CHAPTER ONE

1. INTRODUCTION

1.1 Background of the Subject Statistics

Statistics is a field of mathematics that pertains to data analysis. For the last few centuries, statistics has remained a part of mathematics as the original work was done by mathematicians like Pascal, James Bernoulli, De-Moivre, Laplace, Gauss and others. Till early nineteenth century, statistics was mainly concerned with official statistics needed for the collection of information on revenue, population etc. of a state or kingdom. The science of statistics developed gradually and its field of application widened day by day. In fact, the term statistics is generally used to mean numerical facts and figures.

Statistical thinking has now a day became very essential for different fields of study. Its usefulness has now spread to such diverse fields as physical sciences, engineering, medicine, the social sciences, the life sciences, economics and computer science etc.. For this reason, statistics is now included in the curriculum of many professional and academic study programs. It also helps us in making decision on different problematic situations in our day-to-day life activity.

1.2 Meaning of the Word Statistics

The word statistics seems to have been derived from the Latin word "**status"** or the Italian word "**statista"** or the German word "**Statistik**" each of which means a political state. In ancient times the governments used to collect the information regarding the population and property of wealth of the country- the former enabling the government to have an idea of the manpower of the country (to safeguard itself against external aggression, if any) and the latter providing it a basis for introducing new taxes and levies.

Seventeenth Century saw the origin of vital statistics. Captain John **Graunt** of London known as the father of vital statistics was the first man to study the statistics of birth and death. Computation of mortality table and the calculation of expectation of life at different ages led to the idea of life insurance and the first life insurance institution was founded in London in 1698.

The theoretical development of the so called modern statistics came during the mid - seventeenth century with the introduction of Theory of Probability and Theory of Games and chance. The chief contributors being Pascal, De-Moivre, James Bernoulli, Laplace, Gauss, Sir Francis Galton, Karl Pearson, W. S. Gosset, Helmert, Sir R. A. Fisher.

1.3 Definition of Statistics

Statistics has been defined differently by different authors from time to time. In ancient times statistics was confined only to the affairs of the state but now it embraces almost every sphere of human activity.

Webster defines statistics as classified facts representing the conditions of the people in a state-especially those facts which can be stated in numbers or in any other tabular or classified arrangement. This definition confines statistics only to the data pertaining to the state is inadequate as the domain of statistics is much wider.

Bowley defines statistics as numerical statements of the facts in any department of enquiry placed in relation to each other. He himself defines statistics in three different ways:

I. Statistics may be called as the science of counting.
II. Statistics may rightly be called as the science of averages.
III. Statistics is the science of the measurement of social organism, regarded as a whole in all its manifestations.

The above definitions are inadequate. The first because statistics is not merely confined to the collection of data as other aspects like presentation, analysis and interpretation etc. are also covered by it. The second one is because averages are only a part of the statistical tools used in the analysis of data. These are not only the tools but others being Dispersion, Skewness, Kurtosis, Correlation, Regression analysis, Hypothesis Testing and Estimation etc. The third one is because it restricts the application of statistics to sociology while today the statistics has found its application in almost every field of science and engineering.

Perhaps the best definition seems to be one given by Croxton and Cowden, according to whom statistics may be defined as the science which deals with the collection, analysis and interpretation of numerical data (facts).

Statistics as a subject (field of study): In this sense statistics is defined as the science of collecting, organizing, presenting, analyzing and interpreting numerical data to make decision on the bases of such analysis. (***In singular sense)***

Statistics therefore is defined as the science of collection, compilation, tabulation, analysis and interpretation of quantitative data. It is essentially a branch of applied mathematics i.e. mathematics applied to the observational data. Statistics essentially mean the procedure by which we understand data.

Statistics as a numerical data: In this sense statistics is defined as aggregates of numerically expressed facts (figures) collected in a systematic manner for a predetermined purpose (***Plural sense)***. In this course, we shall be mainly concerned with statistics as a subject, that is, as a field of study.

1.4 Classification of Statistics

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

1. Descriptive statistics and
2. Inferential statistics

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

In short, Descriptive Statistics describes the nature or characteristics of data without making decision or generalization.

Examples of Activities of Descriptive Statistics:

- Recording a student's grades throughout the semester and then finding the average of these grades.
- From sample we have 40% employee suggest positive attitude toward the management of the organization.
- Drawing graphs that show the difference in the scores of males and females.
- Of 50 randomly selected students at computer science department of Unity University 28 of which are female. An example of descriptive statistics is the following statement: "56% of these students are female."

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

Inferential Statistics: Inferential statistics is an area of statistics which consists of generalizing from samples to populations, performing estimations and hypothesis tests, determining relationships among variables, and making predictions etc.

Inferential statistics utilizes sample data to make decision for entire data set (Population).

Examples of Activities of Inferential Statistics:

- There is a definitive relationship between smoking and lung cancer".
- As a result of recent reduction in oil production by oil producing nations , we can expect the price of gasoline to double up in the next year**.**
- As a result of recent survey of public opinion, most Americans are in favor of building additional nuclear power plant.

1.5 Importance of Statistics

The fact that in the modern world statistical methods is universally applicable. It is in itself enough to show how important the science of statistics is. As a matter of fact, there are millions of people all over his world who have not heard a word about statistics and yet who make profuse use of statistical methods in their day-to-day decisions.

Statistical methods are common ways of thinking and hence are used by all types of persons. The importance of statistics in some different disciplines:

(i) Statistics in Computer Science

Statistics play an intrinsic role in computer science and vice versa. Statistics is used for data mining, speech recognition, vision and image analysis, data compression, artificial intelligence, and network and traffic modeling. A statistical background is essential for understanding algorithms and statistical properties that form the backbone of computer science.

Computer scientists tend to focus on data acquisition/cleaning, retrieval, mining, and reporting. They are often tasked with the development of algorithms for prediction and systems efficiency. Focus is also placed on machine learning (an aspect of artificial intelligence), particularly for the purposes of data mining (finding patterns and associations in data for a variety of purposes, such as marketing and finance).

There are a number of ways the roles of statisticians and computer scientists merge; consider the development of models and data mining. Typically, statistical approach to models tends to involve stochastic (random) models with prior knowledge of the data. The computer science approach, on the other hand, leans more to algorithmic models without prior knowledge of the data. Ultimately, these come together in attempts to solve problems.

