UNIT 1: INTRODUCTION TO STATISTICS

1.1 Definition of statistics

Although the term Statistics is defined in a number of ways, all the definitions converge to two basic aspects. That is, Statistics may be defined as Statistical data (plural sense) or it can also be defined as a method (singular sense). Each one of these definitions is treated separately as follows.

Statistics defined as data (*Plural sense****)***

Statistics in the sense statistical data refers to a quantitative aspect of things and is a numerical description of a group of individuals.

Statistic (plural sense): A group of numbers that presents facts or describe situation.

Example*: 1. Ages of the students in your class.*

2. The monthly income of employees in a company.

According to this notion, Prof. Horace Secrist gives the following definition:

"Statistics refer to the aggregates of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a pre-determined purpose and placed in relation to each other."

This definition makes it clear that Statistics (as numeric data) should possess the following characteristics.

i. Statistics should be aggregates of facts

Single and isolated figures are not Statistics for the simple reason that such figures are unrelated and can't be compared. According to this aspect, to be Statistics, data must be in aggregate (mass) and also the individual elements within the aggregate should relate to a common phenomenon so that they can be compared to one another.

ii. Statistics should be affected to a marked extent by multiplicity of causes

Since Statistics are most commonly used in social sciences it is natural that they are affected by a large variety of factors at the same time. Putting differently, Statistics are not as such caused by a single factor (force), rather they are outcomes of a number of (multiple) factors (forces) operating together.

iii. They should be numerically expressed

All Statistics are expressed in numbers. Nevertheless, the converse of this statement is not in general true. That is, statements expressed in terms of numbers may not necessarily be Statistics as there is a possibility for them not to meet the other requirements of the definition given above.

iv. They should be enumerated or estimated according to reasonable standards of accuracy

Numerical statements can either be enumerated, in which case they are supposed to be accurate and precise or else they can be estimated by some expert observers, in which case 100% accuracy is unlikely to be attained. In the process of estimation, reasonable standards of accuracy must, however, be attained.

v. They should be collected in a systematic manner

If data are collected in haphazard manner, then results to be obtained are likely to lead to fallacious conclusions. Therefore, it is essential that Statistics must be collected in a systematic manner so that they may confirm to reasonable standards of accuracy.

vi. They should be collected for a predetermined purpose.

Statistics collected without any predetermined purpose do not serve any useful purpose. Therefore, the purpose of collecting Statistics should be defined clearly before they are collected. Meaning, figures (Statistics) should be collected in view of some goal or target. Moreover, the data should be collected in such a manner that it meets the predetermined needs.

vii. They should be placed in relation to each other

For numerical facts to be called Statistics they should be comparable either period wise or region wise, or in reference to some other means of comparison.

As an example, suppose that the marketing head of a given supermarket in Addis Ababa wants to know the average expenditure of households in the city, among other things, so as to revise his marketing strategy. To achieve his objective, the head collects data on expenditure from a sample of 1000 households' selected using stratification (to be discussed in the next block). Moreover, the head used the interview approach to gather the required information.

Thus, the data collected by the marketing head are Statistics as they fulfill all the requirements of the definition.

First, the data is collected from 1000 households. Hence, it forms an aggregate. Second, data on expenditure are affected by a number of factors like income, taste, season, culture, etc. Third, expenditure of households, as they are expressed in terms of currency, is clearly numeric values. Fourth, as far as expenditures are expressed to the nearest, say birr, the degree of accuracy is more than satisfactory. Fifth, since the investigator uses stratified sampling technique and also the interview data collection method it is reasonable to say that the data is collected in a systematic or scientific manner. The record lack requirement if fulfilled as the investigator has a pre-determined end or goal, which is revising his marketing strategy, based on the average expenditure level and other qualitative and quantitative variables. Finally, it is obvious that expenditure of households can be analyzed (compared) be it region wise or period wise.

Statistics defined as a method (Singular sense)

The second definition of Statistics refers to the science or the methods of Statistics. It is also in the sense of its second definition that we consider Statistics as a subject. With this regard, Statistics may be defined as:

"Statistics is the science which deals with the methods of collecting, classifying, presenting, comparing (analyzing) and interpreting numerical data collected to throw some light on any sphere of enquiry."

"Statistics is the method of judging collective, natural or social phenomei *Seligman*
results obtained from the analysis or enumeration or collection of estimates."

King

"Statistics is the application of the scientific method in the analysis of numerical data for the purpose of making rational decisions."

Berenson and Levin

"Statistics is the collection, presentation, analysis and interpretation of numerical data."

Coroxton and Cowden

Summing up all the above definitions, one can define Statistics preferably as:

Statistics is the study of the principles and methods used in the collection, presentation, analysis and interpretation of numerical data in any sphere of enquiry.

1.2. Basic terminologies in statistics

As a subject (science), Statistics has its own terms and terminologies. Knowing these terms and terminologies is fundamental in understanding the Statistical methods and concepts. To that end, this section presents the meaning of the most commonly used Statistical terms falling under the scope of this textbook.

Variable: A variable is a factor or characteristic that can take on different possible values or outcomes. A variable differs from a constant is that the latter term implies that the values or outcomes are always the same. Income, height, weight, sex, age, etc are examples of variables. In an investigation, data are collected about one or more variables of interest. A variable can be qualitative or quantitative (numeric).

Population: In Statistics the term population is used to mean the totality of causes (items) under consideration in a given investigation or research. In other words, the largest collection of observations on a variable constitutes the population. Population can be finite (limited in its size) or infinite (unrestricted). In finite population, observations are countable- at least in theory. In contrast, infinite population is indefinitely large. The observations cannot be even in theory.

Sample : Any non-empty subset of a population is called a sample. There are different possible samples that can be selected from a single population. Nevertheless, the one that best reflects or represents the behavior of the population is considered to be the

most appropriate one. The critical question is "How to identify and get that best representative sample?" In fact, the whole aim of the theory of sampling is to answer this question.

Parameter : It is a measurable characteristic of the population or it is a numerical result obtained as measuring the population.

Statistic : It is a measurable characteristic of the sample. In short it is a sample result.

Survey: Survey or experiment is a device of obtaining the desired data.

Statistical Design: Statistical design is a process that involves a decision problem and choosing an approach to solving the problem. It is a guide that indicates how an investigation is going to channeled.

1.3. Functions, importance and limitation of statistics

Functions

The main function of Statistics is to collect and present numerical data in a systematic manner so that it may be analyzed in a scientific way. Statistics basically concentrates on the analysis of a phenomenon in a scientific manner, without proving it. The analysis of data, which is the core objective of Statistics, is important because it helps to avoid or replace arbitrary decisions, dogmatism, rule of thumb, tradition, and it tries to increase the custom of making decision based on analyzed quantitative facts.

The following are the major functions of Statistics:

i. It simplifies mass of data (condensation)

It is common that the human mind is not capable of assimilating huge (mass of) facts and figures, as they are complex to understand. Statistical methods simplify this complexity by making the huge data easily intelligible and readily understandable. Meaning, Statistical methods provide the necessary means to condense mass of data and present them with the help of simple figures such as averages, ratios, variations, measures of skew ness, coefficients, etc. More attractive and understandable presentations of data are also made through the help of diagrams and graphs.

ii. It presents facts in a definite form (Definiteness)

Statistics presents facts in a precise and definite (numeric) form and thus helps proper comprehension of what is to be stated. Facts (data) stated in quantitative terms are more precise and simple to understand, analyze and interpret. To give an example, suppose that you read the statement:

"The average rainfall in 1995 is expected to be lower than that of 1994."

This statement does not tell precise and complete information because it doesn't present the quantified rainfall in the year 1995. On the other hand, the statement may be presented as: "**The average rainfall in 1995 is expected to decrease by 5% from that of 5ml recorded in 1994."** Clearly, since the rate of decrement in rainfall is quantitatively expressed (5%) it is possible to have the precise expected rainfall in 1995 in quantitative form.

iii. It facilitates Comparison

The very reason for saying numerical data are more precise is that they are amendable for (lend themselves to) comparison. By furnishing different suitable devices or tools for comparison, like averages and measures of dispersion, Statistics enables better understanding and appreciation of the significance of a series of figures. Moreover, in most cases conclusions or decisions are reached mainly based on the results obtained from different comparisons. For example, the aggregate performance of students in two different sections (classes) can be judged by comparing the average marks for the two classes.

iv. It helps in Predictions

One of the major reasons making Statistical methods so critical in Business is their prediction function.

