CHAPTER ONE: Introduction (3 lecture hours)

1.1	Definition and classification of Statistics
1.2	Stages in statistical investigation
1.3	Definition of some basic terms
1.4	Applications, uses and limitations of Statistics
1.5	Types of variables and measurement scales

At the end of this chapter students will be able to:

- Explain statistics
- identify types of level of measurement
- make a distinction between descriptive statistics and inferential statistics
- identify applications and limitations of statistics

1.1. Definitions and classification of Statistics

Definitions:

Statistics is defined differently by different authors over period of time. In the olden days statistics was confined to only state affairs but in modern days it embraces almost every sphere of human activity. Therefore, a number of old definitions, which were confined to narrow field of enquiry, were replaced by more definitions, which are much more comprehensive and exhaustive.

We can define it in two senses

- **In the plural sense:** statistics are the raw data themselves (Numerical facts) , like statistics of births, statistics of deaths, statistics of students, statistics of imports and exports, etc.
- **In the singular sense:** statistics is concerned with scientific methods for collecting, organizing, summarizing, presenting and analyzing data as well as deriving valid conclusions and making reasonable decisions on the basis of this analysis.

Classifications:

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

1. **Descriptive Statistics**:
   - Is concerned with summary calculations, graphs, charts and tables.
   - Generally characterizes or describes a set of data elements by graphically displaying the information or describing its central tendencies and how it is distributed.
   - In descriptive statistics our objective is to describe a group of data that we have 'in hand' i.e. data that are accessible to us.
   - We are not interested in other data that we are not gathered.

Example 1: the following data refers to the number of malaria patients who have been treated in Black Lion Hospital from 1986 to 1990 (Eth. Calendar).

3645; 4568; 5432; 6751; 7369

If we calculate the average malaria patients from 1986 to 1990 as

$Average = \frac{1}{5}(3645 + 4568 + 5432 + 6751 + 7369) = 5553$ then our work belongs to the domain of descriptive statistics.

If we say that there was an increase of 724 patients from 1986 to 1990, then again this belongs to the domain of descriptive statistics.

Example 2: Suppose that the mark of 6 students in Statistics course for Mathematics is given as 40, 45, 50, 60, 70 and 80. The average mark of the 6 students is 57.5 and it is considered as descriptive statistics.

2. **Inferential Statistics**: consists of generalizing from samples to populations, performing estimations and hypothesis tests, determining relationships among variables, and making predictions. Statistical techniques based on probability theory are required**.**

Example 1: In the above example if we predict the number of malaria patients in the year 1995 to be 9917, then our work belongs to the domain of inferential statistics.

Example 2: Suppose we want to have an idea about the percentage of illiterates in our country. We take a sample from the population and find the proportion of illiterates in the sample. This sample proportion with the help of probability enables us to make some inferences about the population proportion. This study belongs to inferential statistics.

Example 3: In the above example, if we say that the average mark in Statistics course for Mathematics students is 57.5, then we talk about inferential statistics (draw conclusion based on the sample observation).

1.2. 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. This includes formulating the problem, and then collecting, organizing and classifying, presenting, analyzing and interpreting of statistical data. Let's see each stage in detail

I. **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. Ask as many questions as necessary because "*An approximate answer to the right question is worth a great deal more than a precise answer to the wrong question*." - *The first golden rule of applied mathematics*-

Therefore, the first stage in any statistical investigation should be to:

- Get a clear understanding of the physical background to the situation under study;
- Clarify the objectives;
- Formulate the objective in statistical terms

II. ***Proper collection of data***: in order to draw valid conclusions, it is important 'good' data. Data are gathered with aim to meet predetermine objectives. In other words, the data must provide answers to problems. The data itself form the foundation of statistical analyses and hence the data must be carefully and accurately collected. In section 1.6 we will see the methods of data collection.

III. ***Organization and classification of data***: in this stage the collected data organized in a systematic manner. That means the data must be placed in relation to each other. The classification or sorting out of data is, by itself, a kind of organization of data.

IV. ***Presentation of data***: The purpose of putting the organized data in graphs, charts and tables is two-fold. First, it is a visual way to look at the data and see what happened and make

interpretations. Second, it is usually the best way to show the data to others. Reading lots of numbers in the text puts people to sleep and does little to convey information.

V. ***Analyses of data:*** is the process of looking at and summarizing data with the intent to extract useful information and develop conclusions. Data analysis is closely related to data mining, but data mining tends to focus on larger data sets, with less emphasis on making inference, and often uses data that was originally collected for a different purpose. In this stage different types of inferential statistical methods will apply. For instance, hypothesis testing such as χ^2 test of association.

VI. ***Interpretation of data:*** interpretation means drawing valid conclusions from data which form the basis of decision making. Correct interpretation requires a high degree of skill and experience.

Note that: Analyses and interpretation of data are the two sides of the same coin.

1.3. Definition of Some Terms

In this section, we will define those terms which will be used most frequently. These are:

Data: are the values (measurements or observations) that the variables can assume. Data can be facts or figures from which the conclusion can be drawn.

Data set: Facts or figures collected for a particular study. Each value in the data set is called data value or datum.

Raw Data: Data sheets are where the data are originally recorded. Original data are called raw data. Data sheets are often hand drawn, but they can also be printouts from database programs like Microsoft Excel.

Population: The totality of all subjects with certain common characteristics that are being studied in a specified time and place.

Sample: Is a portion of a population which is selected using some technique of sampling. Sample must be representative of the population so that it must be selected by any of the developed technique.

Parameter: Any measure computed from the data of a population.

Example: Populations mean (μ) and population standard deviation (σ)

Statistic: Any measure computed from the sample.

Example: sample mean $(\bar{x})$, sample standard deviation (S)

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

Survey: A collection of quantitative information about members of a population when no special control is exercised over any of the factors influencing the variable of interest.

Sample survey: A survey that include only a portion of the population.

Census: A collection of information about every member of a population

Sample survey has the following advantages over census

- Sample survey saves time and cost
- Has great accuracy
- Avoid wastage of material

Sampling: Is the process of selecting units (e.g., people, households, organizations) from a population of interest so that by studying the sample we may fairly generalize our results back to the population from which they were chosen. There are two types of sampling techniques namely random sampling technique and non-random sampling technique.

Probability sampling technique involves random selection that gives a non- zero chance for all elements of the population to be included in the sample. In other words, there is no personal bias regarding the selection. The five common random sampling techniques are:

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

Non-random sampling technique is mostly known as non-probability sampling techniques and in this case not all elements of a population have a known chance of inclusion or if some outcomes have a zero chance of being selected as a sample. The most familiar examples of non-random sampling techniques are

- Quota sampling
- Convenience sampling
- Volunteer sampling
- Purposive sampling
- Haphazard sampling
- Snow ball sampling etc…

Variable: A **variable** is a characteristic or attribute that can assume different values. Variables whose values are determined by chance are called random variables. Variables are often specified according to their type and intended use and hence variable can be classified in to two namely qualitative and quantitative variables.

- **A quantitative variable** is naturally measured as a number for which meaningful arithmetic operations make sense. Examples: Height, age, crop yield, GPA, salary, temperature, area, air pollution index (measured in parts per million), etc.
- **Qualitative variable:** Any variable that is not quantitative is qualitative. Qualitative variables take a value that is one of several possible categories. As naturally measured, qualitative variables have no numerical meaning. **Examples:** Hair color, gender, field of study, marital status, political affiliation, status of disease infection.

Quantitative variables can be classified as discrete and continuous variable. **Discrete variables** can assume certain numerical values. That is, there are gaps between the possible values. Such as 0, 1, 2...It may be countable finite or countable infinite. For example the number of students in a classroom, number of children a family. **Continuous variable** can take any value within a specified interval with a finite enough measuring device. No gaps between possible values. They are obtained by measuring. For example, consider the heights of two people no matter how close it is we can find another person whose height falls somewhere between the two heights is a continuous variable.