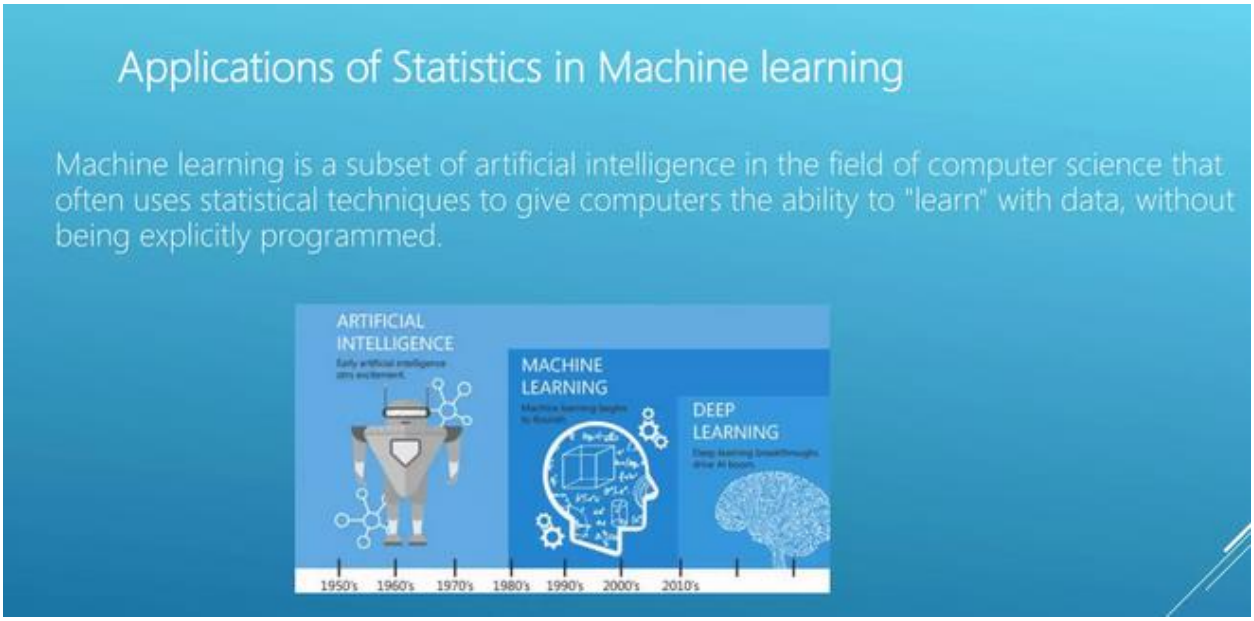
Data mining processes for computer science have statistical counterparts. Consider the following:

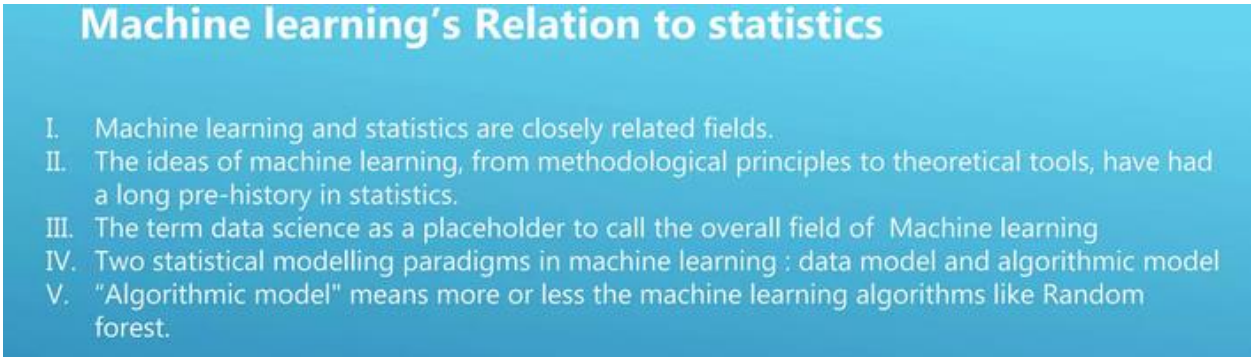
Steps in Computer Science	Steps in Statistics
Data acquisition/enrichment	Experimental design for the collection of data/noise reduction
Data exploration	Discerning the distribution/variability
Analysis and Modeling	Group differences, dimension reduction; prediction; classification
Representation and Reporting	Visualization; communication

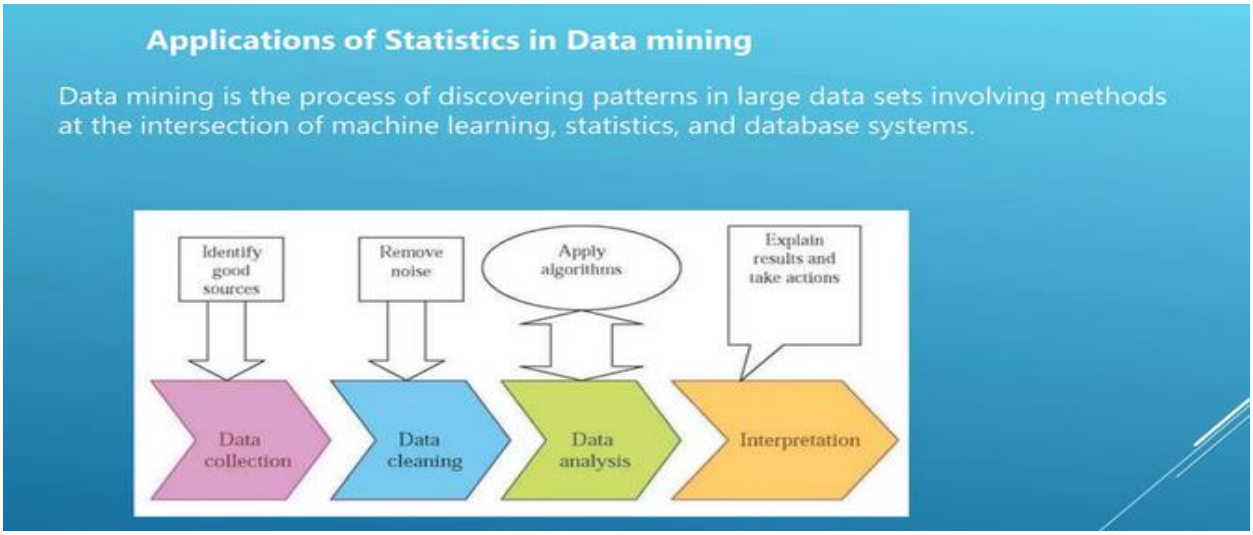
How else is statistics used in computer science? Simulations (used to gain a greater understanding of a variety of systems) are truly a marriage of computing capability and statistics—the use of statistics within programming improves understanding of the underlying