Prediction is the process of making a scientific guess about the future value of a variable. Statistical methods made it possible to predict the likely future value of a variable based on its past trend. Time series and regression analysis are the most commonly used methods towards prediction.

v. It helps in formulation of suitable policies

If carefully handled, Statistics plays invaluable role in the process of policy formulation. Statistical data and Statistical methods help the government in formulating suitable methods help the government in formulating suitable policies with respect to taxation, import-export, budgeting and other socio-economic welfare programs.

vi. It helps in Formulating and Testing hypothesis

This function of Statistics refers to an aspect of inferential Statistics. In inferential Statistics, hypothesis are formulated and tested to make conclusions and in some cases to develop new theories. A Statistical hypothesis is usually a mathematical statement that states about the nature (significance) of variables. In addition, Statistical methods help in testing the correctness of the laws of the different branches of knowledge. Hypothesis testing is commonly used in the fields of medicine, agriculture, economics and business.

Importance of statistics

The fact that the modern world uses Statistical methods is universally what the science of Statistics is. The increasing global economy and the high degree of flexibility provided by Statistical methods has rendered them especially useful and indispensable.

Some of the diverse fields in which Statistical methodology has had extensive applications are:

Business: Estimating the volume of retail sales, designing optimum inventory control system, producing auditing and accounting procedures, improving working conditions in industrial plants, assessing the market for new products.

Economists: Measuring indicators such as volume of trade, size of labor force, and standard of lining, analyzing consumer behavior, computation of national income accounts, formulation of economic laws, etc. Particularly the theory of regression analysis extensively used in the field of economics.

Quality Control: Determining techniques for evaluation of quality through adequate sampling, in process control, consumer survey and experimental design in product development etc. Realizing its importance, large organizations are maintaining their own Statistical quality control department.

Health and Medicine: Developing and testing new drugs, delivering improved medical care, preventing diagnosing, and treating disease, etc. Specifically, inferential Statistics has a tremendous application in the fields of health and medicine.

Biology: Exploring the interactions of species with their environment, creating theoretical models of the nervous system, studying genetical evolution, etc.

Auctorial Sciences: Determining premium rates for different insurance risks, designing pension plans for private and public groups, measuring effectiveness of loss prevention and loss control programs, etc.

Engineering: Working out safer systems of flight control for airports, improving product design and testing product performance, determining reliability and maintainability, etc.

Psychology: Measuring learning ability, intelligence, and personality characteristics, creating psychological scales and abnormal behavior, etc.

Sociology: *Testing theories about social systems, designing and conducting sample surveys to study social attitudes, exploring cross-cultural differences, studying the growth of human population, etc.*

Nevertheless, it has to be emphasized that no volume of Statistics can replace the knowledge and expenditure of the specialist in one or more of the mentioned fields. Rather, in all of the above areas and many others, statisticians work closely with other scientists and researchers to develop new Statistical techniques, adapt existing techniques, design experiments, and direct the analysis of surveys and studies.

Limitations of statistics

The fact that Statistics is applicable in almost all fields of study is not a guarantee for its perfection. Of course, there is no perfect science in the globe. Statistical methods as well have their own limitations. The following are the major limitations:

i. Statistics does not deal with individual items

This is to mean that Statistics deals only with aggregates of facts and no importance is attached to individual items. For instance, age of a single student in a given class in a given year is not a Statistical data. In contrast, the age of all students within a given class in a given year form an aggregate and hence can be considered as data. Alternatively, the semester GPA of a single student for 4 semesters also forms a Statistical data. In short, Statistical methods are suited only to those problems or situations where group characteristics are desired to be studied.

ii. Statistics deals only with quantitatively expressed items

Another limitation of Statistics is that it deals with those subjects of inquiry that are capable of being quantitatively measured and numerically expressed. Accordingly, such qualitative characteristics as health, poverty, honesty and intelligence are not suitable for Statistical analysis however; problems involving such qualitative variables are treated in Statistics indirectly. For example, the variable health may be studied through death rate, which is a quantitative variable. However, these are only indirect methods.

iv. Statistical results are not universally true

As it is often said, Statistical results are true only on the average. Meaning, the results obtained from Statistical data analysis are not true for each member or item within the data for which the analysis is made. Statistical statements or conclusions are not generally true or applicable to individuals, but are applicable to the majority of cases.

v. Statistics is liable to be misused

Misuses of Statistics, unfortunately, are probably as common as valid uses of Statistics. In reality, Statistical methods can be properly used by experienced or trained people, as it

requires skill to draw sensible conclusions from data. It is actually this limitation that hinders the possibility of mass popularity of such a useful and applicable science.

1.4. Types of statistics

Broadly speaking, Statistical methods are classified into two groups or areas based on how data are used. These areas are Descriptive Statistics and Inferential Statistics

Descriptive Statistics: Descriptive Statistics consists of the collection, organization, summarization, and presentation of numerical data. It is concerned with describing certain characteristics of a set of observed data (usually a sample) – that is, what it is shaped like, what number the values tend to cluster (converge) around, how much variation is present in the data, and so forth.

In short, descriptive Statistics describes the nature or characteristics of a data without making conclusion or generalization. The following are some examples of descriptive Statistics.

-The average age of athletes participated in London Marathon was 25 years.

-80% of the instructors in Unity University College are male.

-The marks of 50 students in a mathematics course are found to range from 30 to 85.

Inferential Statistics: Inferential Statistics, also called inductive Statistics, is concerned with the process of drawing conclusions (inferences) about specific characteristics of a population based on information obtained from samples, performing hypothesis testing, determining relationships among variables, and making predictions. The area of inferential Statistics entirely needs the whole aims to give reasonable estimates of unknown population parameters.

The following Statistics are some examples of inferential Statistics:

The result obtained from the analysis of the income of 1000 randomly selected citizens in Ethiopia suggests that the average perception income of a citizen in Ethiopia is 30 Birr.

Note: In the above example we are trying to represent the income of about 65 million population of Ethiopia by a sample of 1000 citizens, hence we are making inference or generalization.

The trend analysis on past data includes that the average exchange rate for a dollar is expected to be 8.30 birr in the coming month.

1.5. Steps in statistical investigation

Recall that according to Coroxton and Cowden, Statistics is defined as the collection, Presentation, analysis and interpretation of numerical data. A bit extension of the above definition leads to the five stages of Statistical investigation. Meaning, in addition to collection, presentation, analysis and interpretation, a Statistical investigation involves one more stage, which is organization of data. These five stages constitute a complete Statistical study or survey. Following are brief explanations about the purpose of each stage.

Stepe 1: ***Data Collection***

Data collection is the process of gathering information or data about the variable of interest. Data are inputs for Statistical investigation. Data may be obtained either from primary source or secondary source.

Stepe 2: ***Organization of Data***

Organization of data includes three major steps. The first step is *editing*. Editing is the process of checking and connecting data for omissions, inconsistencies, irrelevant answers and wrong computations in the collected data. Data collected utmost care simplifies the task of editing and vice versa.

The second step in organizing data is *classification*. Classification is a task of grouping the collected and edited data into different similar categories based on some criterion. Basically, there are four kinds of classifications namely, Geographical, Chronological, Qualitative and Quantitative. Detailed discussions of these terms are made in the third chapter of this text. The third and last step in organization is *tabulation*. The objective of

tabulation is to put the classified data in the form of table. Tabulation of data should be done in such a form that it suits the nature and objectives of the investigation. The importance of proper tabulation is immense because if the tabulation of data is poor, its analysis will not only be difficult but also defective.

Stepe 3: ***Presentation of Data***

The purpose of presentation is to display what is contained in our data in the form of pictures. Diagrams and graphs are the two modes of presentation each containing several different sub-categories (options).

Stepe 4: ***Analysis of Data***

Analysis is the second last stage of Statistical survey where the collected and organized data is manipulated so as to generate or find different quantitative results. It merely involves mathematical operations. Different measures of central tendency (averages), measures of variation regression analysis, etc are only few to be mentioned with in the scope of analysis. In its extreme case, analysis requires the knowledge of advanced mathematics.

Stepe 5: ***Interpretation***

Note that up to the fourth stage what we have is simple calculated figures that need further explanation. As a last stage of Statistical investigation, the task of interpretation associates meaning to the results obtained in the analysis stage.