1.4. Applications, uses and limitations of statistics

I. Applications of statistics

- Apart from helping elicit an intelligent assessment from a body of figures and facts, statistics is indispensable tool for any scientific enquiry-right from the stage of planning enquiry to the stage of conclusion. It applies almost all sciences: pure and applied, physical natural, biological, medical, agricultural and engineering. It also finds applications in social and management sciences, in commerce, business and industry.
- In almost all fields of human endeavor.
- Almost all human beings in their daily life are subjected to obtaining numerical facts.
- Applicable in some process e.g. invention of certain drugs, extent of environmental pollution.
- In industries especially in quality control area.

II. Uses of Statistics

- Statistics presents fact in the form of numerical data
- It condenses and summarizes a mass of data in to a few presentable and precise figures.
- It facilitates comparison of data
- It helps in formulating and testing hypothesis
- It helps in predicting future trend
- It helps in formulating polices.

III. Limitations of Statistics

Statistics with all its wide application in every sphere of human activity has its own limitation. Some of them are given below

- **Statistics is not suitable to the study of qualitative phenomenon:** Since statistics is basically a science and deals with a set of numerical data, it is applicable to the study of only these subjects of enquiry, which can be expressed in terms of quantitative measurements. As a matter of fact, qualitative phenomenon like honesty, poverty, beauty, intelligence etc. …, cannot be expressed numerically and any statistical analysis cannot be directly applied on this qualitative phenomenon. Nevertheless, statistical techniques may be applied indirectly by first reducing the qualitative expressions to accurate quantitative terms. For example, the intelligence of a group of students can be studied on the basis of their marks in a particular examination.
- **Statistics does not study individuals:** Statistics does not give any specific importance to the individual items; in fact it deals with an aggregate of objects. Individual items, when they are taken individually do not constitute any statistical data and do not serve any purpose for any statistical enquiry.
- **Statistical laws are not exact:** It is well known that mathematical and physical sciences are exact. But statistical laws are not exact and statistical laws are only approximations. Statistical conclusions are not universally true. They are true only on an average.
- **Statistics table may be misused:** Statistics must be used only by experts; otherwise, statistical methods are the most dangerous tools on the hands of the inexpert. The use of statistical tools by the inexperienced and untraced persons might lead to wrong conclusions. Statistics can be easily misused by quoting wrong figures of data. As King says aptly ‘statistics are like clay of which one can make a God or Devil as one pleases.’
- **Statistics is one of the methods of studying a problem:** Statistical method does not provide complete solution of the problems because problems are to be studied taking the background of the countries culture, philosophy or religion into consideration. Thus the statistical study should be supplemented by other evidences.

1.5. Scales of 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*. Scales of measurement refer to ways in which variables or numbers are defined and categorized. Each scale of measurement has certain properties which in turn determine the appropriateness for use of certain statistical analyses. The four scales of measurement are nominal, ordinal, interval, and ratio.

Nominal Scales

Nominal scales possess the following properties.

✓ Level of measurement which classifies data into mutually exclusive, all-inclusive categories in which no order or ranking can be imposed on the data.

- ✓ No arithmetic and relational operation can be applied.
- ✓ No quantitative information is conveyed
- ✓ Thus only gives names or labels to various categories.

Examples:

- ✓ Political party preference (Republican, Democrat, or Other,)
- ✓ Sex (Male or Female.)
- ✓ Marital status (married, single, widow, divorce)
- ✓ Country code
- ✓ Regional differentiation of Ethiopia.

2. Ordinal Scales

Ordinal Scales are measurement systems that possess the following properties:

- ✓ Level of measurement which classifies data into categories that can be ranked, however differences between the ranks do not exist.
- ✓ Arithmetic operations are not applicable but relational operations are applicable.
- ✓ Ordering is the sole property of ordinal scale.

Examples:

- ✓ Letter grades (A, B, C, D, F).
- ✓ Rating scales (Excellent, Very good, Good, Fair, poor).
- ✓ Military status.

3. Interval Scales

Interval scales are measurement systems that possess the following properties:

- ✓ Level of measurement which classifies data that can be ranked and differences are meaningful. However, there is no meaningful zero, so ratios are meaningless.
- ✓ All arithmetic operations except division are applicable.
- ✓ Relational operations are also possible.

Examples:

- ✓ IQ, Temperature in F0.

4. Ratio Scales

Ratio scales measurement possesses the following properties: Level of measurement which classifies data that can be ranked, differences are meaningful, and there is a true zero. True ratios exist between the different units of measure.

- ✓ All arithmetic and relational operations are applicable.

Examples:

- ✓ Weight, Height, Number of students, Age

Use of level of measurements

- ✓ Helps you decide how to interpret the data from the variable.
- ✓ Helps you decide what statistical analysis is appropriate on the values that were assigned. For example if a measurement is nominal then you know that you never average the data level.

Assignment: What is the application of statistics for your field of study?

Chapter Two: Methods of Data Collection and Presentation (4 lecture hours)

2.1	Methods of data collection
2.1.1	Sources of data
2.1.2	Methods of collection
2.2	Methods of Data Presentation
2.2.1	Motivating examples
2.2.2	Frequency distributions: qualitative, quantitative
2.2.3	Tabular presentation of data
2.2.4	Diagrammatic display of data: Bar charts, Pie-chart, Cartograms
2.2.5	Graphical presentation of data: Histogram, Frequency Polygon, Ogive Curves

2.1. Methods of data collection

Once it is decide what type of study is to be made, it becomes necessary to collected information about the concerned study, mostly in the form of data. In order to generate valid conclusion from a data, information has to be collected in a systematic manner. Whatever the quality of sampling and analysis method, a haphazardly collected dataset is less likely to produce valuable and generalizable information.

2.1.1 Sources of Data:

There are two types (sources) for the collection of data.

(1) Primary Data **(2)** Secondary Data

(1). Primary data

- The primary data are the first hand information collected, compiled and published by organization for some purpose. They are most original data in character and have not undergone any sort of statistical treatment.
- Refer to those that are collected by conducting survey to meet the specific problem needs at hand.

Example: Population census reports are primary data because these are collected, complied and published by the population census organization.

(2). Secondary data -

- The secondary data are the second hand information which are already collected by someone (organization) for some purpose and are available for the present study. The secondary data are not pure in character and have undergone some treatment at least once.
- Data taken from already available published or unpublished source.

Exercise-2: Write the merits and demerits of secondary data.

2.1.2 Methods of collection

Data collection is the first task to be carried out in statistical analysis. There are two methods of data collection, namely primary and secondary methods of data collection.

The primary method consists of obtaining data or information by any of the following ways:

1. self-administered questionnaire
2. direct investigation-measurement (observation) of the subject and interviewing(face-to-face, telephone, ---)

1. Self-administered questionnaire

Questionnaire is the main data collection instrument in formal sample survey. Before examining the steps in designing a questionnaire we need to review the types of questions used in questionnaires. Depending on the amount of freedom given to respondent in offering responses, there are two basic types of questions that can be used in questionnaires: open-ended questions and closed ended questions.

The type of questions for use will be determined by the form of responses wanted, the nature of the respondents and their ability to answer the questions.

Open-ended questions: - allows the respondent to answer it freely in his or her own words
Example: what do you think are the reasons for a high drop-out rate of village health committee members?

Closed – ended questions:-

Predetermined list of alternate responses is presented to the respondent for checking the appropriate one(s). It implies that the respondent's answers are restricted in some way to a limited range of alternatives.

Advantage

- It is the cheapest and can be conducted by a single researcher.
- Questionnaires can be sending to a wide geographical area.
- There is no interviewer variability

Disadvantage

- Low response rate
- No assurance that the questioners were answered by the right person.
- Mail questionnaire is not suitable for illiterate community

2. direct investigation

I. measurement or/and observation

- data can be obtained through direct observation or measurement
- provides accurate information but it is expensive and inconvenient

eg: Land area measurement, Animal weight gain, Physical examination, direct observation of work.