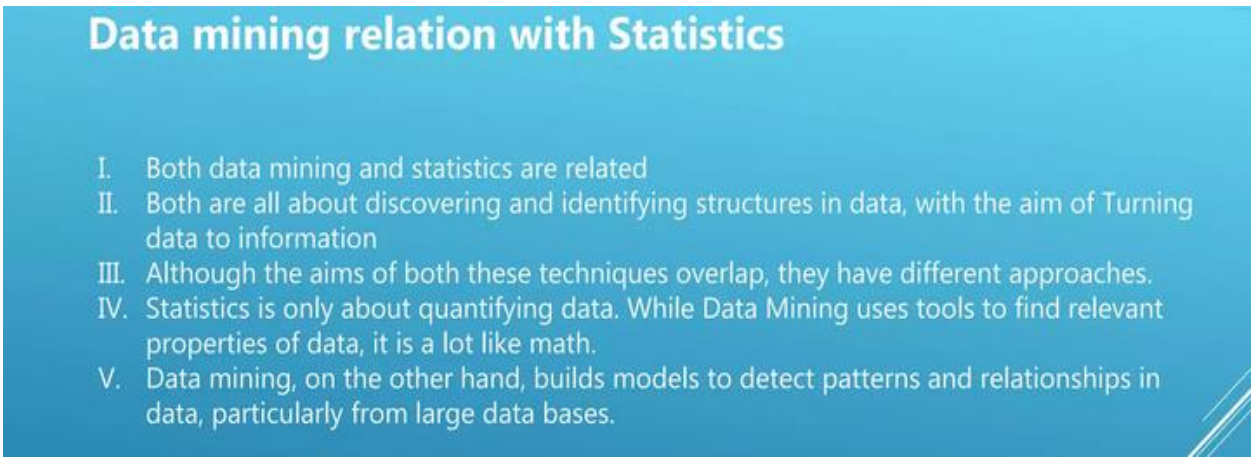
system leading to more meaningful results. Statistics in software engineering leads to more conclusive determinations of quality and optimal performance.

- Application of statistics in Computer Science and Engineering
    - Machine learning
    - Data mining (data management and data analysis)









(ii) Statistics in Planning:

Statistics is indispensable in planning may it be in business, economics, or government level. The modern age is termed as the age of planning and almost all organizations in the government or business or management are resorting to planning for efficient working and for formulating policy decisions.

To achieve this end, the statistical data relating to production, consumption, birth, death, investment, income are of paramount importance. Today efficient planning is a must for almost all countries, particularly the developing economies for their economic development.

(iii) Statistics in Mathematics:

Statistics is intimately related to and essentially dependent upon mathematics. The modern theory of Statistics has its foundations on the theory of probability which in turn is a particular branch of a more advanced mathematical theory of Measures and Integration.

The ever-increasing role of mathematics into statistics has led to the development of a new branch of statistics called Mathematical Statistics. Thus Statistics may be considered to be an important member of the mathematics family. In the words of Connor, Statistics is a branch of applied mathematics which specializes in data.

(iv) Statistics in Research Work:

The job of research work is to present the result of his research before the community. The effect of a variable on a particular problem, under differing conditions, can be known by the research worker only if he makes use of statistical methods.

Statistics are everywhere basic to research activities. To keep alive his research interests and research activities, the researcher is required to lean upon his knowledge and skills in statistical methods. However, in order to draw valid conclusions, a certain standard of accuracy must be maintained. In this way, we can learn about **what is Statistics? Definitions, Meaning, and Characteristics of Statistics.**

1.6 Limitations of Statistics

- The main limitations of statistics are:

(1) Statistics laws are true on average. Statistics are aggregates of facts, so a single observation is not a statistic. Statistics deal with groups and aggregates only.

(2) Statistical methods are best applicable to quantitative data.

(3) Statistics cannot be applied to heterogeneous data.

(4) If sufficient care is not exercised in collecting, analysing and interpreting the data, statistical results might be misleading.

(5) Only a person who has an expert knowledge of statistics can handle statistical data efficiently.

(6) Some errors are possible in statistical decisions. In particular, inferential statistics involves certain errors. We do not know whether an error has been committed or not.

1.7 Stages in Statistical Investigation

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

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

A statistician should be involved at all the different stages of statistical investigation when planning to conduct scientific research. This includes *formulating the problem*, and then *collecting*, *organizing*, *presenting*, *analyzing* and *interpreting* of statistical data. Each step describes as follows.

- **Formulating the problem**: - First research must emanate if there is a problem. At this stage the investigator must be sure to understand the problem and then formulate it in statistical term. Clarify the objectives very carefully. Therefore,
  - Get a clear understanding of the physical background to the situation under study;
  - Clarify the objectives;
  - Formulate the objective in statistical terms
- **Data Collection:** This is a stage where we gather information for our purpose

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

- **Data Organization**: It is a stage where we edit our data .A large mass of figures that are collected from surveys frequently need organization. The collected data involve irrelevant figures, incorrect facts, omission and mistakes. Errors that may have been included during collection will have to be edited .After editing, we may classify (arrange) according to their

common characteristics. Classification or arrangement of data in some suitable order makes the information easer for presentation.

- **Data Presentation:** The organized data can now be presented in the form of tables and diagram. At this stage, large data will be presented in tables in a very summarized and condensed manner. ***The main purpose of data presentation is to facilitate statistical analysis. Graphs and diagrams may also be used to give the data a vivid meaning and make the presentation attractive.***
- **Data Analysis**: This is the stage where we critically study the data to draw conclusions about the population parameter. The purpose of data analysis is to dig out information useful for decision making. Analysis usually involves highly complex and sophisticated mathematical techniques.
- **Data Interpretation:** This is the stage where draw valid conclusions from the results obtained through data analysis. ***Interpretation means drawing conclusions from the data which form the basis for decision making.*** The interpretation of data is a difficult task and necessitates a high degree of skill and experience. If data that have been analyzed are not properly interpreted, the whole purpose of the investigation may be defected and fallacious conclusion be drawn. So that great care is needed when making interpretation.

1.8 Basic Statistical Terms

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

Population: A population is a totality of things, objects, peoples, etc about which information is being collected. It is the totality of observations with which the researcher is concerned.

Parameter: - A parameter is a characteristic or measure obtained by using all the data values from a specific population. Or it is the population measurement used to describe the population.

Sampling: - The process of selecting a sample from the population is called sampling.

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

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

Example: population mean and population standard deviation

Statistic: - It's a measure used to describe the sample. Or it is a value computed from the sample.

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

Data:-Data as a collection of related facts and figures from which conclusions may be drawn.

Variable: A certain characteristic which changes from object to object and time to time. Or it is an item of interest that can take on many different numerical values.

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

Censes survey (studying the whole population without considering samples) requires a great deal of time, money and energy. Trying to study the entire population is in most cases technically and economically not feasible.

To solve this problem, we take a representative sample out of the population on the basis of which we draw conclusions about the entire population.

Therefore, **sampling survey:**

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

1.9 Types of Variables and Scales of Measurement

❖ Types of Variables

– Variables can be classified as qualitative or quantitative.

1. **Qualitative Variables** are non-numeric variables and can't be measured. They are variables that have distinct categories according to some characteristic or attribute. Examples: gender, religious affiliation, and state of birth.
2. **Quantitative Variables** are numerical variables that can be counted and measured. Examples include balance in checking account, number of children in family, the distance between two points and alike. A quantitative variable itself can be classified as ***continuous*** and ***discrete*** variables.

A) **Quantitative Discrete Variables:** - are obtained by counting. A discrete variable takes always whole number values that are counted.