Without Exaggeration, interpretation is by far the most difficult and skill requiring stage. It is at this stage that Statistics seems to be very much viable to be misused. The good news is that connect interpretation of results will lead to a valid conclusion of the study and hence can aid one in taking suitable decisions. On the other hand, the bad news is that improper (incorrect) interpretation may lead to wrong conclusions and makes the whole object of the study useless.

1.6. Level (scale) of measurement

There are four general levels of measurements. These are: Nominal, ordinal, interval and ratio levels of measurements

1. ***Nominal level:*** The terms nominal level of measurements and nominal scaled are commonly used to refer to data that can only be classified in to categories. In the strict sense of the words, however, there are no measurements and no seals involved. In stead, there are just counts.

Nominal level is considered the most primitive, the lowest or the most limited type of measurement

2. ***Ordinal Level*: *Look at the data below.***

Ratings of the company commander on the nurses as " Superior, Good, Average, Poor, and Inferior " is Ordinal Level data. One category is higher than the next one; that is, "Superior" is higher rating than" good", "good" is higher than "average", and so on.

If 1 is substituted for "superior", 2 substituted for 'good' and so on, a 1 ranking is obviously higher than a 2 ranking, and a 2 ranking is higher than a 3 ranking. However it cannot be said that (as an example) a company commander rated good is twice as competent as one rated average, or that a company commander rated superior is twice as competent as one rated good. It can only be said that a rating of superior is greater than a rating of good, and a good rating is greater than an average rating.

The major difference between a nominal level and an ordinal level of measurement is the "greater than" relationship between the ordinal-level categories. Otherwise, the ordinal seal of measurement has the same characteristics as the nominal scale; namely, the categories are mutually exclusive and exhaustive.

3. ***Interval level:*** The interval scale of measurement is the next higher level. It includes all the characteristics of the ordinal scale, but in addition, the distance between values is a constant size. If one observation is greater than another by a certain amount, and the zero point is arbitrary, the measurement is on at least an interval scale. For example, the difference between temperatures of 70 degrees and 80 degrees is 10 degrees. Likewise, a temperature of 90 degrees is 10 degrees more than a temperature

of 80 degrees, and so on. Scores on a statistics or mathematics examination are also examples of the interval scale of measurement.

4. ***Ratio level:*** *Ratio level is the highest level of measurement. This level has all the characteristics of interval level. The distances between numbers are of a known, constant size; the categories are mutually exclusive, and so on.*

The major differences between interval and ratio levels of measurement are these: (1) Ratio-level data has a meaningful zero point and (2) the ratio between two numbers is meaningful. Money is a good illustration having zero dollars has meaning you have none! Weight is another ratio-level measurement.

If the dial on a scale is zero, there is a complete absence of weight. Also, if you earn $40,000 a year and John earns $ 10,000, you earn four times what he does. Likewise, if you weigh 80 kg. and John weight 40 kg., you weigh twice John. But such comparisons are impossible in interval level of measurement.

Summary: The term Statistics may be defined as a data or as a method. Defined as a data, Statistics is the aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a pre-determined purpose and placed in relation to each other. On the other hand, Statistics, as a method, refers to the principles and methods used in the collection, presentation, analysis and interpretation of numerical data in any sphere of enquiry.

Terms like sample, population, variable, etc are the most commonly used ones in discussions related to Statistics. Data may be collected or obtained from two major sources, primary and secondary sources. Data obtained from a primary (original) source is called primary data. On the other hand, data taken from secondary sources, like magazines, newspapers and other publications, are called secondary data. Primary data is collected through different methods including interview, questionnaire, and direct or physical observation.

Broadly speaking there are two types of Statistics. These are descriptive and inferential Statistics. In descriptive Statistics, we simply describe what is contained in our data set. In inferential Statistics, however, we make conclusion or inference about population parameters based on sample results. Inferential Statistics is characterized by making generalization or conclusion.

A number of functions and limitation of Statistics are also discussed in fairly detailed manner. Some of the functions are condensation, definiteness, facilitating comparison, predictions formulation of policies, etc. In contrast, some of the limitations of Statistics are, it deals only with quantitative data, it does not deal with individual observations and is liable to be misused.

Finally, the chapter addressed the different stages of a Statistical investigation namely data collection, organization, presentation, analysis and interpretation.

UNIT 2: Collection, organization and presentation of data

2.1. Collection of data

Meaning of collection of data

Collection of data implies a systematic and meaningful assembly of information for the accomplishment of the objective of a statistical investigation. It refers to the methods used in gathering the required information from the units under investigation.

The quality of data greatly affects the final output of an investigation. Hence, utmost care should be attached to the data collection process and every possible precaution should be taken to ensure accuracy while collecting data. Otherwise, with inaccurate and inadequate data, the whole analysis is likely to be faulty and also the decisions to be taken will also be misleading.

2.1.1. Primary and secondary data

i. **Meaning and distinction between primary and secondary data**

Statistical data may be obtained either from primary or secondary source. A primary source is a source from where first-hand information is gathered. On the other hand, secondary source is the one that makes data available, which were collected by some other agency. Clearly, a source, which is not primary, is necessarily a secondary source. Primary sources are original sources of data.

Data obtained from a primary source is called primary data. Likewise, data gathered from a secondary source is known as secondary data. For example, assume that a simple study is to be conducted to see the age distribution of HIV/AIDS victim citizens. Clearly, the variable of study is age. Data about the age of HIV/AIDS victim citizens may be obtained by making direct interview with the victims. Note, in this specific case, the victim citizens are primary sources. Moreover, the data to be collected from them are primary data. Alternatively, one may use records of hospitals and other related agencies to obtain the age of the victim citizens without the need of tracing the victims personally. Therefore, the records of the hospitals, in our case, are secondary sources and the data copied from such records are secondary data.

In most cases, secondary data is obtained from such sources as census and survey reports, books, official records, reported experimental results, previous research papers, bulletins, magazines, newspapers, web sites, and other publications. Different organizations and government agencies publish information (data) in the form of reports, periodicals, journals, etc. In the case of Ethiopia, the Central Statistics Authority (CSA) is the first to be mentioned in publishing such relevant information (secondary data).

ii. Advantages and Disadvantages of Primary and Secondary data

The following are major advantages of primary data over that of secondary data.

- ✓ The primary data gives more reliable, accurate and adequate information, which is suitable to the objective and purpose of an investigation.
- ✓ Primary source usually shows data in greater detail.

- ✓ Primary data is free from errors that may arise from copying of figures from publications, which is the case in secondary data.

The disadvantages of primary data are:

- ✓ The process of collecting primary data is time consuming and costly.
- ✓ Often, primary data gives misleading information due to lack of integrity of investigators and non-cooperation of respondents in providing answers to certain delicate questions.

Advantage of Secondary data:

- ✓ It is readily available and hence convenient and much quicker to obtain than primary data,
- ✓ It reduces time, cost and effort as compared to primary data,
- ✓ Secondary data may be available in subjects (cases) where it is impossible to collect primary data. Such a case can be regions where there is war.

Some of the disadvantages of Secondary data are:

- ✓ Data obtained may not be sufficiently accurate,
- ✓ Data that exactly suit our purpose may not be found,
- ✓ Error may be made while copying figures.

The choice between primary data and secondary data is determined by factors like nature and scope of the enquiry, availability of financial resources, availability of time, degree of accuracy desired, and the collecting agency. Often, primary data are used in situations where secondary data do not provide adequate basis of analysis. Meaning, when the secondary data do not suit a specific investigation we use primary data. Unless for such cases, most statistical investigations rest up on secondary data since it minimizes cost and saves time. Nevertheless, the following points should be carefully considered while using secondary data in our investigation.

- One should closely examine whether or not the data are suitable for the intended study,

- The source of data should be viewed, keeping in mind whether, at any time, it is reliable or not. If there is any doubt about the reliability of data, it should not be used,
- It should be noted that the data is not obsolete,
- In case the data are based on a sample, one should see whether the sample is a proper representative of the population,
- It should be the case that skilled persons only have handled the primary data carefully.

Finally, it should be clear that primary data in the hands of one person might be secondary in the hands of another. That is why it is often said, "*the difference between primary and secondary data is largely one of degree*."

iii. Methods of collecting primary data

After discussing the two sources of data, primary and secondary, it is logical to say a few words about the methods employed in collecting data from its original or primary source.

