} = \frac{65(110)}{100} = 71.5$ position in Cf (less than type) which is the 3^{rd} class 50-74

$L_{P_{65}} = 49.5, W = 25, f_{P_{65}} = 35, C = 42$

$$P_{65} = L_{P_{65}} + \frac{W}{f_{P_{65}}}\left(\frac{65n}{100} - c\right) = 49.5 + \frac{25}{35}(71.5 - 42) = 49.5 + 21.07 = 70.57$$

$P_{75} = Q_3$

Interpretations of quantiles

1. Q_i is the value below which $(i \times 25)$ percent of the observations in the series are found where $i = 1,2,3$. For instance Q_3 means the value below which 75 percent of observations in the given series are found?
2. D_i is the value below which $(i \times 10)$ percent of the observations in the series are found (where $i = 1,2, \ldots ,9$). For instance, D_4 is the value below which 40 percent of the values are found in the series.
3. P_i is the value below which i percent of the total observations are found (where $i = 1, 2, 3, \ldots, 99$). For example, 73 percent of the observations in a given series are below P_{73}.

Exercise 3.1

1. Write True if the statement is correct or False if the statement is wrong.

a) Measures of central tendency are inferential statistics that identify the best single value for representing a set of data.
b) The descriptive measure of central tendency that is use to locate central value of qualitative data is mode.
c) The mean is the most frequently occurring score.
d) Quartile divides ordered data set in to four equal parts.
e) In decile analysis the 9th decile, D_9tells us 90% of the values of the variable are below it.

2. Fill in the blank space. Give the appropriate solutions for the following blank space questions.

a) If the harmonic mean and geometric mean of two none zero numbers $X \ and \ Y$ be respectively 20 and 10, then their arithmetic mean is _____
b) The mean of 10 observations is 40. If the value of the eleventh observation is 30, then the average of the eleven observations is _____
c) A firm has three branches: first, second and third branch. Twenty percent of the employees work in the first branch, thirty five percent works in the second, and the rest in the third branch. If there are 500 workers in the first branch, then the numbers of employees in the second and third branch respectively are _____ and_____.
d) The average net income of 10 firms was calculated to be 65000 Dollar per month, but latter, it was discovered that one measurement was misread as 40000 Dollar instead of 80000 Dollar. The corrected average net monthly income of the ten firms was _____.
e) The mean annual salary paid to all employees of a company was 50000 Birr. The mean annual salary paid to male and female employees were 52000 and 42000 Birr respectively. The percentage of males and females employed by the company are respectively _____ and _____

3. The number of sick days due to colds and flu last year was recorded by a sample of 15 adults work in a company. The data are:

5	**7**	**0**	**3**	**15**	**6**	**9**
3	8	10	5	2	0	12

Compute the mean, median, mode, 3rd quartile, 8th decile and 85th percentile.

4. A construction company pays its hourly employees \$16.5, \$19, or \$25 per hour. There are 26 hourly employees, 14 of which are paid at the \$16.5 rate, 10 at the \$19 rate, and the rest at the \$25 rate. What is the mean hourly rate paid for the 26 employees?
5. In June an investor purchased 300 shares of stock at \$20 per share. In August she purchased an additional 400 shares at \$25 per share. In November she purchased an additional 400 shares, but the stock declines to \$23 per share. What is the weighted mean price per share?
6. In 1996 a total of 14968000 taxpayers in Country A filed their individual tax returns electronically. By the year 2006 the number increased to 73300000. What is the mean annual increase for the period?

7. The consumer price index is reported monthly by the Bureau of Labor Statistics. It is reporting the change in prices for a market basket of goods from one period to another. The index for 2005 was 148.2; by 2017 it increased to 200.6. What was the mean annual increase for the period?
8. The return on investment earned by a firm for successive years was: 30%, 20%, -40%, and 200%. What is the mean rate of return on investment for the firm?
9. An investment of $1,000 you made 4 years ago was worth $1,200 after the first year, $1,200 after the second year, $1,500 after the third year, and $2,000 today. Compute the annual rates of return of your investment.
10. A curious business man records the time taken to deliver the milk to the processing plant.

Time in minutes	90-100	80-89	70-79	60-69	50-59	40-49	30-39
No. of days	9	32	43	21	11	3	1

Then compute the mean, median, 75^{th} percentile, 9^{th} decile, 1^{st} quartile and modal time taken for the business man to deliver the milk to the processing plant.

CHAPTER 5. MEASURES OF DISPERSION

5.1 Introduction

The measures of central tendency serve to locate the center of the distribution, but they do not reveal how the items are spread out on either side of the center. This characteristic of a frequency distribution is commonly referred to as dispersion. In a series all the items are not equal. There is difference or variation among the values. The degree of variation is evaluated by various measures of dispersion. Small dispersion indicates high uniformity of the items, while large dispersion indicates less uniformity.

Dear learner why you study about dispersion?