Example: Variables such as number of students, number of errors per page, number of accidents on traffic line, number of defective or non-defective items produced in production line.

B) Quantitative Continuous Variables: - are usually obtained by measurement not by counting. These are variables which assume or take any decimal values or fractions when collected. The variables like age, time, height, income, price, temperature, and etc. are all continuous since the data collected from such variables can take decimal values.

❖ **Scales of Measurement**

What is Measurement?

Normally, when one hears the term *measurement*, they may think in terms of measuring the length of something (i.e. the length of a piece of wood) or measuring a quantity of something (i.e. a cup of flour). This represents a limited use of the term measurement. In statistics, the term measurement is used more broadly and is more appropriately termed *scales of measurement*. Each scale of measurement has certain properties which in turn determine the appropriateness for use of certain statistical analyses.

Measurements scales can have **four properties** and the combination of these properties determine what is measured. The properties are

i. differences (e.g. cold-warm, male-female)

ii. magnitude (one attribute is greater than, less than or equal to another instance)

iii. equal intervals between magnitude

iv. a true zero on the scale

Magnitude

Is the quantum or quantity in which the attribute exists in various instances of the phenomena. It allows us to tell whether one instance of the attribute is greater than, less than or equal to another instance of the attribute.

Example: If X gets a score of 20 on an aggressiveness scale and Y a score of 25, we can say that Y is more aggressive than X.

Equal intervals

It denotes that the magnitude of the attribute represented by a unit of measurement on the scale is equal regardless of where on the scale the unit falls.

Example: A difference in heights between 60 inches and 65 inches is equal to the difference in height, between 67 inches and 72 inches.

Absolute zero point

Is a value that indicates zero exists at that point or nothing at the entire attribute being measured exists?

For example, a zero weight indicates "no weight" at all. Keeping in mind these three characteristics of measurements, scales of measurement can be divided in to four different types.

In addition to being classified as qualitative or quantitative, variables can be classified by how they are categorized, counted, or measured. For example, can the data be organized into specific categories, such as area of residence (rural, suburban, or urban)? Can the data values be ranked, such as first place, second place, etc.? Or are the values obtained from measurement, such as heights, IQs, or temperature? This type of classification—i.e., how variables are categorized, counted, or measured—uses **measurement scales**. The four scales of measurement are ***nominal, ordinal, interval,*** **and** ***ratio*****.**

A. ***Nominal Level of Measurement:*** **-** The nominal level of measurement classifies data in to mutually exclusive (non-overlapping) categories in which no order or ranking can be imposed on the data. Sometimes the variable under study is classified by some quality it possesses rather than by an amount or quantity. In such cases, the variable is termed as **attribute.**

Nominal Scale, also called the categorical variable scale, is defined as a scale used for labelling variables into distinct classifications and doesn't involve a quantitative value or order.

In the nominal scale of measurement, numbers are used simply as labels for groups or classes. Calculations done on these variables will be futile as there is no numerical value of the options.

Examples of variables which belongs to this scale:

- Religion affiliation: Christianity, Islam, Hinduism, etc.
- Gender: Male, Female
- Eye color: like- brown, black, etc.

- Blood type: A, B, AB and O.
- ZIP code
- Major program
- Marital status: Single, Divorced, Widowed and Married.
- Political affiliation: Republican, Democrat, or Other.
- Animal or non-animal.
- Country of origin. etc.

B. Ordinal Level of Measurement: -Whenever observations are not only different from category to category, but can be ranked according to some criterion or characteristics. The variables deal with their relative difference rather than with quantitative differences. 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. Ordinal data are data which can have meaningful inequalities. The inequality signs < or > may assume any meaning like 'stronger, softer, weaker, better than etc.

Here in this category, the order of the value is what is important and significant, but the differences between each one is not really known. Ordinal scale reflects only magnitude and does not possess the attribute of equal intervals and an absolute zero point.

Examples of variables which belongs to this scale:

- Patients may be characterized as unimproved, improved & much improved.
- Letter grades (A, B, C, D, F).
- Rating (Lickert*) scales: non-numeric concepts like- satisfaction, happiness etc.

  For example,

  – **How do you feel today?**

  1 - Very unhappy
  2 - Unhappy
  3 - Ok
  4 – Happy
  5- Very happy

  – **How satisfied are you with our service?**

  1 - Very unsatisfied
  2 - Somewhat unsatisfied
  3 - Neutral
  4 - Somewhat satisfied
  5 - Very satisfied

– In each case, we know that a #4 is better than a #3 or #2, but we don't know- and can't quantify how much better it is.

- Military status.

- Order in race
- Individuals may be classified according to socio-economic as low, medium & high. Etc.

C. ***Interval Level Scale:*** The interval scale is a quantitative measurement scale where there is order, the difference between the two variables is meaningful and equal, The interval level of measurement ranks data, however, there is no meaningful zero. Interval data are the types of information in which an increase from one level to the next always reflects the same increase. Possible to add or subtract interval data but they may not be multiplied or divided.

Examples:

Ex1. Temperature

Temperature of zero degrees does not indicate lack of heat. The two common temperature scales; Celsius (C) and Fahrenheit (F). We can see that the same difference exists between 10^{o}C and 20^{o}C as between 25^{o}c and 35^{o}c i.e., the measurement scale is composed of equal-sized interval. But we cannot say that a temperature of 20^{o}c is twice as hot as a temperature of 10^{o}c. because the zero point is arbitrary.

Ex2. IQ
Ex3. SAT scores

Here's the problem with interval scales: they don't have a "true zero." For example, there is no such thing as "no temperature," at least not with Celsius. In the case of interval scales, zero doesn't mean the absence of value, but is actually another number used on the scale, like 0 degrees Celsius. Negative numbers also have meaning. Without a true zero, it is impossible to compute ratios.

Interval scale contains all the properties of the ordinal scale, in addition to which, it offers a calculation of the difference between variables. The main characteristic of this scale is the equidistant difference between objects.