II. Interview

a) Face-to-Face interview

Advantage:-

- Interviewers can observe the surroundings and can use nonverbal communication and visual aids.
- The interviewer can help the respondent if he/she has difficulty in understanding the questions.
- Respondent is likely to answer all the questions alone

Disadvantage:-

- Cost is high
- Interviewer bias is also high
- Untrained interviewer may distort the meaning of the questions

b) Telephone Interview

Advantage:-

- It is less expensive in time and money compared to face to face interviews
- Relatively high response rate
- Reach people who would not open their doors to an interviewer, but might willing to talk on the telephone

Disadvantage:-

- Unrepresentative of the groups which do not have telephones
- Unlisted telephone numbers are excluded from the study.
- Respondent may be substitute by another

The data obtained through these methods are called primary data, as explained above.

The other method of data collection, secondary method, is a method by which we collect data from secondary sources such as administrative records, books, survey results and so on.

- Extracting information from existing resources.
- Is much less expensive than any other two sources
- It is difficult to get the information needed when records are compiled in unstandardized manner.

Example: - Hospital records, professional institutes, Official statistics, - - -

Data collected through such methods are known as secondary data, as has been discussed above.

Editing of Data:

After collecting the data either from primary or secondary source, the next step is its editing. Editing means the examination of collected data to discover any error and mistake before presenting it. It has to be decided before hand what degree of accuracy is wanted and what extent of errors can be tolerated in the inquiry. The editing of secondary data is simpler than that of primary data.

Merits and Demerits of primary data:

1. The collection of data by the method of personal survey is possible only if the area covered by the investigator is small. Collection of data by sending the enumerator is bound to be expensive. Care should be taken twice that the enumerator record correct information provided by the informants.
2. Collection of primary data by framing a schedules or distributing and collecting questionnaires by post is less expensive and can be completed in shorter time.
3. Suppose the questions are embarrassing or of complicated nature or the questions probe into personnel affairs of individuals, then the schedules may not be filled with accurate and correct information and hence this method is unsuitable.
4. The information collected for primary data is more reliable than those collected from the secondary data.

Merits and Demerits of Secondary Data:

1. Secondary data is cheap to obtain. Many government publications are relatively cheap and libraries stock quantities of secondary data produced by the government, by companies and other organizations.
2. Large quantities of secondary data can be got through internet.
3. Much of the secondary data available has been collected for many years and therefore it can be used to plot trends.
4. Secondary data is of value to:
   - The government – help in making decisions and planning future policy.
   - Business and industry – in areas such as marketing, and sales in order to appreciate the general economic and social conditions and to provide information on competitors.
   - Research organizations – by providing social, economic and industrial information.

2.2. Methods of Data organization and Presentation

This topic introduces tabular and graphical 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. Classification of Data

The process of arranging data into homogenous group or classes according to some common characteristics present in the data is called classification.

For Example: The process of sorting letters in a post office, the letters are classified according to the regions and further arranged according to zones, cities, etc.

Bases of Classification:

There are four important bases of classification:

(1) Qualitative Base **(2)** Quantitative Base **(3)** Geographical Base **(4)** Chronological or Temporal Base

(1) Qualitative Base:

When the data are classified according to some quality or attributes such as sex, religion, literacy, intelligence etc.…

(2) Quantitative Base:
When the data are classified by quantitative characteristics like heights, weights, ages, income etc.…
(3) Geographical Base:
When the data are classified by geographical regions or location, like states, provinces, cities, countries etc.…
(4) Chronological or Temporal Base:
When the data are classified or arranged by their time of occurrence, such as years, months, weeks, days etc.… **For Example:** Time series data.
Tabulation of Data
The process of placing classified data into tabular form is known as tabulation. A table is a symmetric arrangement of statistical data in rows and columns. Rows are horizontal arrangements whereas columns are vertical arrangements.
2.2.2. Organization of data: Frequency distribution
A frequency distribution is the organization of raw data in table form, using classes and frequencies. There are three basic types of frequency distributions, and there are specific procedures for constructing each type. The three types are categorical, ungrouped and grouped frequency distributions.
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 charts and graphs for the presentation of data
- To enable the reader to make comparisons between different data set

Some of basic terms that are most frequently used while we deal with 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 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 to the upper class limit and dividing the sum by 2.
- Class width is the difference between two consecutive lower class limits or two consecutive lower class boundaries.

2.2.2.1. 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 data such as political affiliation, religious affiliation, or major field of study would use categorical frequency distribution.
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 graph.
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 is frequency and n is total number of values

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

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

Construct a frequency distribution for the data.

Solution:

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

2.2.2.2 Ungrouped Frequency Distribution

When the data are numerical interested of categorical, the range of data is small and each class is only one unit, this distribution is called an ungrouped frequency distribution.

The major components of this type of frequency distributions are class, tally, 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.

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		

A. Construct ungrouped frequency distribution

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

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

Solution:

A. Since this data set contains only a relatively small number 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

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

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

2.2.2.3. 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. 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.

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

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

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

5. Select the starting point as the lowest class limit. This is usually the lowest score (observation). Add the width to that score to get the lower class limit of the next class. Keep adding until you achieve the number of desired class(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 limit of the first class. Then add the width to each upper class limit to get all upper class limits.

Unit of measurement: Is the next expected upcoming value. For instance, 28, 23, 52, and then the unit of measurement is one. Because take one datum arbitrarily, say 23, then the next upcoming value will be 24. Therefore,$U = 24 - 23 = 1$. If the data 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 **C**lass **B**oundary $=$ **L**ower **C**lass **L**imit $- \frac{U}{2}$

and **U**pper **C**lass **B**oundary $=$ **U**pper **C**lass **L**imit $- \frac{U}{2}$. 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 type of cumulative frequency namely less than cumulative frequency and more than cumulative frequency. Less than cumulative frequency is obtained by adding successively the frequencies of all the previous classes including the class against which it is written. The cumulate is started from the lowest to the highest size. More than cumulative frequency is obtained by finding the cumulate total of frequencies starting from the highest to the lowest class.

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

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

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

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

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

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

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 limit.

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

7. Find the class boundaries. Take the formula in step 6.
   $LCB_1 = LCL_1 - 0.5$ and $UCB_1 = UCL_1 + 0.5$

Class Boundaries	*5.5-11.5*	*11.5-17.5*	*17.5-23.5*	*23.5-29.5*	*29.5-35.5*	*35.5-41.5*

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

2.2.2.4. 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.

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

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

2.2.3. 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

2.2.3.1. Diagrammatic display of data: Bar charts, Pie-chart, Cartograms

I. 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 is360^o. The angles of each component are calculated by the formula.

$$\text{Angle of Sector} = \frac{\text{Component Part}}{\text{Total}} \text{x} 360^o$$

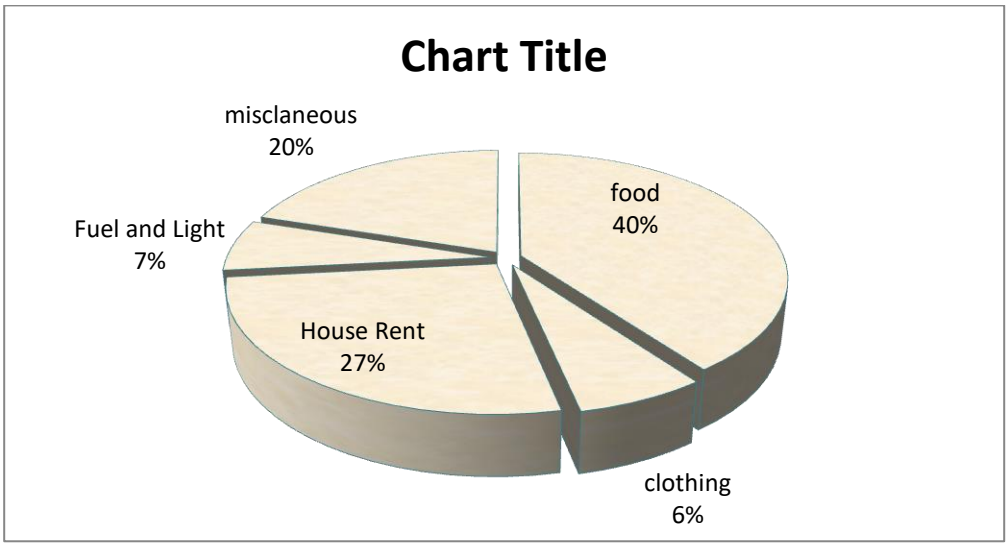
These angles are made in the circle by mean of a protractor to show different components. The arrangement of the sectors is usually anti-clock wise.