Measures of central tendency; mean, median and mode indicate the central position of a series. They indicate the general magnitude of the data but fail to reveal all the peculiarities and characteristics of the series. In other words, they fail to reveal the degree of the spread out or the extent of the variability in individual items of the distribution. This can be explained by certain other measures, known as 'Measures of Dispersion' or Variation. We can understand variation with the help of the following example:

	Series I	Series II	Series III
	10	2	10
	10	8	12
	10	20	8
Sum	30	30	30

In all three series, the value of arithmetic mean is 10. On the basis of this average, we can say that the series are alike. If we carefully examine the composition of the three series, we find the following differences:

1. In case of 1stseries, three items are equal; but in 2nd and3rdseries, the items are unequal and do not follow any specific order.
2. The magnitude of deviation, item-wise, is different for the 1st, 2nd and 3rd series. But all these deviations cannot be ascertained if the value of simple mean is taken into consideration.
3. In these three series, it is quite possible that the value of arithmetic mean is 10; but the value of median may differ from each other. This can be understood as follows;

Series I	Series II	Series III
10	2	10
10 (median)	8 (median)	12 (median)
10	20	8

The value of Median' in 1st series is 10, in 2nd series is8 and in 3rd series is 10. Therefore, the value of the mean and median are not identical.

4. Even though the average remains the same, the nature and extent of the distribution of the size of the items may vary. In other words, the structure of the frequency distributions may differ even though their means are identical.

In general measures of variation are needed for the following basic objectives.

- To judge the reliability of a measure of central tendency
- To compare two or more sets of data with regard to their variability
- To control variability itself like in quality control, body temperature, etc.
- To make further statistical analysis or to facilitate the use of other statistical measures.

Dear learner what is dispersion?

Simplest meaning that can be attached to the word 'dispersion' is a lack of uniformity in the sizes or quantities of the items of a group or series. According to Reiglemen, "Dispersion is the extent to which the magnitudes or quantities of the items differ, the degree of diversity." The word dispersion may also be used to indicate the spread of the data. In all these definitions, we can find the basic property of dispersion as a value that indicates the extent to which all other values are dispersed about the central value in a particular distribution.

Properties of a good measure of Dispersion

Since there are different measures of dispersion there are certain pre-requisites for a good measure of variation generally fulfills. Some of them are:

i. It should be simple to understand.
ii. It should be easy to compute.
iii. It should be rigidly defined.
iv. It should be based on each individual item of the distribution.
v. It should be capable of further algebraic treatment.
vi. It should have sampling stability.
vii. It should not be unduly affected by the extreme items.

Types of Dispersion

The measures of dispersion can be either 'absolute' or "relative". Absolute measures of dispersion are expressed in the same units in which the original data are expressed. For example, if the series is expressed as Birr in a particular subject; the absolute dispersion will provide the value in Birr. The only difficulty is that if two or more series are expressed in different units, the series cannot be compared on the basis of dispersion. 'Relative' or 'Coefficient' of dispersion is the ratio or the percentage of a measure of absolute dispersion to an appropriate average. The basic advantage of this measure is that two or more series can be compared with each other despite the fact they are expressed in different units. Theoretically, 'Absolute measure' of dispersion is better. But from a practical point of view, relative or coefficient of dispersion is considered better as it is used to make comparison between series.

Various measures of dispersions are in use. The most commonly used measures of dispersions are:

Absolute measure

1. Range
2. Interquartile Range
3. Quartile deviation
4. Mean deviation
5. Standard deviation
6. Variance

Relative measure

1. Coefficient of Range
2. Coefficient of Quartile deviation
3. Coefficient of Mean deviation
4. Coefficient of variation
5. Standard score

Dear learner how can you measure Dispersion?
There are two methods of measuring variation in data. These are:

- **Mathematical Methods**

We can study the 'degree' and 'extent' of variation by these methods. In this category, commonly used measures of dispersion are: range, interquartile range, quartile deviation, mean deviation, variance, standard deviation, coefficient of range, coefficient of quartile deviation, and coefficient of mean deviation, coefficient of variation and standard score.

- **Graphic Methods**

When we want to study only the extent of variation, whether it is higher or lesser a Lorenz-curve is used. Lorenz curve is a graph on which the cumulative percentage of total national income (or some other variable) is plotted against the cumulative percentage of the corresponding population (ranked in increasing size of share).

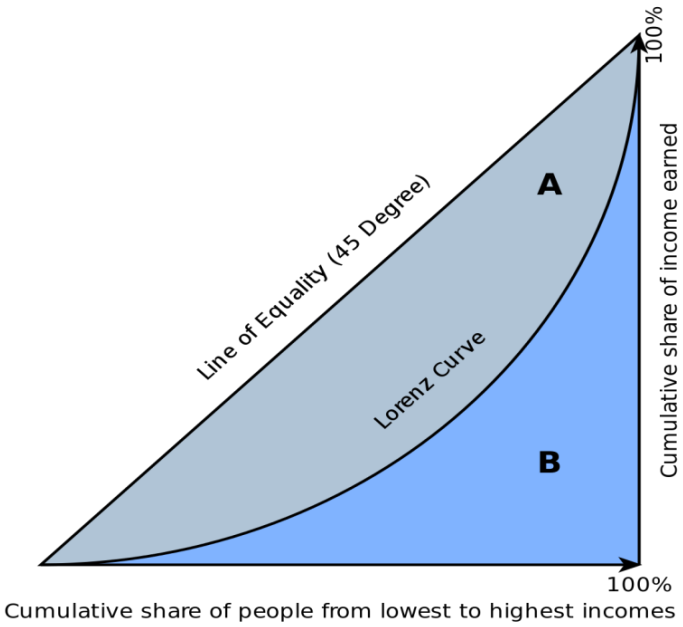


Figure 3.1. Lorenz curve

The extent to which the curve sags below a straight diagonal line indicates the degree of inequality of distribution. It was developed by Max O. Lorenz in 1905 for representing inequality of the wealth distribution.