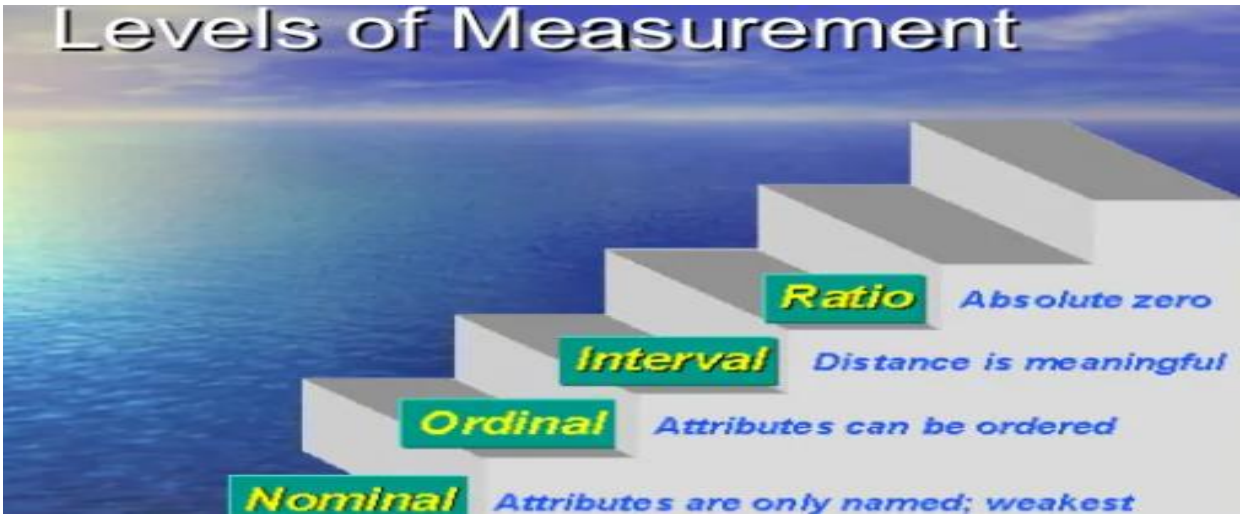
D. ***Ratio Level of Measurement:*** **-** The ratio level of measurement possesses all the characteristics of interval measurement, and there exists a true zero. In addition, true ratios exist when the same variable is measured on two different members of the population or sample. Typical examples of ratio scales are measures of time or space. For example, if one

person can lift 200 pounds and another can lift 100 pounds, and then the ratio between them is 2 to 1. Put another way, the first person can lift twice as much as the second person. Interval scales do not have the ratio property. Most statistical data analysis procedures do not distinguish between the interval and ratio properties of the measurement scales.

Ratio Scale: is defined as a variable measurement scale that not only produces the order of variables but also makes the difference between variables known along with information on the value of true zero. It is calculated by assuming that the variables have an option for zero, the difference between the two variables is the same and there is a specific order between the options.

Examples of variables which belongs to this scale:

- Age
- Height
- Length
- Number of students
- Response speed
- Rate
- Percentage
- Time
- Amount of rainfall, etc.



❖ **Summary table for measurement scale**

Provides:	**Nominal**	**Ordinal**	**Interval**	**Ratio**
- *"Counts," aka "frequency of distribution"*	√	√	√	√
- The sequence of variables is established		√	√	√
- *Mode,*	√	√	√	√

- *Median*	√	√	√
- *Mean*		√	√
- *The "order" of values is known*	√	√	√
- Can quantify the difference between each value		√	√
- Can add or subtract values		√	√
- Can multiple and divide values			√
- Has "true zero"			√

- The table below will help to clarify the fundamental differences between the four scales of measurement

Scale	Indications Difference	Indicates Direction of Difference (Ranking, order or Scaling)	Indicates Amount of Difference (Equal Interval)	Absolute Zero
Nominal	√			
Ordinal	√	√		
Interval	√	√	√	
Ratio	√	√	√	√

CHAPTER TWO

2. Methods of Data Collection and Presentation

2.1 Method of Data Collection

2.1.1 Source of Data

We have already explained what it means by statistical data. The statistical data may be already available or may have to be collected by an investigator or an agency.

- The data termed as **primary** (**first hand**) data when the reference is to data collected for the first time by the investigator and,
- The data termed as **secondary** (**second hand**) data when the data are taken from records or data already available for the same as well as other purposes.

Characteristics of Primary Data

- Measurements observed and recorded as part of an original study.
- The data required for a particular study can be found neither in the internal records of the enterprise, nor in published sources,
- It may become necessary to collect original data to conduct first hand investigation.
- The work of collecting original data is usually limited by
    - Time,
    - Money and
    - Manpower available for the study.

Characteristics of Secondary Data

- The investigator need not begin from the very beginning,
- It is a data which has already been collected by others.
- It may be a must to take into account what has already been discovered by others.
- Secondary data can be obtained from
    - Journals,
    - Different Public and Private Organizations,
    - Reports,
    - Government publications,

- Publications of research organizations, or sometime unpublished resources etc.

Secondary data must be used with utmost care, because:

- Such data may be full of errors because of bias,
- Inadequate size of the sample,
- Arithmetical errors, etc.
- Even if there is no error, secondary data may not be suitable and adequate for the purpose of inquiry.

2.1.2 Method of Primary Data Collection

In primary data collection, you collect the data by yourself using methods such as interviews, observations, laboratory experiments, diary and questionnaires. The key point here is that the data you collect is unique to you and your research and, until you publish, no one else has access to it. There are many methods of collecting primary data and the main methods include:

a) Self-administered Questionnaire: It is a set of questions relating to the enquiry, but is difficult to design and often require many rewrites before an acceptable questionnaire is produced.

Advantages:

- Can be posted, e-mailed or faxed.
- Can cover a large number of people or organizations.
- Wide geographic coverage.
- Relatively cheap.
- No prior arrangements are needed.
- Avoids embarrassment on the part of the respondent.
- No interviewer bias.

Disadvantages:

- Historically low response rate (although inducements may help).
- Time delay whilst waiting for responses to be returned.
- Assumes no literacy problems.
- No control over who completes it.
- Respondent can read all questions beforehand and then decide whether to complete or not. For example, perhaps because it is too long, too complex, uninteresting, or too personal.

b) ***Personal Interviewing*** is a technique that is primarily used to gain an understanding of the underlying reasons and motivations for people's attitudes, preferences or behavior. Interviews can be undertaken on a *personal one-to-one* basis or *in a group*. They can be conducted at work, at home, in the street or in a shopping center, or some other agreed location. In this method, It is essential that the interviewer should be polite, tactful and has a sense of a response.

Advantages:

- Serious approach by respondent resulting in accurate information.
- Good response rate.
- Completed and immediate.
- Possible in-depth questions.
- Interviewer in control and can give help if there is a problem.
- Can use recording equipment.
- It is most useful when the area of investigation is very small.
- Characteristics of respondent assessed – tone of voice, facial expression, hesitation, etc.

Disadvantages:

- Need to set up interviews.
- Time consuming.
- Geographic limitations.
- Can be expensive.

- Embarrassment possible if personal questions.
- Transcription and analysis can present problems– subjectivity.
- The chances of personal bias are greater.
- If many interviewers, training required.

c) Observation: It involves recording the behavioral patterns of people, objects and events in a systematic manner.

d) Laboratory experiment: Conducting laboratory experiments on fields of chemical, biological, engineering, agricultural sciences and so on.

2.1.3 Secondary Data Collection

Secondary data analysis can be literally defined as ***second-hand*** analysis and is the analysis of data or information that was either gathered by someone else (e.g., researchers, institutions, other NGOs, etc.) or for some other purpose than the one currently being considered, or often a combination of the two.

The investigator need not begin from the very beginning; It may be a must to take into account what has already been discovered by others.