Example2.4: The following table gives the details of monthly budget of a family. Represent these figures by a suitable diagram.

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

Solution: The necessary computations are given below:

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



II. Bar Charts

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

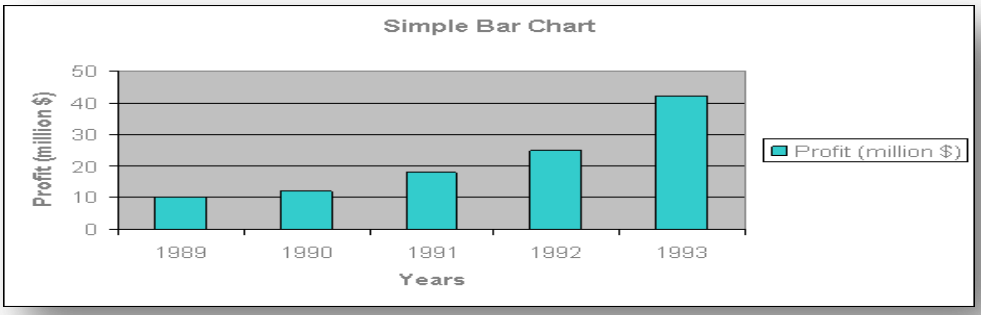
- *Make the bars the same width*
- *Make the units on the axis that are used for the frequency equal in size*

A. *A **simple bar chart** is used to represents data involving only one variable classified on spatial, quantitative or temporal basis.* In simple bar chart, we make bars of equal width but variable length, i.e. the magnitude of a quantity is represented by the height or length of the bars. 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 2.5: Draw simple bar diagram to represent the profits of a bank for 5 years.

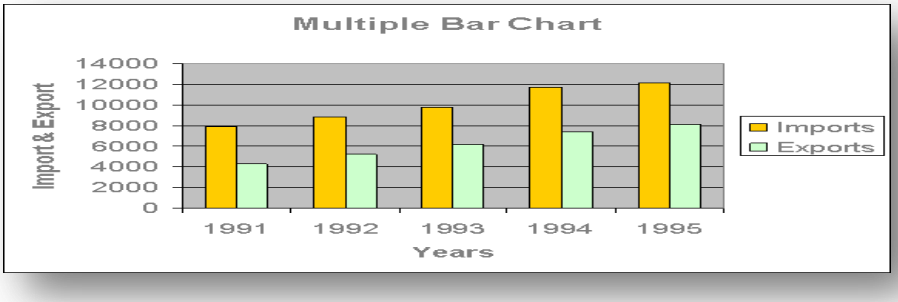
Years	1989	1990	1991	1992	1993
Profit (million $)	10	12	18	25	42



B. **Multiple bar charts** are used two or more sets of inter-related data are represented (multiple bar diagram facilities comparison between more than one phenomenon). The technique of simple bar chart is used to draw this diagram but the difference is that we use different shades, colors, or dots to distinguish between different phenomena.

Example 2.6: Draw a multiple bar chart to represent the import and export of Canada (values in $) for the years 1991 to 1995.

Years	Imports	Exports
1991	7930	4260
1992	8850	5225
1993	9780	6150
1994	11720	7340
1995	12150	8145



C. **Stratified (Stacked or component) 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 or staked chart.

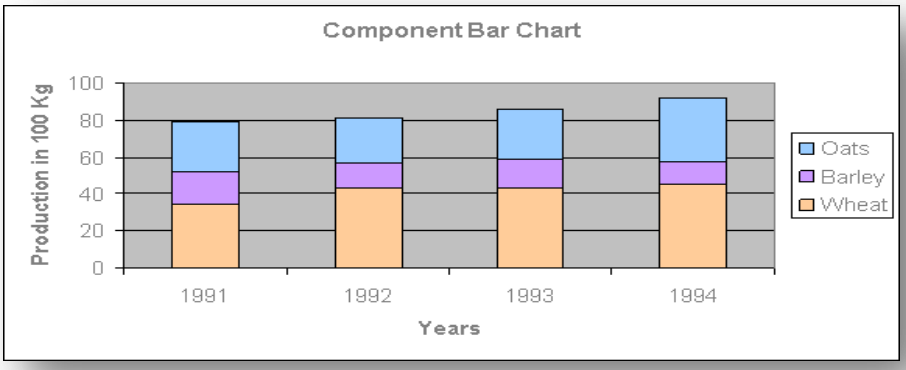
Example 2.7: The table below shows the quantity in hundred kgs of Wheat, Barley and Oats produced on a certain form during the years 1991 to 1994. Draw stratified bar chart.

Years	Wheat	Barley	Oats
1991	34	18	27
1992	43	14	24
1993	43	16	27
1994	45	13	34

Solution: To make the component bar chart, first of all we have to take year wise total production.

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

The required diagram is given below:



2.2.3.2. Graphical presentation of data: Histogram, Frequency Polygon, Ogive Curves

Statistical graphs can be used to describe the data set or to analyze it. Graphs are also useful in getting the audience's attention in a publication or a speaking presentation.

They can be used to discuss an issue, reinforce a critical point, or summarize a data set. They can also be used to discover a trend or pattern in a situation over a period of time.

The three most commonly used graphs in research are

1. The histogram.

2. The frequency polygon.

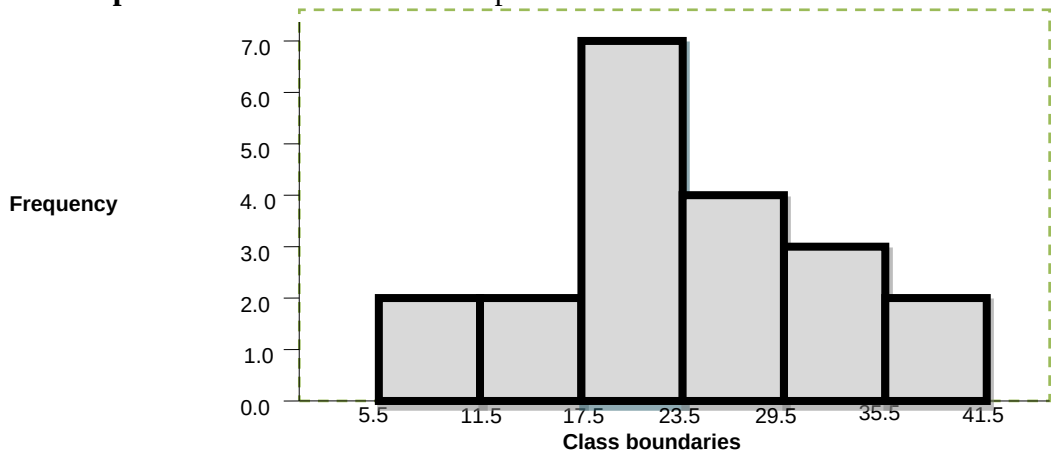
3. The cumulative frequency graph, or Ogive (pronounced o-jive)

(1). 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.

Example2.8: Take the data in example 2.3.

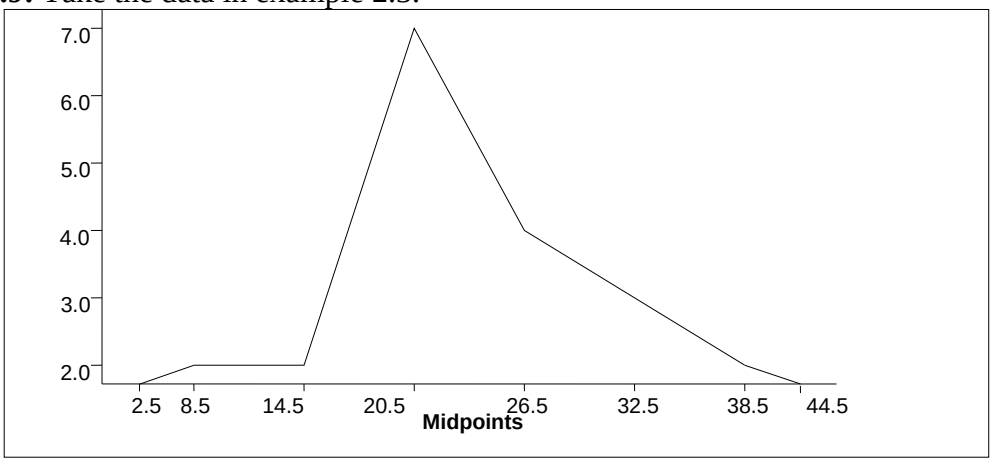


Relative frequency histogram has the same shape and horizontal (y axis) scale as a histogram, but the vertical (x axis) scale is marked with relative frequencies instead of actual frequencies.