Many authors commonly state three methods of collecting primary data. These are:

a. Personal Enquiry Method (Interview method)
b. Direct Observation
c. Questionnaire method

a) Personal Enquiry Method (Interview method)

In personal enquiry method, a question sheet is prepared which is called schedule. The schedule contains all the questions, which would extract a complete report from a respondent. Usually, schedules are pre-tested so as to remove certain discrepancies like ambiguities of the questions and irrelevant questions. This pre-testing process is called a pilot survey. It is worth mentioning that the schedule is not directly given to the respondent. Rather, it is the interviewer who asks those questions on the schedule and jot down the interviewee's (respondent's) response. Depending on the nature of the interview, personal enquiry method is further classified into two types.

- ***Direct Personal Interview***: It is a type of personal enquiry where there is a face-to-face contact with the persons from whom the information is to be obtained. In other words, the investigator contacts each respondent personally, without the interference of third party, and asks questions given in the schedule one by one and notes down respondent's replies on the schedule.
- ***Indirect Personal Enquiry (Interview):*** It is the second type of personal enquiry where the investigator contacts third parties called witnessed who are capable of supplying the necessary information. Here, the information is not collected directly from the respondent but from a third person who knows the respondent well. Such an approach is useful in case where the respondent is expected to conceal information about him or her. For example, if an enquiry about the habit of using condoms is distributed in a village, most of the villagers may not provide the correct information. Thus, it would be wiser to get the required information from other parties, like the nearby condom dealing shop.

b) Direct Observation

In this approach, an investigator stays at the place of survey and notes down the observation himself. There is no enquires in the case of direct observation. For example, an investigator making a study on nutritional status of children may directly (physically) measure the weight, height, and other required parameters himself/herself. Direct observation is more experimental and usually applied in scientific studies. It is time consuming and also costly.

c) Questionnaire Method

Under this method, a list of questions related to the survey is prepared and sent to the various respondents by post, Web sites, e-mail, etc. However, this method cannot be used if the respondent is illiterate. It is a method that is often used in many statistical investigations.

The following are the major points that we need to take into account while preparing a questionnaire.

- The number of questions should be small. Naturally, respondents are not comfortable with lengthy questionnaires. Lengthy questionnaires usually bore respondents. Hence, fifteen to twenty five questions in a questionnaire are optimal. If a lengthy questionnaire is unavoidable, it should preferably be divided into two or more parts.

- The questions should be short, clear, simple and unambiguous. Moreover, the questions must be arranged in a logical order so that natural and spontaneous reply to each is induced. For instance, it is not appropriate to ask a person how many packets of cigarette he/she smokes before asking whether he/she smokes or not.

- Questions of sensitive nature should be avoided. Sensitive questions are those questions that are too personal and pecuniary like " Sources of income", "Drinking habit", etc. The logic here is that respondents do not willingly answer sensitive questions. Such information, if necessary, may be gathered through interview or through other indirect questions.
- Questions should be capable of objective answers. As much as possible, avoid subjective questions and keep to questions of fact. To this end, multiple answer questions can be used.

- Mail questionnaires should be accomplished by a covering letter, which should state the purpose of the questionnaire, promise of confidentiality of responses, etc.

Furthermore, the questions preferably designed in such a way can easily be answered as yes/no.

2.2. Classification of data

After collecting relevant information (data) for the purpose of statistical investigation, the next important task is classification and presentation of this data. It is difficult to group the meaning of any considerable volume of numerical data unless their mass is some hours reduced to relatively few convenient classes or categories and presented with the help of some kinds of visual aid.

This section discusses classification of data. Presentation of data using graphs and charts will be seen in the next unit.

Definition of classification of data:- is the process of arranging things in groups or classes according to their resemblance.

Purposes of Classification: -

- To eliminate unnecessary detail.
- To bring out clearly points of similarity & dissimilarity
- To enable one to form mental pictures of objects on measurements
- To enable one to make comparisons and draw inferences

Basically there are four types of classification.

1. **Geographical Classification**: - Data are arranged according to places like continents, regions, and countries

Example.

Region	Common Language Spoken
1	Tigrigna
2	Afar
3	Amharic
4	Oromifa

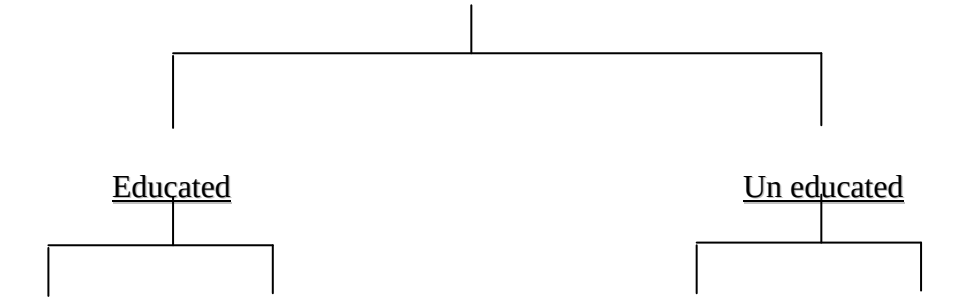
2. **Chronological Classification**:- Data are arranged according to time like month, year

Example.

Year (in EC)	Population (in million)
1974	30
1986	52
1991	60

3. **Qualitative Classification**: - Data are arranged according to attributes like color, religion, marital-status, sex, educational background, etc.

Example. ***Employees in a Factory*** X



4. **Quantitative Classification**:- In this type of classification, the statistical data is classified according to some quantitative variables. The variable may be either discrete or continuous.

Example**.**

Mr. x	Height (X) in cm
A	160
B	182
C	175
D	178

Note: There are two kinds of variables, which can have values: Discrete Variable and Continuous Variable.

a. Discrete Variables – are variables that are associated with enumeration or counting

Example: i. Number of students in a class

ii.Number of children in a family, etc

b. Continuous Variables – are variables associated with measurement.

Example : i. The weights of 10 students,

ii. The heights of 12 persons,

iii. Distance covered by a car between two stations etc.

2.3. Frequency distribution

When the raw data have been collected, they should be put in to an ordered array in an ascending or descending order so that it can be looked at more objectively. Then this data must be organized in to a "FD" which simply lists the values or classes with their corresponding frequencies in a tabular form. Here, frequency refers to the number of observations a certain value occurred in a data.

The tabular representation of values of a variable together with the corresponding frequency is called a Frequency Distribution (FD).

Definition: A frequency distribution is the organization of raw data in table form, using classes and frequencies. Frequency distribution is of two kinds.

a. Ungrouped Frequency Distribution (UFD)

Shows a distribution where the values of a variable are linked with the respective frequencies.

Example . Consider the number of children in 15 families.

1	0	3	2	0
2	4	1	3	1
4	1	2	2	3

Construct ungrouped FD for the above data.

Sol.:

No. of Children (Values)	No. of Family (Tallies)	Frequency
0	/ /	2
1	/ / / /	4
2	~~/ / / /~~	5
3	/ / /	3
4	/ /	2
		Total 16

Example: Consider the following scores in a statistics test obtained by 20 students in a given class. 10, 4, 4, 7, 5, 7, 7, 8, 5, 7, 8, 5, 10, 8, 7, 5, 7, 8, 7, 4.

a. Prepare an ungrouped FD

b. Grouped Frequency Distribution (GFD)

Raw Data: Data collected in original form.

Frequency: The number of values in a specific class of the distribution or the number of times a value occurs in the distribution.

Cumulative Frequencies: refer to the total frequency of all values up to and including the upper boundary of the class interval that is under consideration.

Frequency Distribution: A table showing classes or values with their corresponding frequencies.

Class: In set refers to a group of data considered as one item in a frequency distribution.

Range: Means the difference between the largest and the smallest values in a set of data.

Class Interval: Refers to difference between class limits (boundaries).

Class Limits: Means limits of different classes in a frequency distribution.

Class Boundaries: Boundaries that are obtained by adding and subtracting half of unit of measurement.

If the mass of the data is very large, it is necessary to condense the data in to an appropriate number of classes or groups of values of a variable and indicate the number of observed values that fall in to each class. Therefore, a GFD is a frequency distribution where values of a variable are linked in to groups & corresponded with the number of observations in each group.