Before using secondary data, we have to examine the following aspects:

- Whether the data are **suitable** for the purpose of investigation.
- Whether the data are **adequate** for the purpose of the investigation. (For example, if our object is to study the wage rates of the workers in the cotton industry in Ethiopia and if the available data covers only single industry, it would not solve the purpose).
- Whether the data are **reliable:** (to determine the reliability of secondary data is perhaps the most important and at the same time most difficult job).

Some of the sources of secondary data are ***government document***, ***official statistics***, ***technical report***, ***scholarly journals***, ***trade journals***, ***review articles***, ***reference books***, ***research institutes***, ***universities***, **hospitals**, ***libraries***, ***library search engines***, ***computerized data base*** **and** ***world wide web*** **(*WWW*).**

Advantage of secondary data

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

Disadvantage of secondary data

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

2.2 Methods of Data Presentation

So far you know how to collect data. So, what do we do with the collected data next? Now you have to present the data you have collected so that they can be of use. Thus, the collected data also known as **raw data** are always in an unorganized form and need to be organized and presented in a meaningful and readily comprehensible form in order to facilitate further statistical analysis.

This chapter introduces ***tabular, graphical and diagrammatic methods*** commonly used to summarize both qualitative and quantitative data. Tabular and graphical summaries of data can be obtained in annual reports, newspaper articles and research studies. Everyone is exposed to these types of presentations, so it is important to understand ***how they are prepared and how they will be interpreted***.

Modern statistical software packages provide extensive capabilities for summarizing data and preparing graphical presentations. **MINITAB**, **SPSS** and **STATA** are three packages that are widely available.

2.2.1 Frequency Distribution

A frequency distribution is the organization of raw data in table form, using classes, frequencies, percentage and cumulative frequencies.

The reasons for constructing a frequency distribution are as follows;

- To organize the data in a meaningful, intelligible way.
- To enable the reader to determine the nature or shape of the distribution
- To facilitate computational procedures for measures of average and spread
- To enable the researcher to draw the charts and graphs for the presentation of data
- To enable the reader to make comparisons between different data set

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

A. Categorical Frequency Distribution

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

The major components of categorical frequency distribution are ***class, tally*** and ***frequency***. Moreover, even if *percentage* is not normally a part of a frequency distribution, it will be added since it is used in certain types of graphical presentations, such as pie chart.

Steps of constructing categorical frequency distribution

1. You have to identify that the data is in nominal or ordinal scale of measurement
2. Make a table as show below

A	B	C	D
Class	Tally	Frequency	Percent

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

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

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

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

Construct a frequency distribution for the above data

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

B. Discrete/Ungrouped Frequency Distribution

When the data are *numerical* instead of categorical, *the range of data is small* and each class is only *one uni*t, this distribution is called an ***ungrouped frequency distribution.***

The major components of this type of frequency distributions are class, tally, frequency, relative frequency and cumulative frequency. The steps are almost similar with that of categorical frequency distribution.

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

Example 2.2: The following data represent the number of days of sick leave taken by each of 50 workers of a company over the last 6 weeks.

2	0	0	5	8	3	4	1	0	0	7 1	7
1	5	4	0	4	0	1	8	9	7	0	1
7	2	5	5	4	3	3	0	0	2	5	1
3	0	2	4	5	0	5	7	5	1	1	0
2											

i. Construct ungrouped frequency distribution

ii. How many workers had at least 1 day of sick leave?

iii. How many workers had between 3 and 5 inclusive days of sick leave?

Soln.

i. Since this data set contains only a relatively small number (9) of distinct or different values, it is convenient to represent it in a frequency table which presents each distinct value along with its frequency of occurrence.

Class	Frequency	Cumulative frequency
0	12	12
1	8	20
2	5	25
3	4	29
4	5	34
5	8	42
7	5	47
8	2	49
9	1	50

ii. Since 12 of the 50workers had no days of sick leave, the answer is 50-12=38

iii. The answer is the sum of the frequencies for values 3, 4 and 5 that is 4+5+8=17

Note that: If the number of possible values of a discrete variable is very large the discrete frequency distribution will not more be condensed presentation, then the data handled as continuous variable and distributed in to classes.

C. **Continuous/Grouped Frequency Distribution**

When the range of the data is large, the data must be grouped in which each class has more than one unit in width. Some of basic terms that are most frequently used while we deal with grouped frequency distribution are the following:

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

While we construct this frequency distribution, we have to follow the following steps.

1. Find the highest and the lowest values

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

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

I. Use Struge's rule. That is, $K = 1 + 3.32 \log n$ where K is the number of class and n is the number of observations. OR

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

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

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

4. Find the class width by dividing the range by the number of classes which means:

$$\mathrm{W} = \frac{\mathrm{R}}{\mathrm{K}} \quad \text{or} \qquad \text{Width} = \frac{\text{Range}}{\text{NumberofClasses}}$$

Note that: Round the answer up to the nearest whole number if there is a reminder. For instance, $\mathbf{4.7 \approx 5}$ ***and*** $\mathbf{4.12 \approx 5}$.

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

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

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

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

7. Find the class boundaries. $Lower\,Class\,Boundary = Lower\,Class\,Limit - \frac{U}{2}$ and $Upper\ Class\,Boundary = Upper\,Class\,Limit - \frac{U}{2}$. In short, $LCB = LCL - \frac{U}{2}$ and $UCB = UCL + \frac{U}{2}$.

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

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

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

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

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

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

***Example 2.4**:* Consider the following set of data and construct the frequency distribution.

11 29 6 33 14 21 18 17 22 38

31 22 27 19 22 23 26 39 34 27

Steps

1. *Highest value=39, Lowest value=6*
2. $R = 39 - 6 = 33$
3. $K = 1 + 3.32\log 20 = 5.32 \approx 6$
4. $W = \frac{R}{K} = \frac{33}{6} = 5.5 \approx 6$
5. Select starting point. Take the minimum which is 6 then add width 6 on it to get the next class LCL.

6	12	18	24	30	36

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

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

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

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

8. 9 and 10

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

Example 2.5 The following data are on age of 20 women who attended health education in a certain hospital. Construct frequency distribution by using sturge's rule.

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

Solution (Exercise)!!!

D. Relative Frequency Distribution

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

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

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

Exercises

1. **FM Radio Stations** A random sample of 30 states shows the number of low-power FM radio stations for each state. Find the variance and standard deviation for the data.

Class limits	Class Boundary	Frequency (f)	Class Mid-point (m_i)	Cumulative frequency		Relative frequency
				Less than	More than	
1–9		5				
10–18		7				
19–27		10				
28–36		3				
37–45		3				
46–54		2				

- Complete the above continuous frequency distribution

2. A random sample was taken of the thickness of insulation in transformer windings, and the following thicknesses (in millimeters) were recorded:

18	21	22	29	25	31	37	38	41	39
44	48	54	56	57	47	38	35	56	36
29	37	32	42	43	40	48	36	37	37
36	38	40	41	44	39	38	34	24	32
39	44	42	30	37	30	42	37	34	37
32	24	42	36	49	39	23	34	36	40

- Prepare an appropriate frequency distribution for the above information.