The curve is a graph showing the proportion of overall income or wealth assumed by the bottom X% of the people, although this is not rigorously true for a finite population (see above). It is often used to represent income distribution, where it shows for the bottom x% of households, what percentage (Y%) of the total income they have. The percentage of households is plotted on the x-axis, the percentage of income on the y-axis. It can also be used to show distribution of assets. In such use, many economists consider it to be a measure of social inequality.

The concept is useful in describing inequality among the size of individuals in ecology and in studies of biodiversity, where the cumulative proportion of species is plotted against the cumulative proportion of individuals. It is also useful in business modeling: e.g., in consumer finance, to measure the actual percentage y% of delinquencies attributable to the x% of people with worst risk scores.

5.2. Mathematical Measures of Variation

5.2.1 Range, Interquartile Range, Mean Deviation
Range and relative range

Range is the simplest possible measure of dispersion and is defined as the difference between the largest and smallest values of the variable.

$$Range\ (R) = L - S. Where$$

$$L = Largest\ value.$$
$$S = Smallest\ value.$$

In case grouped data, range is found by taking the difference between the upper-class boundary of the last class and the lower-class boundaries of the first class or the difference between the class mark of the last class and that of the first class.

$$R = UCB_{last} - LCB_{first}, where$$

$$UCB_{last} = the\ upper\ class\ boundary\ of\ the\ last\ class$$

$$LCB_{first} = the\ lowwer\ class\ boundary\ of\ the\ first\ class$$

or

$$R = CM_{last} - CM_{first}, where$$

$$CM_{last} = the\ class\ mark\ of\ the\ last\ class$$

$$CM_{first} = the\ class\ mark\ of\ the\ first\ class$$

Relative range (RR), also known as *coefficient of range*, is given by

$$Relative\ Range\ (RR) = \frac{L - S}{L + S}, for\ ungrouped\ data$$

$$RR = \frac{UCB_{last} - LCB_{first}}{UCB_{last} + LCB_{first}}, \qquad for\ grouped\ data$$

Properties of Range and Relative Range

- Range and relative range are easy to calculate and simple to understand.
- Both cannot be computed for grouped data with open ended classes.
- They do not tell us anything about the distribution of values in the series.

Example: Find the range and relative range for the monthly salary of ten workers in a certain paint factory given below.

462	480	534	624	498	552	606	588	516	570

Solution:

$$Range\ (R) = L - S = 624 - 462 = 162birr$$

$$RR = \frac{L - S}{L + S} = \frac{624birr - 462birr}{624birr - 462birr} = \frac{162birr}{1086birr} = 0.149$$

Example 2: Find the range and relative range for the following frequency distribution: which shows the distribution of the sales volume in million Birr of firms.

Sales volume in million Birr	Number of firms
93 – 97	2
98 – 102	5
103 – 107	12
108 – 112	17
113 – 117	14
118 – 122	6
123 – 127	3
128 – 132	1

Solution:

$$R = UCB_{last} - LCB_{first} = 132.5 million\ birr - 92.5 million\ birr = 40 million\ birr$$

$$RR = \frac{UCB_{last} - LCB_{first}}{UCB_{last} + LCB_{first}} = \frac{132.5 - 92.5}{132.5 + 92.5} = \frac{40\ million\ birr}{225 million\ birr} = 0.178$$

Inter-quartile range

The difference between the third and the first quartile is called inter-quartile range (IQR)

$$IQR = Q_3 - Q_1, where$$
$$Q_3 = the\ third\ quartile$$
$$Q_1 = the\ first\ quartile$$

Many times, the IQR is reduced in the form of semi-inter quartile range or quartile deviation.

Quartile deviation (semi-inter quartile range) denoted by QD is the half of inter-quartile range.

$$QD = \frac{IQR}{2} = \frac{Q_3 - Q_1}{2}$$

Coefficient of quartile deviation, denoted by CQD, is given by

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

Remark: IQR, Q.D and C.Q.D include only the middle 50% of the observation.

Example: Compute (a) Inter-quartile range, (b) Semi-quartile range, and (c) Coefficient of quartile deviation from the following data:

Farm size (acres)	No. of farms	Cumulative frequency
0-40	394	394
41-80	461	855
81-120	391	1246
121-160	334	1580
161-200	169	1749
201-240	113	1862
241-280	148	2010

Solution

Let Q_1, Q_2, and Q_3 respectively be the 1st , 2nd and 3rd quartile of the data

$Q_1 = 49.9143\ acre$, $Q_2 =$, and $Q_3 = 151.8174\ acres$

a) $Inter\ quartile\ range = IQR = Q_3 - Q_1 = 151.8174 - 49.9143 = 101.9031$ acre

b) Semi quartile range = SQR = quartile deviation = QD = $\frac{Q_3-Q_1}{2} = \frac{151.8174-49.9143}{2} = \frac{101.9031}{2} = 50.9515$ acre

c) $Coefficient\ of\ quartile\ deviation = CQD = \frac{Q_3-Q_1}{Q_3+Q_1} = \frac{151.8174-49.9143}{151.8174-49.9143} = \frac{101.9031}{201.7317} = 0.5051$