Example

Values (x_i)	1 - 25	26 - 50	51 - 75	76 - 100
Frequency (f_i)	3	10	18	6

Common terminologies in a GFD

i. **Class**:- group of values of a variable between two specified numbers called lower class limit (LCL) & upper class limit (UCL) . In the above Example**,**

The GFD contains four classes: that are 1 – 25, 26 – 50, 51 – 75, and 76 – 100

$LCL_1 = 1$, $UCL_1 = 25$ $LCL_3 = 51$, $UCL_3 = 75$

$LCL_2 = 26$, $UCL_2 = 50$ $LCL_4 = 76$, $UCL_4 = 100$

ii. Class Frequency (or Simply Frequency): refers to the number of observations corresponding to a class.

In the above example the class frequency of the 1^{st}, 2^{nd}, 3^{rd}, & 4^{th} classes are respectively 3, 10, 18 and 6.

iii. Class Boundaries: are boundaries obtained by subtracting half of the unit of measurement (u) from the lower limits or by adding ½ (u) on the upper limits of a class.

i.e $UCB_i = UCL_i + ½ (u)$ and $LCB_i = LCL_i - ½ (u)$

Where UCB_i = Upper Class Boundaries and LCB_i = Lower Class Boundaries

Remark: The unit of measurement (u) is the gap between any two successive classes. i.e

u = lower limit of a class – upper limit of the preceding class.

In the above example, consider the 2^{nd} class, 26 – 50, since u = 26 – 25 = 1, $LCL_2 = 26$ and $UCL_2 = 50$ $LCB_2 = 26 - ½(1) = 25.5$ $UCB_2 = 50 + ½(1) = 50.5$

iv. Class Width (size of a class or class interval): it is the difference between the upper and lower class limits or the difference between the upper and lower class boundaries of any class.

Remarks:

1. If both the LCL & UCL are included in a class, it is called an inclusive class. For inclusive classes, Class width (cw) = $UCB_i - LCB_i$

2. If LCL is included and the UCL is not included in a class, it is called an exclusive class. For exclusive classes $cw = UCL_i - LCL_i$

To be consistent, we use inclusive classes.

v. Class Mark (cm): it is the mid-point (center) of a class

$$cm_i = \frac{UCB_i + LCB_i}{2}$$

Note:- The difference between any two successive class marks is equal to the width of a class.

vii. Range (R) : is the difference between the largest (L) and the smallest (S) values in a data , $R = L - S$.

Example: Consider the following GFD

Class	Frequency (*f*)
5 – 9	2
10 – 14	6
15 – 19	12
20 – 24	7
25 – 29	3
	Total 30

a. What is the class frequency of the 3rd class?

b. How many observations (items) are linked into the last class?

c. Find i. The LCL and UCL of the fourth class

ii. The UCB and LCB of the third class

iii. The class interval (class width) of the fifth class

iv. The class mark (mid-point) of the second class

Rules for forming a grouped frequency distribution

To construct a GFD the following points should be considered

1) The classes should be clearly defined. That is each observation should fall in to on e & only one class.
2) The number of classes neither should either to be too larger nor should be too small. Normally, 5 to 20 classes are recommended
3) All the classes should be of the same width. An approximate suitable class width can be obtained as:

$$cw \approx \frac{Range}{Number\ of\ Classes} \quad i.e \quad cw = \frac{R}{n} = \frac{L - S}{n}$$

Example 1. Let $\frac{R}{n} \approx 6.8263$

If all the observations are whole numbers, cw = 7

If all the observations are to one decimal places, cw = 6.8

If all the observations are to two decimal places, cw = 6.83, etc.

Note that a suitable number of classes can be obtained by using the formula

$$n \approx 1 + 3.322 \log N.$$

up/down to the nearest whole number, where N is the total number of observations.

Remark Unequal class intervals create problem in graphing and computing some statistical measures

4) Determine the class limits

i. Determine the lower class limit of the first class (LCL_1), then

$LCL_2 = LCL_1 + cw$, $LCL_3 = LCL_2 + cw$,… $LCL_{i+1} = LCL_i + cw$

ii. Determine the upper class limit of the first class (UCL_1) i.e.

$UCL_1 = LCL_1 + cw - u$, where u = the unit of measurement, then

$UCL_2 = UCL_1 + cw$, UCL_3 UCL_2, … , $UCL_{i+1} = UCL_i + cw$

5) Complete the GFD with the respective class frequencies.

Example 2. The number of customers for consecutive 30 days in a supermarket was listed as follows:

20	48	65	25	48	49
35	25	72	42	22	58
53	42	23	57	65	37
18	65	37	16	39	42
49	68	69	63	29	67

a. construct a GFD with a suitable number of classes

b. complete the distribution obtained in (a) with class boundaries & class marks

Sol.: i. Range = Largest value – smallest value

= 72 – 16 = 56

ii. N = 30 (total number of observations)

$\therefore$ number of classes, $n = 1 + 3.322 \log 30$

$\Rightarrow n = 1 + 3.322 \log 30$

$= 1 + 3.322 (1.4771)$

$= 5.9$

Hence a suitable number of class n is chosen to be 6

iii. Class width = $\frac{Range}{n} = \frac{56}{6} = 9.33 =$ **cw**

For the sake of convenience, take **cw** to be 10 (note that it is also possible to choose the **cw** to be 9).

iv. Take lower limit of the 1st class (LCL_1) to be 16 & u = 1

i.e. $LCL_1 = 16$ and $UCL_1 = LCL_1 + cw - u = 16+10-1 = 25$

$LCL_2 = LCL_1 + cw = 16 + 10 = 26$ $UCL_2 = UCL_1 + cw = 25 + 10 = 35$

$LCL_3 = LCL_2 + cw = 26 + 10 = 36$ $UCL_3 = UCL_2 + cw = 35 + 10 = 45$

Therefore, the GFD would be

a)

Class (x_i)	Frequency (f_i)
16 – 25	7
26 – 35	2
36 – 45	6
46 – 55	5
56 – 65	6
66 – 75	4

b)

Class (x_i)	Frequency (f_i)	CB_i	cm_i
16 – 25	7	15.5 – 25.5	2.05
26 – 35	2	25.5 – 35.5	30.5
36 – 45	6	35.5 – 45.5	40.5
46 – 55	5	45.5 – 55.5	50.5
56 – 65	6	55.5 – 65.5	60.5
66 – 75	4	65.5 – 75.5	70.5

Exercise: Construct a grouped frequency distribution for the following ages of 50 persons with 6 classes.

37	40	69	35	36	70	72	62	36	72
65	64	47	59	55	42	45	50	46	65
54	63	51	50	61	60	58	58	56	58
55	45	49	51	50	56	44	60	70	44
52	43	55	46	42	62	57	48	60	55

Cumulative frequency distribution (CFD)

It is the collection of values of a variable above or below specified values in a distribution. GFD is of two types.

a. **'Less Than' Cumulative Frequency Distribution (<CFD):** shows the collection of cases lying below the upper class boundaries of each class.

b. **'More Than' Cumulative Frequency Distribution (>CFD):** shows the collection of cases lying above the lower class boundaries of each class.

Remark: The frequency distribution does not tell us directly the number of units above or below specified values of the classes this can be determined from a "cumulative Frequency Distribution'

Example 3: Consider the frequency distribution in example 2

Class (x_i)	Frequency (f_i)	Less than Cumulative Frequency ($<cf_i$)	More than Cumulative Frequency ($>cf_i$)
3 - 6	4	4	30
7 – 10	7	11	26
11 – 14	10	21	19
15 – 18	6	27	9
19 – 22	3	30	3

This means that from 'less than' cumulative frequency distribution there are 4 observations less than 6.5, 11 observations below 10.5, etc and from 'more than' cumulative frequency distribution 30 observations are above 2.5, 25 above 6.5 etc.

Relative frequency distribution (RFD)

It enables the researcher to know the proportion or percentage of cases in each class. Relative frequencies can be obtained by dividing the frequency of each class by the total frequency. It can be converted in to a percentage frequency by multiplying each relative frequency by 100%. i.e. $Rf_i = \frac{f_i}{n}$

Where Rf_i – is the relative frequency of the i^{th} class

f_i – is the frequency of the i^{th} class

n – is the total number of observations

Note: $\boldsymbol{Pf_i = Rf_i \times 100\%}$ **w**here Pf_i is percentage frequency of each class.