2.2.2 Diagrammatic Presentation of data

We have discussed the techniques of classification and tabulation that help us in organizing the collected data in a meaningful fashion. However, this way of presentation of statistical data does not always prove to be interesting to a layman. Too many figures are often confusing and fail to convey the massage effectively.

One of the most effective and interesting alternative way in which a statistical data may be presented is through diagrams and graphs. There are several ways in which statistical data may be displayed pictorially such as different types of graphs and diagrams.

General steps in constructing graphs

1. Draw and label the x and y axes

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

3. Represent the class boundaries for the histogram or Ogive or the mid-point for the frequency polygon on the x axis.

4. Plot the points

5. Draw the bars or lines

1. Pie Chart

Pie chart can used to compare the relation between the whole and its components. Pie chart is a circular diagram and the area of the sector of a circle is used in pie chart. Circles are drawn with radii proportional to the square root of the quantities because the area of a circle is πr^2.

To construct a pie chart (sector diagram), we draw a circle with radius (square root of the total). The total angle of the circle is 360^0 . The angles of each component are calculated by the formula

$$Angle\,of\ Sector = \frac{Component\,Part}{Total} x360^0$$

These angles are made in the circle by mean of a protractor to show different components. The arrangement of the sectors is usually anti-clock wise.

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

Item of Expenditure	Family Budget
Good	$ 600
Closing	$ 100
House Rent	$ 400
Fuel and lighting	$ 100
Miscellaneous	$ 300
Total	$ 1500

Solution: The necessary computations are given below:

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

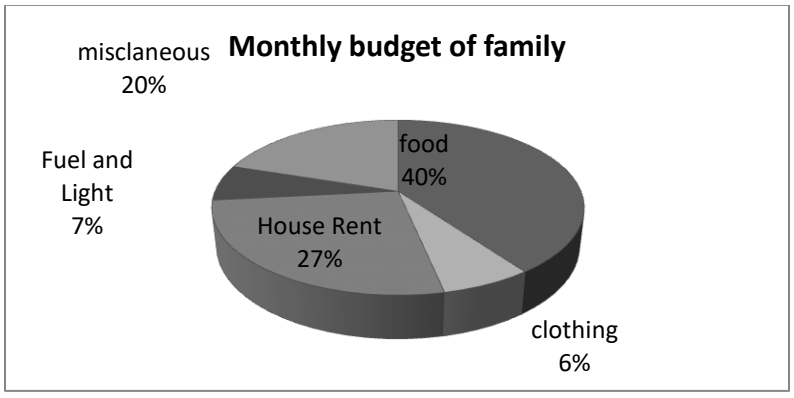


Figure 2.1 pie chart

2. Pictogram

Data can be presented by means of some picture symbols. Once we decide about a suitable picture or symbols to represent, the number of units in each category represent the number of things in each category.

The symbols used are thematically linked to the nature of the data.

Example 2.7

The table below shows the number of medals in different sporting categories tha competitors from a particular country won in an international games event.

Number of medals won by type of sport

Type of sport	*Medals won*
Track and field	6
Swimming	3
Team games	2

Show this set of data in the form of a pictograph.

Type of sport	Medals won
Track and field	🏅 🏅 🏅 🏅 🏅 🏅
Swimming	🏅 🏅 🏅
Team games	🏅 🏅

Each symbol represents one medal.

3. Bar Charts

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

- Make the bars the same width.
- Make the units on the axis that are used for the frequency equal in size.
- the gap between successive bars should be remains the same.

A. Simple bar chart

It's used to represents data involving only one variable classified on spatial, quantitative or temporal basis. In simple bar chart, we make bars of equal width but variable length,

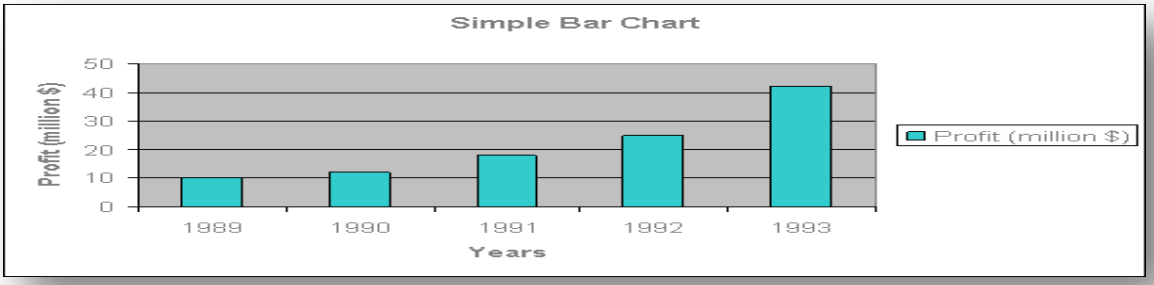
i.e. the magnitude of a quantity is represented by the height or length of the bars.

Following steps are undertaken in drawing a simple bar diagram:

- Draw two perpendicular lines one horizontally and the other vertically at an appropriate place of the paper.
- Take the basis of classification along horizontal line (X-axis) and the observed variable along vertical line (Y-axis) or vice versa.
- Marks signs of equal breath for each class and leave equal or not less than half breath in between two classes.
- Finally, marks the values of the given variable to prepare required bars.

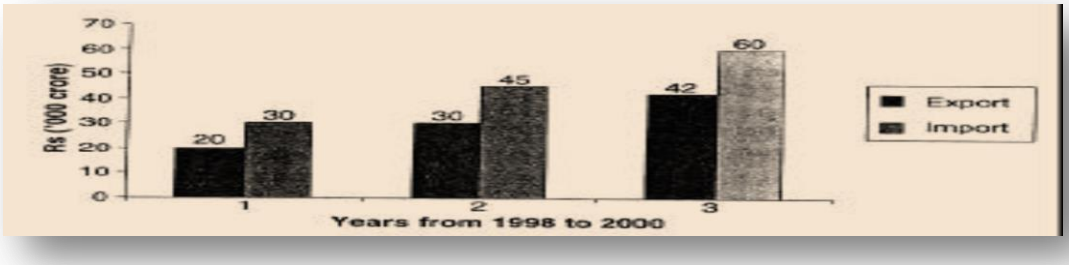
Example 2.7*:* Draw simple bar diagram to represent the profits of a bank for 5 years.

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



B. Multiple Bars

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



Example2.8: Draw simple bar diagram to represent the number of students in a given school for three consecutive years. The data is given bellow;

Students	**Year**			**Total**
	2001	**2002**	**2003**	
Female	95	145	210	450
Male	207	350	570	1127
Total	302	495	780	1577

solution (exercise)!!!