The Mean Deviation and Coefficient of Mean Deviation

The mean deviation of a set of items is defined as the arithmetic mean of the values of the absolute deviations from a given average. It tells us how far, on average, all values are from the middle. Depending up on the type of averages used we have different mean deviations.

i) Mean Deviation about the mean

It tells us how far, on average, all values are from the mean. It is denoted by $MD(\bar{x})$ is given by:

$$MD(\bar{x}) = \frac{\sum|x_i - \bar{x}|}{n}$$

For the case of frequency distribution, it is given as:

$$MD(\bar{x}) = \frac{\sum f_i|x_i - \bar{x}|}{n}$$

Where x_i is the class mark of the i^{th} class, f_i is the frequency of the i^{th} class and $n = \sum f_i$.

ii) The mean deviation about the median

It tells us how far, on average, all values are from the median. It is denoted by $MD(\tilde{x})$ and given by

$$MD(\tilde{x}) = \frac{\sum|x_i - \tilde{x}|}{n}$$

For the case of frequency distribution, it is given as:

$$MD(\tilde{x}) = \frac{\sum f_i|x_i - \tilde{x}|}{n}$$

iii) Mean Deviation about the mode

It tells us how far, on average, all values are from the mode. It is denoted by $MD(\hat{x})$ and given by

$$MD(\hat{x}) = \frac{\sum|x_i - \hat{x}|}{n}$$

For the case of frequency distribution, it is given as:

$$MD(\hat{x}) = \frac{\sum f_i|x_i - \hat{x}|}{n}$$

The coefficient of mean deviation *(CMD)* is the ratio of the mean deviation of the observations to their corresponding measure of central tendency. The coefficient of mean deviation about the mean is denoted by $CMD(\bar{x})$ and given by

$$CMD(\bar{x}) = \frac{MD(\bar{x})}{\bar{x}}$$

The coefficient of mean deviation about the median is denoted by $CMD(\tilde{x})$ and given by

$$CMD(\tilde{x}) = \frac{MD(\tilde{x})}{\tilde{x}}$$

And the coefficient of mean deviation about the mode is denoted by $CMD(\hat{x})$ and given by

$$CMD(\hat{x}) = \frac{MD(\hat{x})}{\hat{x}}$$

Example: Compute (a) the mean deviation about the mean, median and the mode, (b) the coefficient of mean deviation about the mean, median and the mode from the following data:

Farm size (acres)	No. of firms
0-40	394
41-80	461
81-120	391
121-160	334
161-200	169
201-240	113
241-280	148

Solution

a) To find the mean deviation about the mean, median and mode first we have to find the mean, median and mode of the given data

Farm size (acres)	Class boundary	Class mark (m_i)	No. of firms (f_i)	Cumulative frequency (F_i)
0-40	0-40.5	20.25	394	394
41-80	40.5-80.5	60.5	461	855
81-120	80.5-120.5	100.5	391	1246
121-160	120.5-160.5	140.5	334	1580
161-200	160.5-200.5	180.5	169	1749
201-240	200.5-240.5	220.5	113	1862
241-280	240.5-280.5	260.5	148	2010

$$mean = \bar{x} = {\sum_{i=1}^{7} f_i m_i}\Big/{\sum_{i=1}^{7} f_i} = 107.4958\ acre$$

$$mode = \hat{x} = 60.0620\ acre$$

$$median = \tilde{x} = 95.8453\ acre$$

Secondly, we have to find the deviation of the class mark from the mean, median and mode and take the absolute values.

In the third step we have to multiply the absolute deviation of the class mark from the mean, median and mode by the corresponding frequencies.

class mark (m_i)	Frequency(f_i)	$\lvert m_i - \bar{x}\rvert$	$f_i\lvert m_i - \bar{x}\rvert$	$\lvert m_i - \hat{x}\rvert$	$f_i\lvert m_i - \hat{x}\rvert$	$\lvert m_i - \tilde{x}\rvert$	$f_i\lvert m_i - \tilde{x}\rvert$
20.25	394	87.2458	34374.8452	39.812	15685.928	75.5953	29784.5482
60.5	461	46.9958	21665.0638	0.438	201.918	35.3453	16294.1833
100.5	391	6.9958	2735.3578	40.438	15811.258	4.6547	1819.9877
140.5	334	33.0042	11023.4028	80.438	26866.292	44.6547	14914.6698
180.5	169	73.0042	12337.7098	120.438	20354.022	84.6547	14306.6443
220.5	113	113.0042	12769.4746	160.438	18129.494	124.6547	14085.9811
260.5	148	153.0042	22644.6216	200.438	29664.824	164.6547	24368.8956
Sum			117550.476		126713.736		115574.91