(2). 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.

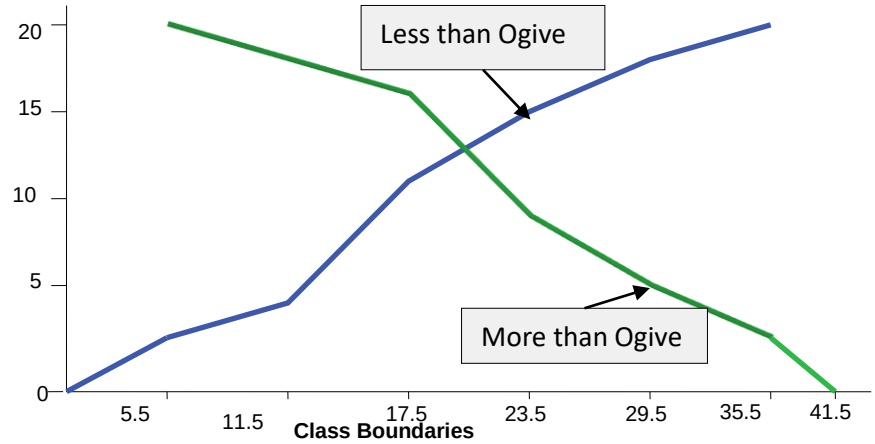
Example 2.9: Take the data in example 2.3.



(3). Ogive Graph

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

Example 2.10: Take the data in example 2.3 and draw less than and more than Ogive



3.1 Motivating examples
3.2 Objectives of measures of central tendency
3.3 Important characteristics of a good average
3.4 Summation notation
3.5 Mean (Arithmetic Mean)
3.6 Median
3.7 Mode

3.1 Introduction and objectives of measure of central tendency

Measures of central tendency are measures of the location of the middle or the center of a distribution. The definition of "middle" or "center" is purposely left somewhat vague so that the term "central tendency" can refer to a wide variety of measures.

❖ The tendency statistical data to get concentrated at certain value is called central tendency. And various methods that determine the actual value at which the data tend to concentrate are called measure of central tendency. One of the most important objectives of statistical analysis is to get one single value that describes the characteristics of the entire data. Such a value is called the central value or average.

❖ When we want to make comparison between groups of numbers it is good to have a single value that is considered to be a good representative of each group. This single value is called the average of the group.

❖ Averages are also called measures of central tendency.

❖ An average which is representative is called typical average and an average which is not representative and has only a theoretical value is called a descriptive average.

3.2 Objectives:

- To comprehend the data easily - to condensed the mass of data in to one single value.
- To facilitate comparison.
- To make further statistical analysis.

3.3 Characteristic of a good measure of central tendency (A typical average should possess the following):

- It should be defined rigidly which means that it should have a definite value.
- It should be based on all observation under investigation.
- It should be not be affected by extreme observations.
- It should be capable of further algebraic treatment.
- It should be as little as affected by fluctuations of sampling or should be stable with sampling.
- It should be ease to calculate and simple to understand.
- It should be unique and always exist.

Note:-There is no measure satisfied all the above condition!

3.4 The Summation Notation:

- Let X1, X2, ,X3 …XN be a number of measurements where N is the total number of observation and Xi is ith observation.
- Very often in statistics an algebraic expression of the form $X_1+X_2+X_3+...+X_N$ is used in a formula to compute a statistic. It is tedious to write an expression like this very often, so mathematicians have developed a shorthand notation to represent a sum of scores, called the summation notation.
- The symbol $\sum_{i=1}^{N} X_i$ is mathematical shorthand for $X1_1+X_2+X_3+...+X_N$

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

The expression is read, "the sum of X sub i from i equals 1 to N." It means "add up all the numbers."

Example: Suppose the following were scores made on the first homework assignment for five students in the class: 5, 7, 7, 6, and 8. In this example set of five numbers, where N=5, the summation could be written:

$$\sum_{i=1}^{5} X_i = X_1 + X_2 + X_3 + X_4 + X_5 = 5 + 7 + 7 + 6 + 8 = 33$$

The "i=1" in the bottom of the summation notation tells where to begin the sequence of summation. If the expression were written with "i=2", the summation would start with the second number in the set.

For example: $\sum_{i=2}^{5} X_i = X_2 + X_3 + X_4 + X_5 = 7 + 7 + 6 + 8 = 28$

The "N" in the upper part of the summation notation tells where to end the sequence of summation. If there were only three scores then the summation and example would be:

$$\sum_{i=1}^{3} X_i = X_1 + X_2 + X_3 = 5 + 7 + 7 = 19$$

Sometimes if the summation notation is used in an expression and the expression must be written a number of times, as in a proof, then a shorthand notation for the shorthand notation is employed. When the summation sign " $\sum$ " is used without additional notation, then "i=1" and "N" are assumed

PROPERTIES OF SUMMATION

1. $\sum_{i=1}^{n} K = nK$, Where k is any constant
2. $\sum_{i=1}^{n} KX_i = K\sum_{i=1}^{n} X_i$, Where k is any constant
3. $\sum_{i}^{n}(a + bX_i) = na + b\sum_{i=1}^{n} X_i$, where a and b are any constant
4. $\sum_{i=1}^{n}(X_i + Y_i) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i$
5. $\sum_{i=1}^{N}(X_i \times Y_i) = X_1 \times Y_1 + X_2 \times Y_2 + --- + X_N \times Y_N$

Example 3.1: considering the following data determine

X	Y
5	6
7	7
7	8
6	7
8	8

a) $\sum_{i=1}^{5} X_i$ b) $\sum_{i=1}^{5} Y_i$ c) $\sum_{i=1}^{5} 11$ d) $\sum_{i=1}^{5}(X_i + Y_i)$ e) $\sum_{i=1}^{5}(X_i - Y_i)$

f) $\sum_{i=1}^{5} X^{2}{}_i$ g) $\sum_{i=1}^{5} X_i Y_i$ h) $\sum_{i=1}^{5} X_i + \sum_{i=1}^{5} Y_i$ g) $\sum_{i=1}^{5} X_i \sum_{i=1}^{5} Y_i$

Types of measures of central tendency

There are several different measures of central tendency; each has its own advantage and disadvantage.

- The Mean
- The Mode
- The Median

The choice of these averages depends up on which best fit the property under discussion.

3.5. The Mean

There are three types of mean which are suitable for a particular type of data. They are

a) Arithmetic mean or Average
b) Geometric mean
c) Harmonic mean

The Arithmetic Mean:

Divided in to two i.e. simple arithmetic mean and the weighted arithmetic mean

1) Simple Arithmetic Mean:

The mean is defined as the sum of the magnitude of the items divided by the number of items

- The mean of $X_1, X_2, X_3 \ldots X_n$ is denoted by A.M, or $\overline{X}$ and is given by:

$$\overline{X} = \frac{X_1 + X_2 + - - - + X_n}{n} = \sum_{i=1}^{n} X_i / n$$

- When the data are arranged or given in the form of frequency distribution i.e. there are k variate values such that a value X_i has a frequency f_i (i=1,2,---,k) ,then the Arithmetic mean will be

$\overline{X} = \frac{\sum_{i=1}^{k} f_i X_i}{\sum_{i=1}^{k} f_i}$ Where k is the number of classes and $\sum_{i=1}^{k} f_i = n$

Arithmetic Mean for Grouped Data

If data are given in the shape of a continuous frequency distribution, then the arithmetic mean is obtained as follows:

$\overline{X} = \frac{\sum_{i=1}^{n} f_i Y_i}{\sum_{i=1}^{n} f_i}$ Where Y_i = the class mark of the i^{th} class and fi = the frequency of the i^{th} class

Example 3.2:

1) Daily cash earnings of 15 workers working in different industries are as follows: 11.63,8.22,12.56,12.14,29.23,18.23,11.49,11.30,17.00,9.16,8.64,27.56,8.23,19.77,12.81.Find the average daily earning of a worker?