C. *Stratified (Stacked) Bar Chart*

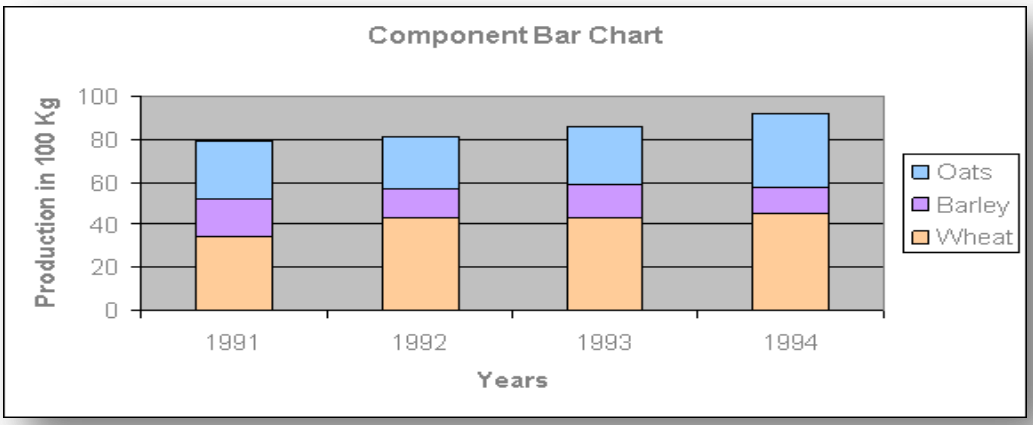
It should be noted that multiple bar diagrams are particularly suitable where some comparison is involved.

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

Example 2.9: The table below shows the quantity in hundred kgs of Wheat, Barley and Oats produced on a certain form during the years 1991 to 1994. Draw stratified bar chart.

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

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



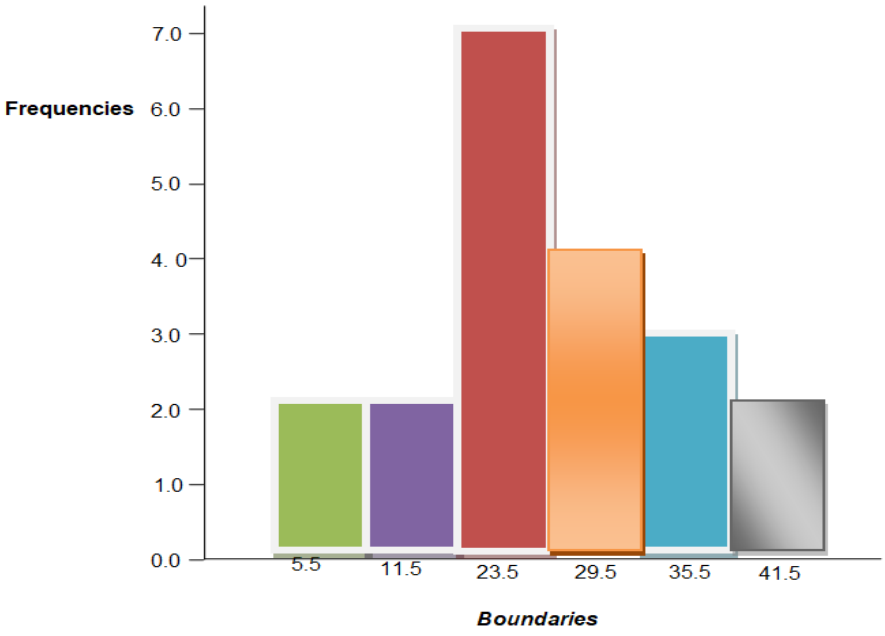
2.2.3 Graphical Presentation of Data

A. Histogram

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

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

Example2.10: Take the data in *example 2.4*, the histogram for the data can be of the form:

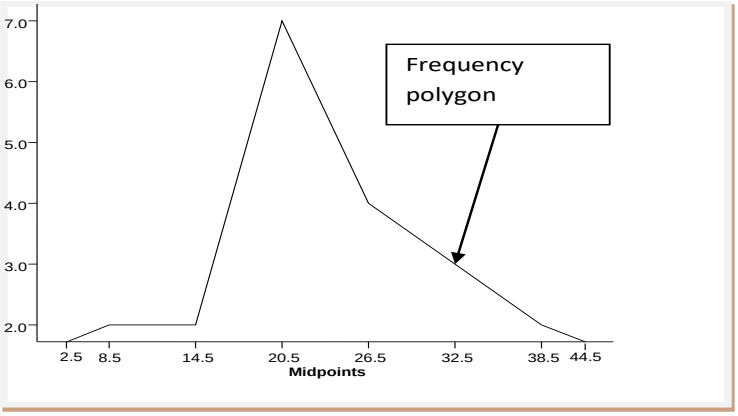


Relative frequency histogram has the same shape and horizontal ($y-$axis) scale as a histogram, but the vertical ($x-$axis) scale is marked with *relative frequencies* instead of actual frequencies.

B. Frequency Polygon

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

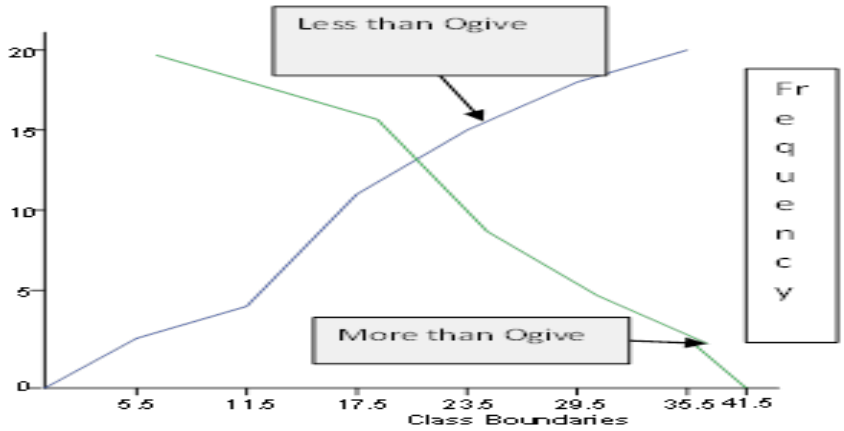
Example 2.11: Take the data in *example 2.4*, the frequency polygon for the data can be of the form:



C. Ogive Graph

An *Ogive* (pronounced as "oh-jive") is a line that depicts *cumulative* frequencies, just as the cumulative frequency distribution lists cumulative frequencies. Note that the Ogive uses class boundaries along the 0.horizontal scale, and graph begins with the lower boundary of the first class and ends with the upper boundary of the last class. Ogive is useful for determining the number of values below some particular value. There are two type of Ogive namely *less than Ogive* and *more than Ogive*. The difference is that less than Ogive uses less than cumulative frequency and more than Ogive uses more than cumulative frequency on y axis.

Example 2.12: Take the data in *example 2.4* and draw less than and more than Ogive curve.


Less than Ogive
More than Ogive
Frequency
20
15
10
5
0
5.5
11.5
17.5
23.5
29.5
35.5
41.5
Class Boundaries