In the fourth step sum all absolute deviations and find mean deviation about the mean, median and mode as:

$$mean\ deviaation\ about\ mean = MD(\bar{x}) = \frac{\sum_{i=1}^{7} f_i|m_i - \bar{x}|}{\sum_{i=1}^{7} f_i} = \frac{117550.476}{2010} = 58.4828\ acre$$

$$mean\ deviaation\ about\ mode = MD(\hat{x}) = \frac{\sum_{i=1}^{7} f_i|m_i - \hat{x}|}{\sum_{i=1}^{7} f_i} = \frac{126713.736}{2010} = 63.04166\ acre$$

$$mean\ deviaation\ about\ median = MD(\hat{x}) = \frac{\sum_{i=1}^{7} f_i|m_i - \tilde{x}|}{\sum_{i=1}^{7} f_i} = \frac{115574.91}{2010} = 57.49996\ acre$$

b) To find the coefficient mean deviation about the mean, mode and median you can proceed as

$$coefficient\ of\ mean\ deviaation\ about\ mean = CMD(\bar{x}) = \frac{MD(\bar{x})}{\bar{x}} = \frac{58.4828\ acre}{107.4958\ acre} = 0.5441$$

$$coefficient\ of\ mean\ deviaation\ about\ mode = CMD(\hat{x}) = \frac{63.04166\ acre}{60.062\ acre} = 1.04961$$

$$coefficient\ of mean\ deviaation\ about\ median = CMD(\hat{x}) = \frac{57.49996\ acre}{95.8453\ acre} = 0.599925$$

Properties of Mean Deviation and coefficient of mean deviation

- It is easy to understand and compute, and it is based on all observations
- It is not affected very much by the values of extreme value(s).
- It is not capable of further mathematical treatments and it is not a very accurate measure of dispersion.

Activity: Calculate mean and coefficient of mean deviation about mean from the following data:

Profit less than (in million dollars)	Number of firms
10	4
20	10
30	20
40	40
50	50
60	56
70	60

5.2.2 Variance and Standard Deviation
The Variance

We know that variance is a measure of how spread out a data set is. It is calculated as the average squared deviation of each number from the mean of a data set. Like that of mean we have population variance and sample variance.

Population Variance (σ^2)

- **For ungrouped data**

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N} = \frac{\sum_{i=1}^{N} x_i{}^2 - N\mu^2}{N}, \mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

whereμ the population arithmetic mean and N is the total number of observations in the population.

- **For grouped data**

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i(x_i-\mu)^2}{N} = \frac{\sum_{i=1}^{k} f_i x_i{}^2 - N\mu^2}{N}$$

Where μ is the population arithmetic mean, $x_i$is the class mark of the i^{th} class, f_i is the frequency of the i^{th} class and $N = \sum_{i=1}^{k} f_i$ *and k is the number of classes*

Sample Variance (S^2)

- **For ungrouped data**

$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i{}^2 - n\bar{x}^2}{n-1}, \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, Where $\bar{x}$ is the sample arithmetic mean and *n* is the total number of observations in the sample.

- **For grouped data**

$s^2 = \frac{\sum_{i=1}^{k} f_i(x_i-\bar{x})^2}{n-1} = \frac{\sum_{i=1}^{k} f_i x_i{}^2 - n\bar{x}^2}{n-1}$Where $\bar{x}$ is the sample arithmetic mean, $x_i$is the class mark of the i^{th} class, f_i is the frequency of the i^{th} class and $n = \sum f_i$.

Example

The following are data on farm size in hectare and the number of firms in a place. Compute the variance for the following data

Farm size	3	6	9	12	15
No. of firms	1	4	10	3	2

Solution:

Farm size	3	6	9	12	15	total
No. of firms	1	4	10	3	2	20
$f_i x_i$	3	24	90	36	30	183
$x_i - \mu$	-6.15	-3.15	-0.15	2.85	5.85	
$(x_i-\mu)^2$	37.8225	9.9225	0.0225	8.1225	34.2225	
$f_i(x_i-\mu)^2$	37.8225	39.69	0.225	24.3675	68.445	170.55

$$\bar{x} = \frac{\sum f_i x_i}{n} = \frac{183}{20} = 9.15, \quad where\ n = \sum_{i=1}^{5} f_i = 20$$

And $S^2 = \dfrac{\sum f_i \left(x_i - \bar{x}\right)^2}{n-1} = \dfrac{170.55}{19} = 8.976$

Properties of the Variance

- It removes most of the demerits or drawbacks of the measures of dispersion discussed so far.
- Its unit is the square of the unit of measurement of values. For example, if the variable is measured in *kg*, the unit of variance is kg^2.
- It is calculated based on all the observations/data in the series.
- It gives more weight to extreme values and less to those which are near to the mean.

The Standard Deviation

Standard deviation is the positive square root of the variance. The standard deviation is extremely useful in judging the representativeness of the mean. The concept of standard deviation, which was introduced by Karl Pearson, has a practical significance because it is free from all defects, which exists in a range, quartile deviation or average deviation.
Standard deviation is calculated as the square root of average of squared deviations taken from actual mean. It is also called root mean square deviation. The square of standard deviation called 'variance'.

Population Standard Deviation (σ)

- $\sigma = \sqrt{\sigma^2}$, where σ^2 is the population variance.

Sample Standard Deviation (S)

- $S = \sqrt{S^2}$, where S^2 is the sample standard deviation.