Example 4: The relative and percentage of frequency distribution of Example 3 is :

x_i	f_i	Rf_i	%freq. (Pf_i)
3 – 6	4	4/30	4/30 × 100
7 – 10	7	7/30	7/30 × 100
11 – 14	10	10/30	10/30 × 100
15 – 18	6	6/30	6/30 × 100
19 – 22	3	3/30	3/30 × 100
Total	30	1	100%

Summary: This unit discussed the definitions of classification of data and a frequency distribution. In order to describe situations, draw conclusions or make

inferences about random events, one must organize the data in some meaningful way. The most convenient method of organizing data is to construct a frequency distribution.

Therefore, a frequency distribution was seen as a distribution showing the correspondence of values or classes with their respective frequencies.

Presentation of data

Definition: Presentation is a statistical procedure of arranging and putting data in a form of tables, graphs, charts and/or diagrams.

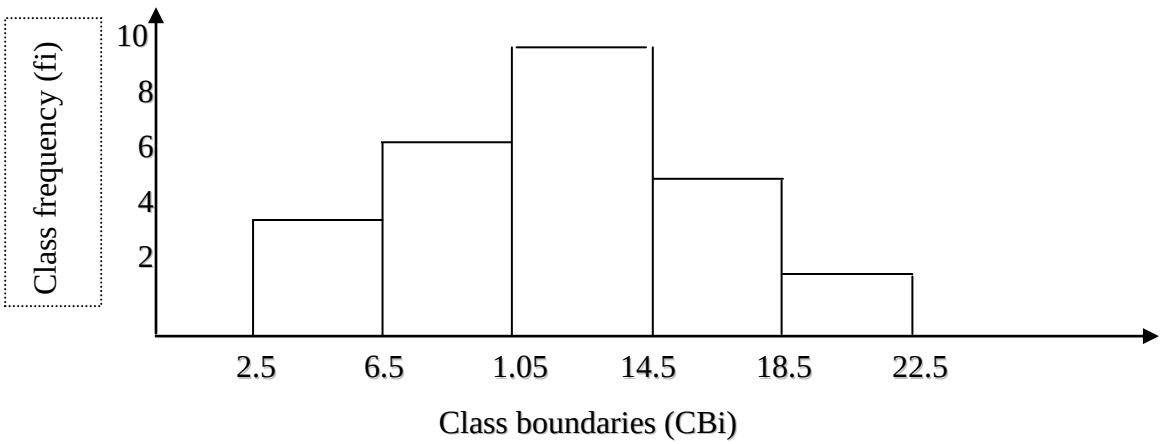
Histogram: After you complete a frequency distribution, your next step will be to construct a "picture" of these data values using a histogram. A histogram is a graph consisting of a series of adjacent rectangles whose bases are equal to the class width of the corresponding classes and whose heights are proportional to the corresponding class frequencies. Here, class boundaries are marked along the horizontal axis (x – axis) and the class frequencies along the vertical axis (y – axis) according to a suitable scale. It describes the shape of the data. You can use it to answer quickly such questions a,s are the data symmetric? And where do most of the data values lie?

Example1. Considers the following GFD and construct a histogram

Class (x_i)	Frequency (f_i)
3 – 6	4
7 – 10	7
11 – 14	10
15 – 18	6
19 - 22	3
Total 30	

Sol.:

Histogram for the above distribution



Example 1: Construct a histogram for the following distribution

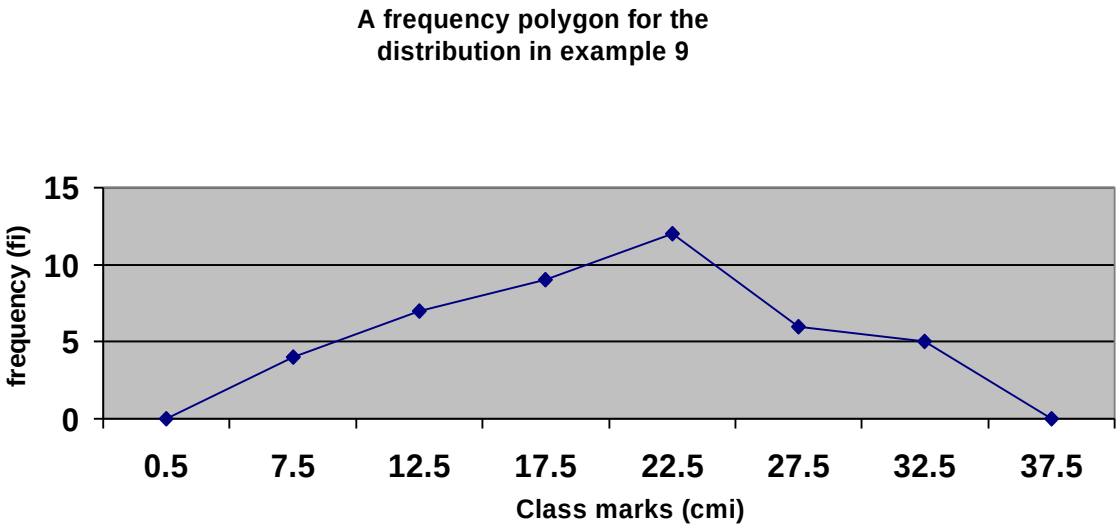
Class (x_i)	Frequency (f_i)
5 – 10	4
10 – 15	7
15 – 20	9
20 – 25	12
25 - 30	6
30 – 35	5

Frequency polygon

It is a line graph of frequency distribution. Although a histogram does demonstrate the shape of the data, perhaps the shape can be more clearly illustrated by using a frequency polygon. Here, you merely connect the centers of the tops of the histogram bars (located at the class midpoints) with a series of straight lines. The resulting figure is a frequency polygon. Here the class marks are plotted along the x – axis and the class frequencies along the y – axis. Empty classes are include at each end so that the curve will anchor with the x – axis.

Example 2. Construct a frequency polygon for the frequency distribution given in example1

Sol.:



Example 3. Construct a frequency polygon for the frequency distribution given under Example 1

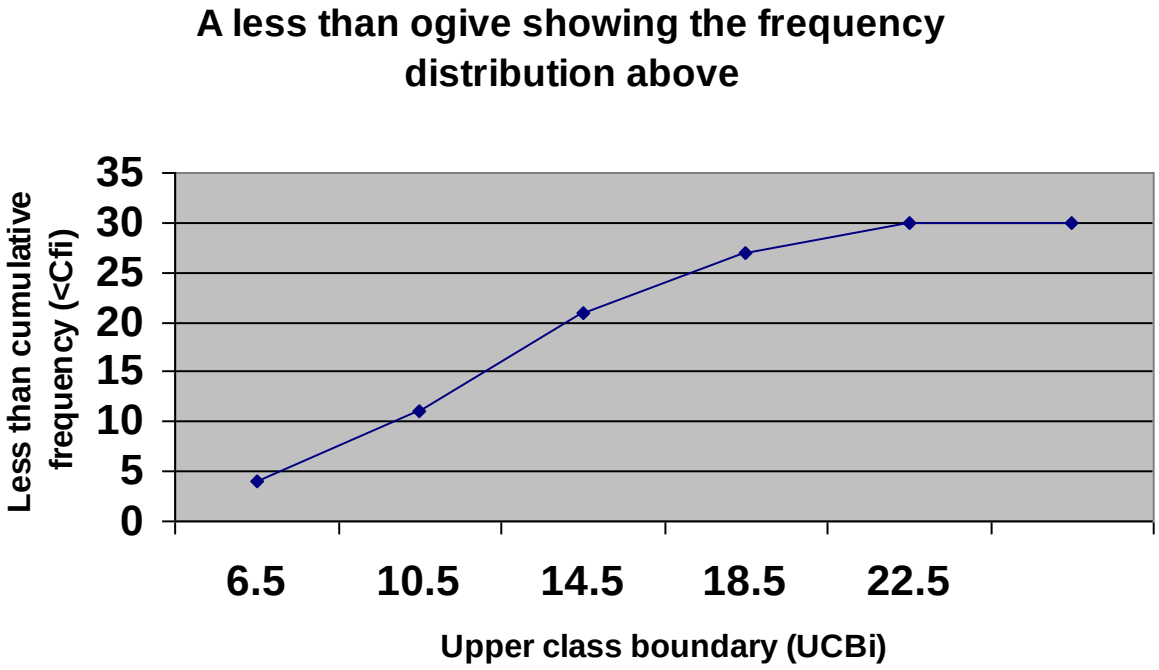
Cumulative frequency curve, (o-give)

It is the graphic representation of a cumulative frequency distribution Ogives are of two kinds. 'Less than' ogive and 'more than' Ogive ie < Ogive and > Ogive.