2) The distribution of age at first marriage of 130 males was as given below

Age in years(X):	18	19	20	21	22	23	24	25	26	27	28	29
No. of males (f):	2	1	4	8	10	12	17	19	18	14	13	12

Compute the average age of males at first marriage?

3) Calculate the mean for the following age distribution.

Class	frequency
6- 10	35
11- 15	23
16- 20	15
21- 25	12
26- 30	9
31- 35	6

Special properties of Arithmetic mean

1.The sum of the deviations of a set of items from their mean is always zero. i.e.

$$\sum_{i=1}^{n}(X_i - \overline{X}) = 0$$

2.The sum of the squared deviations of a set of items from their mean is the minimum. i.e.

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 < \sum_{i=1}^{n}(X_i - A)^2$$

3.If $\overline{X}_1$ is the mean of observations n_1

If $\overline{X}_2$ is the mean of observations n_2

.

.

.

If $\overline{X}_k$ is the mean of observations n_k

Then the mean of all the observation in all groups often called the combined mean is given by:

$$\overline{X}_c = \frac{\sum_{i=1}^{n} n_i \overline{X}_i}{\sum_{i=1}^{n} n_i}$$

4. If a wrong figure has been used when calculating the mean the correct mean can be obtained without repeating the whole process using:

$$Corrected\ mean = Wrong\ mean + \frac{(correct\ value - Wrong\ value)}{n}$$

Where n is total number of observations.

5.The effect of transforming original series on the mean.

a) If a constant *k* is added/ subtracted to/from every observation then the new mean will be *the old mean*± *k* respectively i.e. $\overline{X}_{new} = \overline{X}_{old} \pm K$

b) If every observations are multiplied by a constant *k* then the new mean will be k**old mean i.e.* $\overline{X}_{new} = \overline{X}_{old} \times K$.

Example 3.4:

1) In a class there are 30 females and 70 males. If females averaged 60 in an examination and boys averaged 72, find the mean for the entire class?

2) An average weight of 10 students was calculated to be 65.Latter it was discovered that one weight was misread as 40 instead of 80 k.g. calculate the correct average weight?

3) The mean of n Tetracycline Capsules $X_1, X_2, \ldots, X_n$ are known to be 12 gm. New set of capsules of another drug are obtained by the linear transformation $Y_i = 2X_i - 0.5$ (i = 1, 2, …, n) then what will be the mean of the new set of capsules?

4) The mean of a set of numbers is 500.

a.If 10 is added to each of the numbers in the set, then what will be the mean of the new set?

b.If each of the numbers in the set are multiplied by -5, then what will be the mean of the new set?

2) Weighted Mean

- When a proper importance is desired to be given to different data a weighted mean is appropriate.
- Weights are assigned to each item in proportion to its relative importance.

- Let $X_1, X_2, \ldots X_n$ be the value of items of a series and $W_1, W_2, \ldots W_n$ their corresponding weights , then the weighted mean denoted $\overline{X}_w$ is defined as:

$$\overline{X}_w = \frac{\sum_{i=1}^{n} X_i W_i}{\sum_{i=1}^{n} W_i}$$

Example 3.5:

1.Example: A student obtained the following marks in his examinations: English 60, Biology 75, Physics 59 and Chemistry 55. find the students weighted mean if weights 1, 2, 1, 3 and 3 respectively allotted to the subjects.

Solution:

$$\overline{X}_w = \frac{\sum_{i=1}^{n} X_i w_i}{\sum_{i=1}^{n} w_i} = \frac{60\times1+75\times2+63\times1+59\times3+55\times3}{1+2+1+3+3} = \frac{615}{10} = 61.5$$

2.A teacher allots weights 2 to homework, 3 to mid-exam and 5 to final exam. If students score 90, 50, and 60 for HW, MID and FIN respectively. What is his/her academic performance?

Merits and Demerits of Arithmetic Mean

Merits:

- It is rigidly defined.
- It is based on all observation.
- It is suitable for further mathematical treatment.
- It is stable average, i.e. it is not affected by fluctuations of sampling to some extent.
- It is easy to calculate and simple to understand.

Demerits:

- It is affected by extreme observations.
- It cannot be used in the case of open end classes.
- It cannot be determined by the method of inspection.
- It cannot be used when dealing with qualitative characteristics, such as intelligence, honesty, beauty.
- It can be a number which does not exist in a serious.
- Sometimes it leads to wrong conclusion if the details of the data from which it is obtained are not available.
- It gives high weight to high extreme values and less weight to low extreme values.

3.6. The Median

- In a distribution, median is the value of the variable which divides it in to two equal halves.
- In an ordered series of data median is an observation lying exactly in the middle of the series. It is the middle most value in the sense that the number of values less than the median is equal to the number of values greater than it.

-If $X_1, X_2, \ldots X_n$ be the observations, then the numbers arranged in ascending order will be $X_{[1]}, X_{[2]}, \ldots X_{[n]}$, where $X_{[i]}$ is i^{th} smallest value.

➢ $X_{[1]} < X_{[2]} < \ldots < X_{[n]}$

-Median is denoted by. $\tilde{X}$

Median for ungrouped data:

$$X_{\frac{n+1}{2}}, \quad if\ n\ is\ odd$$

$$\frac{1}{2}(X_{\frac{n}{2}} + X_{\frac{n}{2}+1}), if\ n\ is\ even$$

Example 3.8: 1) Actual waiting time for the first job on the selected sample of nine people having different field of specializations was given below.

Waiting time (in month):11.6, 11.3, 10.7, 18.0, 3.3, 9.2, 8.3, 3.8, 6.8

Calculate the median of the waiting time?

2) The export of agricultural products in million dollars from a country during eight quarters in 1974 and 1975 was, 29.7, 16.6, 2.3, 14.1, 36.6, 18.7, 3.5, 21.3.

Find the median of the given set of values?

Median for grouped data.

-If data are given in the shape of continuous frequency distribution, the median is defined as:

$$\tilde{X} = L_{med} + \frac{W}{f_{med}}(\frac{n}{2} - f_c)$$

Where: L_{med} =lower class boundary of the median class.

f_{med}= *The frequency of the median class*

f_c= *The comulative frequency less than type preceding the median class*.

W=the size of the median class.

n=total number of observation.

Note: The median class is the class with the smallest cumulative frequency (less than type) greater than or equal to n/2.

Example 3.9: Find the median of the following distribution.

Class	Frequency
40-44	7
45-49	10
50-54	22
55-59	15
60-64	12
65-69	6
70-74	3

Merits and Demerits of Median

Merits:

• Median is a positional average and hence not influenced by extreme observations.

• Can be calculated in the case of open end intervals.

• Median can be located even if the data are incomplete.

Demerits:

• It is not a good representative of data if the number of items is small.

• It is not amenable to further algebraic treatment.

• It is susceptible to sampling fluctuations.

3.7. The Mode

- Mode is a value which occurs most frequently in a set of values
- The mode may not exist and even if it does exist, it may not be unique.
- In case of discrete distribution the value having the maximum frequency is the model value.
- If in a set of observed values, all values occur once or equal number of times, there is no mode

Examples:

1.Find the mode of 5, 3, 5, 8, and 9

Mode =5

2.Find the mode of 8, 9, 9, 7, 8, 2, and 5.

It is a bimodal Data: 8 and 9

3. Find the mode of 4, 12, 3, 6, and 7.

No mode for this data.

- The mode of a set of numbers $X_1, X_2, \ldots X_n$ is usually denoted by $\hat{X}$.