Standard deviation is the measure of spread most commonly used in statistical practice when the mean is used to calculate central tendency. Thus, it measures spread around the mean. Because of its close links with the mean, standard deviation can be greatly affected if the mean gives a poor measure of central tendency.

Standard deviation is also influenced by outliers one value could contribute largely to the results of the standard deviation. In that sense, the standard deviation is a good indicator of the presence of outliers. This makes standard deviation a very useful measure of spread for symmetrical distributions with no outliers.

Standard deviation is also useful when comparing the spread of two separate data sets that have approximately the same mean. The data set with the smaller standard deviation has a narrower spread of measurements around the mean and therefore usually has comparatively fewer high or low values. An item selected at random from a data set whose standard deviation is low has a better chance of being close to the mean than an item from a data set whose standard deviation is higher.

Generally, the more widely spread the values are, the larger the standard deviation is. For example, imagine that we have to separate two different sets of exam results from a class of 30 students the first exam has marks ranging from 31% to 98%, the other ranges from 82% to 93%. Given these ranges, the standard deviation would be larger for the results of the first exam.

Standard deviation might be difficult to interpret in terms of how big it has to be in order to consider the data widely spread. The size of the mean value of the data set depends on the size of the standard deviation. When you are measuring something that is in the millions, having measures that are "close" to the mean value does not have the same meaning as when you are measuring the weight of two individuals. For example, a measure of two large companies with a difference of $10,000 in annual revenues is considered pretty close, while the measure of two individuals with a weight difference of

30 kilograms is considered far apart. This is why, in most situations, it is useful to assess the size of the standard deviation relative to the mean of the data set.

Although standard deviation is less susceptible to extreme values than the range, standard deviation is still more sensitive than the semi-quartile range. If the possibility of high values (outliers) presents itself, then the standard deviation should be supplemented by the semi-quartile range.

Example
The following are data of rejected items per operator during a production process

Number of rejects (per operator)	Number of operators (f_i)
21-25	5
26-30	15
31-35	28
36-40	42
41-45	15
46-50	12
51-55	3

Compute the mean and standard deviation.
Solution

Number of rejects (per operator)	Number of operators (f_i)	Class mark(x_i)	$f_i{*}x_i$	$d_i = x_i - mean$	d_i^2	$d_i^2 * f_i$
21-25	5	23	115	-13.96	194.8816	974.408
26-30	15	28	420	-8.96	80.2816	1204.224
31-35	28	33	924	-3.96	15.6816	439.0848
36-40	42	38	1596	1.04	1.0816	45.4272
41-45	15	43	645	6.04	36.4816	547.224
46-50	12	48	576	11.04	121.8816	1462.579
51-55	3	53	159	16.04	257.2816	771.8448
Sum	120		4435			5444.792

$$mean = \bar{x} = \frac{\sum_{i=1}^{7} f_i x_i}{\sum_{i=1}^{7} f_i} = \frac{4435}{120} = 36.9583$$

$$sample\ standard\ deviation = s = \sqrt{\frac{\sum_{i=1}^{7} f_i (x_i - \bar{x})^2}{\sum_{i=1}^{7} f_i - 1}} = \sqrt{\frac{\sum_{i=1}^{7} f_i d_i^2}{\sum_{i=1}^{7} f_i - 1}} = \sqrt{\frac{5444.792}{120 - 1}}$$

$$= \sqrt{\frac{5444.792}{119}} = \sqrt{45.75455} = 6.764211$$

Properties of Standard Deviation

- It is considered to be the best measure of dispersion.
- It is calculated based on all the observations/data in the series. Standard deviation is capable of further algebraic treatment.
- Standard deviation is as such neither easy to calculate nor to understand.
- Similar to the variance, standard deviation gives more weight to extreme values and less to those which are near to the mean.
- Standard deviation is only used to measure spread or dispersion around the mean of a data set.
- Standard deviation is never negative.

- Standard deviation is sensitive to outliers. A single outlier can raise the standard deviation and in turn, distort the picture of spread.
- For data with approximately the same mean, the greater the spread, the greater the standard deviation.
- If all values of a data set are the same, the standard deviation is zero (because each value is equal to the mean).

Special properties of Standard deviations

1. Empirical rule: For normal (symmetric distribution) the following holds. Let x be a value in the data set, $\bar{x}$ be the sample mean and S be the standard deviation of the data set, then
    - Approximately 68.27% of the data values fall within one standard deviation of the mean. i.e. with in $(\bar{x} - S < x < \bar{x} - S)$
    - Approximately 95.45% of the data values fall within two standard deviations of the mean. i.e. with in $(\bar{x} - 2S < x < \bar{x} - 2S)$
    - Approximately 99.73% of the data values fall within three standard deviations of the mean. i.e. with in $(\bar{x} - 3S < x < \bar{x} - 3S)$

2. Chebyshev's Theorem

For any data set, no matter what the pattern of variation, the proportion of the values that fall within k standard deviations of the mean or $(\bar{x} - kS, \bar{x} - kS)$ will be at least $1 - \frac{1}{k^2}$, where k is an number greater than 1 i.e. the proportion of items falling beyond k standard deviations of the mean is at most $\frac{1}{k^2}$.