A) '**Less than' ogive**: here, upper class boundaries are plotted against the 'less than' cumulative frequencies of the respective class & they are joined by adjacent lines.

Example 4. Draw a 'less than' ogive for the frequency distribution in example 1.

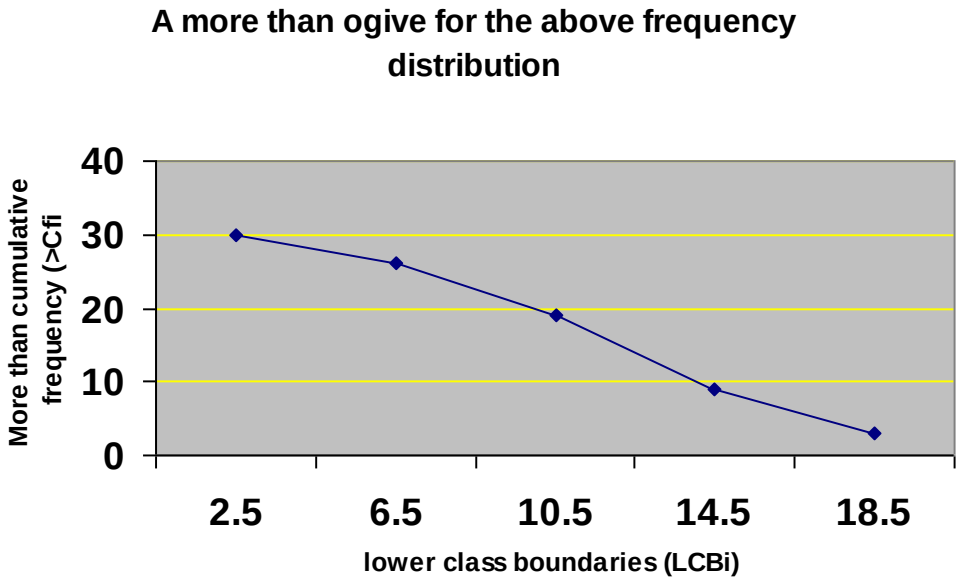
Sol.:



B) '**More than' ogive**: here, lower class boundaries are plotted against the 'more than' cumulative frequencies of their respective class and they are joined by adjacent lines.

Example 5. Draw a 'More than' ogive for the frequency distribution in example 1.

Sol.:



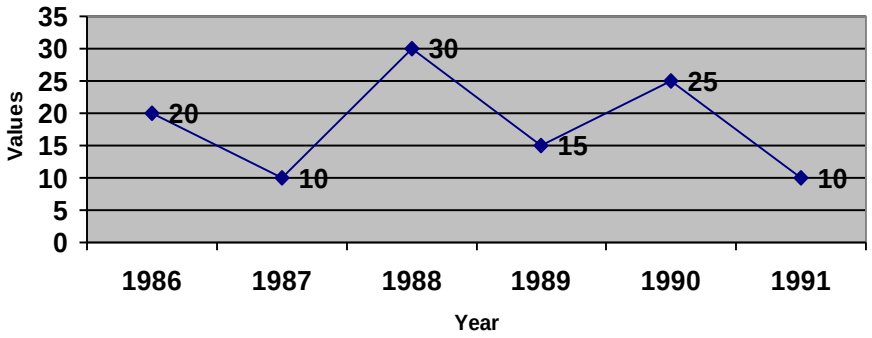
Line graph

It represents the relationship between time (on the x-axis) and values of variable (on the y-axis). The values are recorded with respect to the time of occurrence.

Example : Draw a line graph for the following time series.

Year	1986	1987	1988	1989	1991
Values	20	10	30	15	1

A line graph showing the above time series



Sol.:

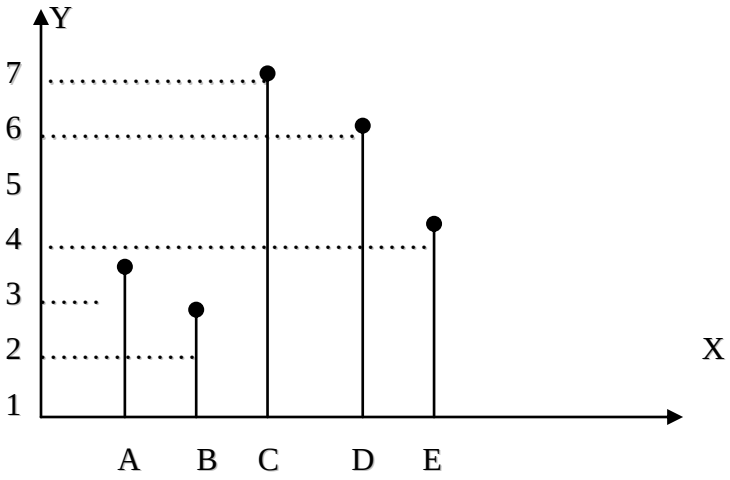
Vertical line graph

Is a graphical representation of discrete data (or characteristics expressed with whole numbers) with respect to the frequencies? Vertical solid lines are used to indicate the frequencies.

Example : Draw a vertical line graph for the following data

Family	A	B	C	D	E
Number of children	3	2	7	6	4

Sol.:



Vertical line graph showing number of children in family A, B, C, D and E

Bar chart (Bar diagram)

Histogram, Frequency polygon, ogives are used for data having an interval or ratio level of measurement. The other kinds of presenting statistical data suitable for a particular kind of situations are bar charts, pie chart and pictograph.

Bar chart is a series of equally spaced bars of uniform width where the height (length) of a bar represents the amount (magnitude) of frequency corresponding with a category. Bars may be drawn horizontally or vertically. Vertical bar graphs are preferred as they allow comparison with other bars.

Types of bar charts

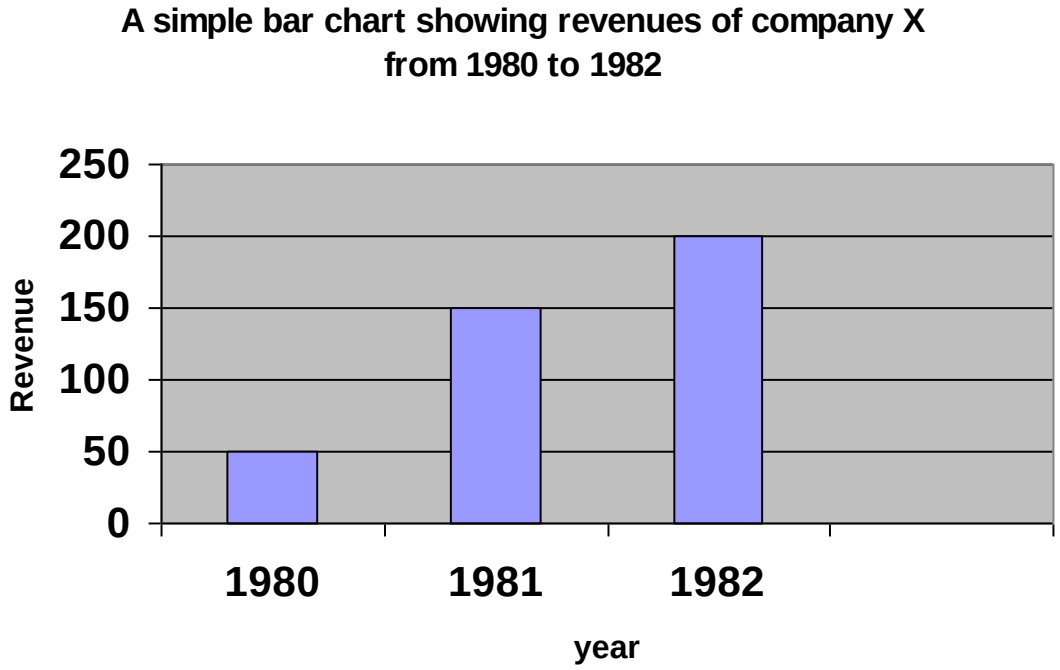
a. Simple Bar Chart:

It represents a single set of data (variable) classified in different categories. Singular bars are drawn with the respective frequencies.