Mode for Grouped data

If data are given in the shape of continuous frequency distribution, the mode is defined as:

$$\hat{X} = L_{\text{mod}} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right)W$$

Where: $\hat{x}$ = the mode of the distribution, L_{mo}= the lower class boundary of the modal class

$$\Delta_1 = f_{mo} - f_1$$

$$\Delta_2 = f_{mo} - f_2$$

f_{mo}= frequency of the modal class

f_1 = frequency of the class preceding the modal class

f_2 = frequency of the class succeeding the modal class W=the size of the modal class

Note: The modal class is a class with the highest frequency

Example 3.7: The following is the distribution of the size of certain farms selected at random from a district. Calculate the mode of the distribution.

Size of farms	5- 15	15- 25	25- 35	35- 45	45- 55	55- 65	65- 75
No. of farms	8	12	17	29	31	5	3

Merits and Demerits of Mode

Merits:

- It is not affected by extreme observations.
- Easy to calculate and simple to understand.
- It can be calculated for distribution with open end class.
- Can be used for qualitative data as well.

Demerits:

- It is not rigidly defined.
- It is not based on all observations
- It is not suitable for further mathematical treatment.
- It is not stable average, i.e. it is affected by fluctuations of sampling to some extent.
- Often its value is not unique.

Empirical relationship between $\bar{X}, \hat{X}, and\ \tilde{X}$

- $\bar{X} = \hat{X} = \tilde{X}$, for symmetrical distribution
- $\bar{X} - \hat{X} = 3\left(\bar{X} - \tilde{X}\right)$, for unimodal skewed or asymmetrical frequency distribution.

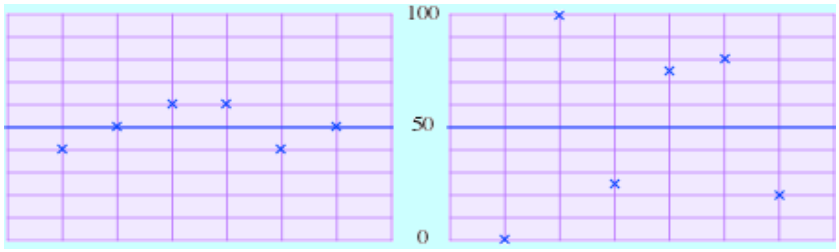
4.1 Motivating examples
4.2 Objectives of measures of variation
4.3 Measures of Dispersion (Variation)
4.3.1 Range and Relative Range
4.3.2 Variance, Standard Deviation and Coefficient of Variation
4.3.3 Standard Scores
4.4 Moments (about the origin and about the mean)
4.5 Skewness
4.6 Kurtosis

4.1 Introduction

Consider the following two sets of scores:

Set 1: 40, 50, 60, 60, 40, 50
Set 2: 0, 100, 25, 75, 80, 20

Both these sets have the same mean (50), but the second set is a lot more widely dispersed ("scattered") than the first.



Set 1 **Set 2**

Measure of central tendency alone does not adequately describe a set of observation unless all observations are the same. So we need some additional information like

1) The extent to which the items in a particular distribution are scatters around the central tendency i.e. measure of dispersion.
2) The direction of scattered ness whether more items are attached towards higher or lower values i.e. measure of skew ness.
3) The extent to which the distribution is more peaked or more flat toped than the normal distribution i.e. measure of kurtosis.

Definition:

- The scatter or spread of items of a distribution is known as dispersion or variation. In other words the degree to which numerical data tend to spread about an average value is called dispersion or variation of the data.
- Measures of dispersions are statistical measures which provide ways of measuring the extent in which data are dispersed or spread out.

4.2 Objectives of measuring variation

- To determine the reliability of an average by pointing out as how far an average is representative of the entire data.
- To determine the nature and cause of variation in order to control the variation itself.
- Enable comparison of two or more distribution with regard to their variability.
- Measuring variability is of great importance to other statistical analysis. E.g., it is the basis of statistical quality control.

A good measure of variation possess

- It should be easy to compute and understand.
- It should be based on all observations.
- It should be Uniquely defined
- It should be capable of further algebraic treatment.
- It should be as little as affected by extreme values

Absolute and relative measures

Measures of dispersion may be either absolute or relative

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

2. ***Relative measures of dispersion (RMD)***: when one desires to compare the dispersion in two sets of data, however comparing the two AMDs may lead to fallacious results. It may be that the two variables involved are measured in different units. For example, we may wish to know, for a certain population, whether serum cholesterol levels, measured in milligrams per 100 ml, are more variable than body weight, measure in kilograms.

Furthermore, although the same unit of measurement is used, the two MCT (means) may be quite different. If we compare the AMD of weights of first grade children with the AMD of the weights of high school freshmen, we may find that the latter AMD is numerically larger than the former, because the weights themselves are larger, not because the AMD is larger.

What is needed in situation like these is a measure of relative variation rather than absolute variation. It is the ratios of absolute dispersion to an appropriate average such as co- efficient of Standard Deviation or Co-efficient of Mean Deviation.

4.3 Types of Measures of Dispersion

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

Absolute measure	**Relative measures**
Range	Relative range
Variance	Coefficient of variation
Standard deviation	Standard scores

4.3.1 The Range (R)

The range is the largest score minus the smallest score. It is a quick and dirty measure of variability, although when a test is given back to students they very often wish to know the range of scores. Because the range is greatly affected by extreme scores, it may give a distorted picture of the scores. The following two distributions have the same range, 13, yet appear to differ greatly in the amount of variability.

Distribution 1:	32	35	36	36	37	38	40	42	42	43	43	45
Distribution 2:	32	32	33	33	33	34	34	34	34	34	35	45

For this reason, among others, the range is not the most important measure of variability.

For ungrouped data:

$$R = x_{\max} - x_{\min}, \quad where\ x_{\max} = \max imum\ \ value$$

$$X_{\min} = \min imum value$$

For grouped data:

$$R = x_{\max} - x_{\min}, where\ x_{\max} = UCB\ \ for the last class$$

$$X_{\min} = LCB\ \ for the\ \ first class$$

Relative Range (RR)

It is also sometimes called coefficient of range and given by:

For ungrouped data: $RR = \frac{X_{\max} - X_{\min}}{X_{\max} + X_{\min}}$

For grouped data: $RR = \frac{UCB_{last} - LCB_{first}}{UCB_{last} + LCB_{first}}$

Merits and Demerits of range

Merits:

- It is rigidly defined.
- It is easy to calculate and simple to understand.

Demerits:

- It is not based on all observation.
- It is highly affected by extreme observations.
- It is affected by fluctuation in sampling.
- It is not liable to further algebraic treatment.
- It cannot be computed in the case of open end distribution.
- It is very sensitive to the size of the sample.

Example 1

For raw data, 5, 6,8,4,5, 3,9,8,7,3,5,6,8,11

R=11-3=8

$$coefficient\ of\ range = \frac{11-3}{11+3} = \frac{8}{14} = 0.57$$

Example 2:

Height (in)	Number of Students
Less than 59.5	0
Less than 62.5	5
Less than 65.5	23
Less than 68.5	65
Less than 71.5	92
Less than 74.5	100

R=74.5-56.5=18

$$coefficient\ range = \frac{x_{\max} - x_{\min}}{x_{\max} - x_{\min}} = \frac{74.5-56.5}{74.5+56.5} = 0.135$$

Example 4.1:1) Find the R, and RR and then identify which data is more dispersed?

a) For the month income of 10 workers X_i: 347, 420, 500,600,696,710, 835, 850, and 900.

b) For the following age distribution.

Class	frequency
6- 10	35
11- 15	23
16- 20	15
21- 25	12
26- 30	9
31- 35	6

2. If the range and relative range of a series are 4 and 0.25 respectively. Then what is the value of:

a) Smallest observation

b) Largest observation

4.3.2 The Variance

Population Variance

If we divide the variation by the number of values in the population, we get something called the population variance. This variance is the "average squared deviation from the mean".

$$Population Variance = \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - u)^2}{N}, \; i = 1,2,3,..., N$$

Sample Variance

One would expect the sample variance to simply be the population variance with the population mean replaced by the sample mean. However, one of the major uses of statistics is to estimate the corresponding parameter. This formula has the problem that the estimated value isn't the same as the parameter. To counteract this, the sum of the squares of the deviations is divided by one less than the sample size.

$$Sample Variance = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

i.e. The **sample variance**, denoted by s^2, of a set of n *o*bserved values having a mean $\bar{x}$ is the sum of the squared deviations divided by $n-1$.