Example: Suppose a distribution has mean 50 and standard deviation 6. What percent of the numbers are:

a) Between 38 and 62?
b) Between 32 and 68?

Solution:

a. 38 and 62 are at equal distance from the mean,50 and this distance is 12

$$kS = 12$$
$$k = \frac{12}{S} = \frac{12}{6} = 2$$

Applying the above theorem, at least $\left(1 - \frac{1}{2^2}\right) * 100\% = 75\%$ of the number lie between 38 and 62.

b. 32 and 68 are at equal distance from the mean, 50 and this distance is 18

$$kS = 18$$
$$k = \frac{18}{S} = \frac{18}{6} = 3$$

Applying the above theorem, at least $\left(1 - \frac{1}{3^2}\right) * 100\% = 88.89\%$of the number lie between 32 and 68.

5.2.3 Coefficient of Variation and Standard Scores (Z- scores)

Coefficient of Variation

The standard deviation is an absolute measure of dispersion. The corresponding relative measure is known as the *coefficient of variation (CV)*. In probability theory and statistics, the **coefficient of variation** (**CV**), also known as **relative standard deviation** (**RSD**), is a standardized measure of dispersion of a probability distribution or frequency distribution. It is often expressed as a percentage, and is defined as the ratio of the standard deviation to the mean. The coefficient of variation (CV) is a statistical measure of the dispersion of data points in a data series around the mean. The coefficient of variation represents the ratio of the standard deviation to the mean, and it is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from one another. Coefficient of variation is used in such problems where we want to compare

the variability of two or more different series. Coefficient of variation is the ratio of the standard deviation to the arithmetic mean, usually expressed in percent.

$CV = \frac{S}{\bar{x}} \times 100$, where *S* is the standard deviation of the observations.

A distribution having less coefficient of variation is said to be less variable or more consistent or more uniform or more homogeneous. If coefficient of variation (C.V.) is more for one series as compared to the other, there will be more variations in that series, lesser stability or consistency in its composition. If coefficient of variation is lesser as compared to other series, it will be more stable or consistent. Moreover, that series is always better where coefficient of variation or coefficient of standard deviation is lesser.

Example

Suppose we want to compare two firms where the salaries of the employees are given as follows per day:

	Firm A	Firm B
Number of workers	100	100
Mean salary	100 Rs.	80Birr
Standard deviation	40 Rs.	45Birr

Solution: We can compare these firms with the help of coefficient of variation. If we use coefficient of variation, then we shall apply the formula:

$$cv\ of\ firm\ A = \frac{s_A}{\bar{x}_A} * 100 = \frac{40}{100} * 100 = 40\%$$

$$cv\ of\ firm\ B = \frac{s_B}{\bar{x}_B} * 100 = \frac{45}{80} * 100 = 56.25\%$$

Because the coefficient of variation is lesser for firm A than firm B, therefore, firm A is less variable and more stable.

Example: Last semester, students of Economics and Management Departments took *Introduction to Statistics* course. At the end of the semester, the following information was recorded.

Department	**Economics**	**Management**
Mean score	79	64
Standard deviation	23	11

Compare the relative dispersions of the two departments' scores using the appropriate way.

Solution:

Economics Department	Management Department
$CV = \frac{S}{\bar{x}} \times 100$ $= \frac{23}{79} \times 100 = 29.11\%$	$CV = \frac{S}{\bar{x}} \times 100$ $= \frac{11}{64} \times 100 = 17.19\%$

Since coefficient of variation of Economics Department students is greater than that of Management Department students, we can say that there was more dispersion relative to the mean in the distribution of Economics students' scores compared with that of Management students.

Example: The following table illustrates the frequency distribution of yearly profit in millions birr of 100 leather companies' worldwide.

profit ($millions)	60-62	63-65	66-68	69-71	72-74
No. of companies	5	18	42	27	8

Find: a) the variance b) the standard deviation c) the coefficient of variation
d) Calculate mean deviation?

Solution:

Profit ($millions)	60-62	63-65	66-68	69-71	72-74	total
No. of companies(f_i)	5	18	42	27	8	100
class mark (x_i)	61	64	67	70	73	
$f_i x_i$	305	1152	2814	1890	584	6745
$f_i x_i^2$	18605	73728	188538	132300	42632	455803
$\lvert x_i - \bar{x} \rvert$	6.45	3.45	0.45	2.55	5.55	
$f_i \lvert x_i - \bar{x} \rvert$	32.25	62.1	18.9	68.85	44.4	226.5

$\sum_{i=1}^{5} f_i m_i = 6745$, $\sum_{i=1}^{5} x_i^2 = 455803$, $n = \sum_{i=1}^{5} f_i = 100$

and $\bar{x} = \frac{\sum_{i=1}^{5} f_i m_i}{n} = \frac{6745}{100} = 67.45$

a) $S^2 = \frac{\sum_{i=1}^{5} f_i x_i^2 - n\bar{x}^2}{n-1} = \frac{455803 - 100(67.45)^2}{99} = 8.61$

b) $S = \sqrt{S^2} = \sqrt{8.61} = 2.93$

c) $CV = \frac{S}{\bar{x}} * 100 = \frac{2.93}{67.45} * 100 = 4.344$

d) $MD = \frac{\sum_{i=1}^{5} f_i \lvert x_i - \bar{x} \rvert}{n} = \frac{226.5}{100} = 2.265$

The Standard Scores (*Z*-Scores)

A standard score is a measure that describes the relative position of a single score in the entire distribution of scores in terms of the mean and standard deviation. It also gives us the number of standard deviations a particular observation lie above or below the mean in standard deviation units.