Example: Revenue (in millions of Birr) of company x from 1980 to 1982 is given below

Year	Revenue
1980	50
1981	150
1982	200

Sol.:



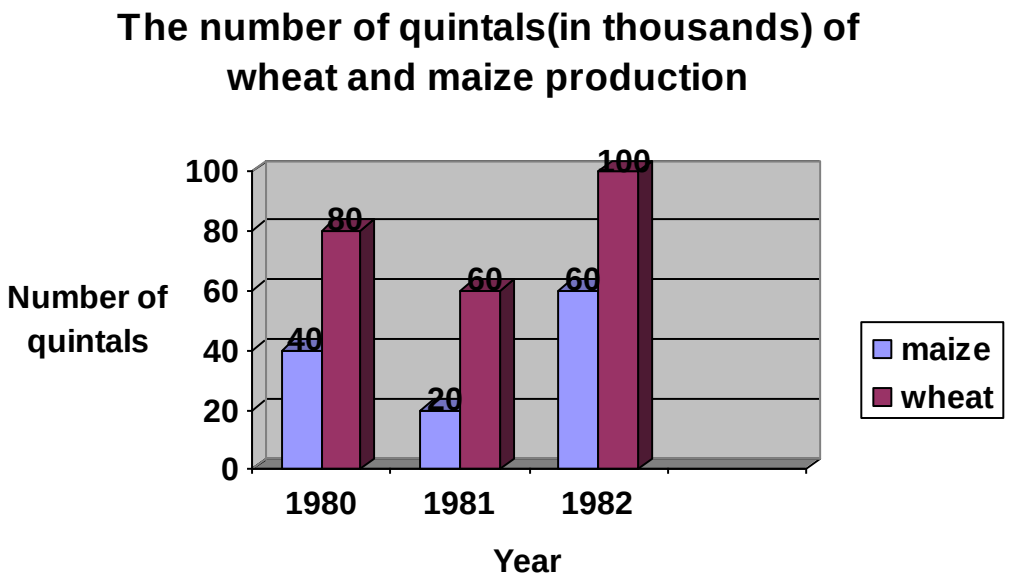
b. **Multiple Bar Chart:**

Here two or more bars are grouped with the corresponding frequency to represent two or more interrelated data in each category. The bars of related variables are kept adjacent to each other for every set of values. These charts can be used if the overall total is not required and each bar is shaded or colored separately and a key is given to distinguish them.

Example: The following table shows the production of wheat and maize in hundreds of quintals.

Year	Maize	Wheat
1980	40	80
1981	20	60
1982	60	100

Sol.:



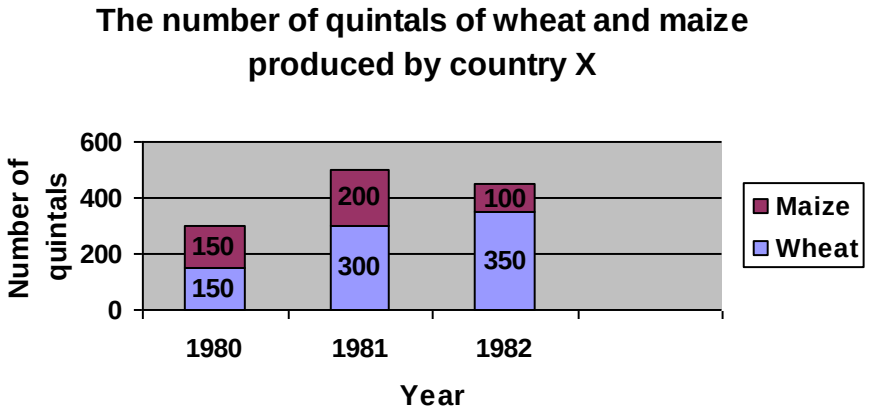
c. Subdivided Bar Chart:

It is used to present data by subdividing a single bar with respect to the proportional frequency. Each portion of the bar is then shaded or colored and a key is give to distinguish them.

Example: The number of quintals of wheat and maize (in millions of quintals) produced by country x in the indicated years.

Year	Wheat	Maize
1980	150	150
1981	300	200
1982	350	100

Sol.:



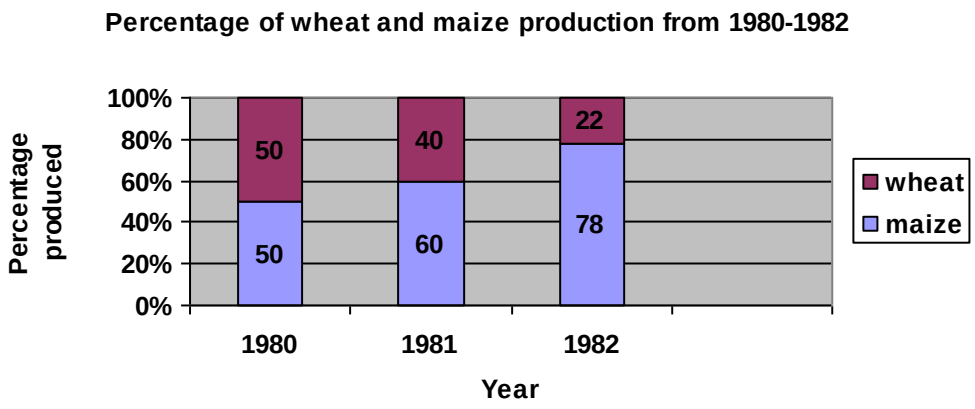
d. Percentage Bar Chart:

It is a subdivided bar chart where percentages are used in each classification rather than the actual frequencies.

Example: construct percentage bar chart for the data in Example 19.

Sol.:

Year	% of Wheat Production	% of Maize Production
1980	150/300 × 100 = 50	150/300 × 100 = 50
1981	300/500 × 100 = 60	200/500 × 100 = 40
1982	350/450 × 100 = 78	100/450 × 100 = 22



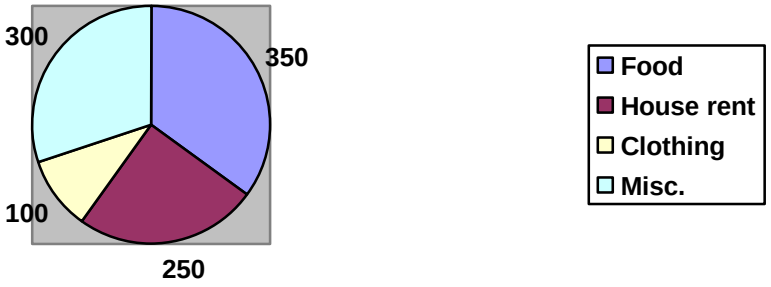
Pie chart

A pie chart is a circle divided in to various sectors with areas proportional to the value of the component they represent. It shows the components in terms of percentages not in absolute magnitude. The degree of the angle formed at the center has to be proportional to the values represented.

Example : the monthly expenditure of a certain family is given below.

Items	Expenditure	% Proportion (Pf_i)	Degrees ($360^o\ Rf_i$)
Clothing	100	$100/1000 \times 100 = 10$	$100/1000 \times 360^o = 36$
Food	350	$350/1000 \times 100 = 35$	$350/1000 \times 360^o = 126$
House Rent	250	$250/1000 \times 100 = 25$	$250/1000 \times 360^o = 90$
Miscellaneous	300	$300/1000 \times 100 = 30$	$300/1000 \times 360^o = 108$
Total	1000	100%	360^o

Sol.: The pie chart for the above expenditure is as follows



.

Pictograph (pictogram):

A pictograph is a graph that uses symbols or pictures to represent data.

Example: In comparing the population of a country from 1990 to 1992, we simply draw pictures of people where each picture may represent 1000,000 people.

1992 - 🧍 🧍 🧍 🧍 🧍 🧍 🧍 Key: 🧍 = 1000,000

1991 - 🧍 🧍 🧍 🧍 🧍 🧍 🧍 🧍 🧍 🧍

1990 - 🧍 🧍 🧍 🧍 🧍 🧍 🧍 🧍 🧍 🧍 🧍 🧍 🧍 🧍 🧍

Summary: This unit discussed how to present the organized data. Once a frequency distribution is constructed, the representation of the data by using graphs is a simple task. The most commonly used graphs in research statistics are the histograms, frequency polygon, an o'give, and other graphs and diagrams, like the bar charts, pie charts, pictograms can also be used. And some of these graphs are seen frequently in newspapers, magazines, and various statistical reports.