Population standard score

$Z = \frac{x - \mu}{\sigma}$ where x is the value of the observation, μ and σ are the mean and standard deviation of the population respectively.

Sample standard score

$Z = \frac{x - \bar{x}}{S}$, where x is the value of the observation, $\bar{x}$ and S are the mean and standard deviation of the sample, respectively.

Example: Two sections were given an exam in a course. The average score was 72 with standard deviation of 6 for *section 1* and 85 with standard deviation of 5 for *section 2*. Student A from *section 1* scored 84 and student B from *section 2* scored 90. Who performed better relative to his/her group?

Solution: Section 1: $\bar{x} = 72$, $S = 6$ and score of student A from Section 1; $x_A = 84$

Section 2: $\bar{x} = 85$, $S = 5$ and score of student B from Section 2; $x_B = 90$

Z-score of student A: $Z = \frac{x_A - \bar{x}_1}{S_1} = \frac{84 - 72}{6} = 2.00$

Z-score of student B: $Z = \frac{x_B - \bar{x}_2}{S_2} = \frac{90 - 85}{5} = 1.00$

From these two standard scores, we can conclude that student A has performed better relative to his/her section students because his/her score is two standard deviations above the mean score of *selection 1* while the score of student B is only one standard deviation above the mean score of *section 2* students.

Exercise 3.2

1. Write True if the statement is correct or False if the statement is wrong.
   a) A measure of central tendency tells us, using a single value, the best representation for an entire set of scores while a measure of variability tells us if the mean is greater than the mode.
   b) The range is measure of dispersion that is influenced most by extreme values.
   c) The interquartile range is descriptive measure of dispersion that is based on the concept of a deviation about the mean.
   d) A graphical measure of variation is used when we want to study only the extent of variation, whether it is higher or lesser.
   e) Coefficient of mean deviation is absolute measures of variation.
2. Fill in the blank space.

a) The annual salaries of the employees of a chain of computer stores produced a mean and standard deviation of $28,000 and $3,000, respectively.
   i. At least 75% of the annual salaries lie between _____ and _____.
   ii. At least 88.9% of the annual salaries lie between _____ and _____.

b) After an analysis of the returns on an investment, a statistics practitioner discovered that the frequency distribution of the data is bell shaped or symmetric with mean and standard deviation are 10% and 8%, respectively. Which of the following are true about the way the returns are distributed?
   i. Approximately 68% of the returns lie between _____ and _____.
   ii. Approximately 95% of the returns lie between _____ and _____.
   iii. Approximately 99.7% of the returns lie between _____ and _____.

c) If quartile range is 24 then quartile deviation is _____.

d) Suppose the mark distribution has mean 50 with standard deviation 6. The percentage of the marks between 32 and 68 is at least _____.

e) Consider a very large number of companies doing business. And suppose the mean profit on the real estate section is 570 million dollar with a standard deviation of 14million dollar.
   i. The standard score for a company in which its profit was 600million dollar is _____.
   ii. A company is told that its z-score on this area is -1.5 then its actual profit on real estate investment was _____

3. The following are the number of tools a firm produced per day.

7	**3**	**2**	**5**	**8**	**11**	**5**	**15**	**3**	**6**

Find the range, the relative range, the inter quartile range, the quartile deviation, mean deviation and coefficient of mean deviation.

4. The following are the daily electricity costs of a firm.

23.8	**11.89**	**15.98**	**26.11**	**31.79**	**39.93**	**12.27**	**40.06**	**21.38**	**18.65**

Find the sample variance and sample standard deviation.

5. Jemilla and Shallom, owners of J&S's Photo Shop, are considering two investment alternatives, asset A and asset B. They are not sure which of these two single assets is better, and they ask SelamYbeltal, a financial planner, for some assistance. Selam knows that the standard deviation,*S*, is the most common single indicator of the risk or variability of a single asset. In financial situations the fluctuation around a stock's actual rate of return and its expected rate of return is called the *risk* of the stock. The standard deviation measures the variation of returns around an asset's mean. Selam obtains the rates of return on each asset for the last 5 years and calculates the means and standard deviations of each asset. Her results are given in Table3.1.

Table3.1 Rates of Return: Asset A and Asset B

	ASSET A	*ASSET B*
Mean Rate of Return	12.2%	12.2%
Standard Deviation in Rate of Return	0.63%	3.12%

Which asset do you think is a riskier investment? Why?

6. Jemilla and Shallom, owners of J&S's Photo Shop, are considering purchasing shares on two stocks, stock A and stock B. From the closing prices of both stocks over the last several months, the standard deviations were found to be considerably different, with standard deviation of stock A is $2.00 and standard deviation of stock B is $8.00. The mean closing price for stock A is $4.00 and the mean closing price for stock B is $80.00. Should stock A be purchased, since the standard deviation of stock B is larger?