The following steps are used to calculate the sample variance:

1. Find the arithmetic mean.
2. Find the difference between each observation and the mean.
3. Square these differences.
4. Sum the squared differences.
5. Since the data is a sample, divide the number (from step 4 above) by the number of observations minus one, i.e., n-1 (where n is equal to the number of observations in the data set).

For the case of frequency distribution it is expressed as:

$$s^2 = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^2}{n-1}$$

We usually use the following short cut formula.

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}, \quad for\ raw\ data$$

$$s^2 = \frac{\sum_{i=1}^{n} f_i x_i^2 - n\bar{x}^2}{n-1}, \quad for\ frequency\ distribition,\ where \sum f_i = n$$

Standard Deviation

There is a problem with variances. Recall that the deviations were squared. That means that the units were also squared. To get the units back the same as the original data values, the square root must be taken.

$$population\ standard\ deviation = \sigma = \sqrt{\sigma^2}$$

$$sample\ standard\ deviation = s = \sqrt{s^2}$$

Examples: Find the variance and standard deviation of the following sample data

1. 5, 17, 12, 10,8
2. The data is given in the form of frequency distribution.

Class	Frequency
40-44	7
45-49	10
50-54	22
55-59	15
60-64	12
65-69	6
70-74	3

Coefficient of Variation (C.V)

• Is defined as the ratio of standard deviation to the mean usually expressed as percents.

$$C.V = \frac{S}{\bar{X}} * 100\%$$

• The distribution having less C.V is said to be less variable or more consistent.

Examples:

1. An analysis of the monthly wages paid (in Birr) to workers in two firms A and B belonging to the same industry gives the following results

Value	Firm A	Firm B
Mean wage	52.5	47.5
Median wage	50.5	45.5
Variance	100	121

In which firm A or B is there greater variability in individual wages?

Solutions:

Calculate coefficient of variation for both firms.

$$C.V_A = \frac{S_A}{\bar{X}_A} * 100\% = \frac{10}{52.5} * 100\% = 19.05\%$$

$$C.V_B = \frac{S_B}{\bar{X}_B} * 100\% = \frac{11}{47.5} * 100\% = 23.16$$

Since $C.V_A < C.V_B$, in firm B there is greater variability in individual wages.

2. A meteorologist interested in the consistency of temperatures in three cities during a given week collected the following data. The temperatures for the five days of the week in the three cities were

City 1	25	24	23	26	17
City2	22	21	24	22	20
City3	32	27	35	24	28

Which city have the most consistent temperature, based on these data?

(Exercise)

4.3.3 Standard Scores (Z-scores)

- If X is a measurement from a distribution with mean *X* and standard deviation S, then its value in standard units is

$$Z_i = \frac{X_i - u}{\sigma}\ ,\ for\ the\ population$$

$$Z_i = \frac{X_i - \bar{x}}{s}\ ,\ for the sample$$

- Z gives the deviations from the mean in units of standard deviation.
- Z gives the number of standard deviation a particular observation lie above or below the mean.
- It is used to compare two observations coming from different groups.

Examples:

1. Two sections were given introduction to statistics examinations. The following information was given.

Value	Section 1	Section 2
Mean	78	90
Stand. deviation	6	5

Student A from section 1 scored 90 and student B from section 2 scored 95. Relatively speaking who performed better?

Solutions:

Calculate the standard score of both students.

$$Z_1 = \frac{X_1 - \overline{X_1}}{S_1} = \frac{90-78}{6} = 2$$

$$Z_2 = \frac{X_2 - \overline{X_2}}{S_2} = \frac{95-90}{5} = 1$$

$\Rightarrow$ Student A performed better relative to his section because the score of student A is two standard deviation above the mean score of his section while, the score of student B is only one standard deviation above the mean score of his section.

2. Two groups of people were trained to perform a certain task and tested to find out which group is faster to learn the task. For the two groups the following information was given:

Value	Group one	Group two
Mean	10.4 min	11.9 min
Stand.dev.	1.2 min	1.3 min

Relatively speaking:

a) Which group is more consistent in its performance

b) Suppose a person A from group one take 9.2 minutes while person B from Group two take 9.3 minutes, who was faster in performing the task? Why?

Solutions:

a) Use coefficient of variation.

$$C.V_1 = \frac{S_1}{\overline{X_1}} * 100\% = \frac{1.2}{10.4} 100\% = 11.54\%$$

$$C.V_2 = \frac{S_2}{\overline{X_2}} * 100\% = \frac{1.3}{11.9} * 100\% = 10.92\%$$

Since $C.V_2 < C.V_1$, group 2 is more consistent.

b) Calculate the standard score of A and B

$$Z_A = \frac{X_A - \overline{X_1}}{S_1} = \frac{9.2 - 10.4}{1.2} = -1$$

$$Z_B = \frac{X_B - \overline{X_2}}{S_2} = \frac{9.3 - 11.9}{1.3} = -2$$

$\Rightarrow$ Child B is faster because the time taken by child B is two standard deviation shorter than the average time taken by group 2 while, the time taken by child A is only one standard deviation shorter than the average time taken by group 1.

4.4 Moments

Moments are statistical tools used in statistical investigation. The moments of a distribution are the arithmetic mean of the various powers of the deviations of items from some number. In our course, we shall use it in the study of *Skewness* and *Kurtosis* of statistical distribution.

Moments about the origin

$$M_r = \frac{\sum (X_i)^r}{n}$$

Where $r = 0, 1, 2, 3, ...$

Moments about the origin for *grouped frequency distribution* and for *ungrouped frequency distribution*

$$M_r = \frac{\sum f_i (X_i)^r}{n}$$

Where f_i is the frequency of X_i. X_i is the midpoint in the case of *grouped frequency distribution* or *class value* in the case of *ungrouped frequency distribution*.

Note that: $M_1 = \bar{X}$, $M_0 = 1$

Moments about the Mean (Central Moments)

$$M_r' = \frac{\sum (X_i - \bar{X})^r}{n}$$

Moments about the mean for *grouped frequency distribution* and for *ungrouped frequency distribution*

$$M_r' = \frac{\sum f_i (X_i - \bar{X})^r}{n}$$

Where f_i is the frequency of X_i. X_i is the midpoint in the case of *grouped frequency distribution* or *class value* in the case of *ungrouped frequency distribution*.

Note that: $M_r' = SD^2$ if it is assumed $n = n - 1$

Moments about any arbitrary constant A

$$M_r' = \frac{\sum (X_i - A)^r}{n}$$

Moments about any arbitrary constant A for *grouped frequency distribution* and for *ungrouped frequency distribution*

$$M_r' = \frac{\sum f_i (X_i - A)^r}{n}$$

Example 4.11: Find the first four moments about the mean for the following individual series

X	3	6	8	10	18

Solution:

S.No	$\boldsymbol{X}$	$(\boldsymbol{X}-\bar{\boldsymbol{X}})$ $\bar{\boldsymbol{X}}=\mathbf{9}$	$(\boldsymbol{X}-\bar{\boldsymbol{X}})^{\mathbf{2}}$	$(\boldsymbol{X}-\bar{\boldsymbol{X}})^{\mathbf{3}}$	$(\boldsymbol{X}-\bar{\boldsymbol{X}})^{\mathbf{4}}$
1	3	-6	36	-216	1296
2	6	-3	9	-27	81
3	8	-1	1	-1	1
4	10	1	1	1	1
5= $\boldsymbol{n}$	18	9	81	729	6561
Total	$\sum X = 45$	$\sum (X-\bar{X}) = 0$	$\sum (X-\bar{X})^2 = 128$	$\sum (X-\bar{X})^3 = 486$	$\sum (X-\bar{X})^4 = 7940$

Now

$$\bar{X} = \frac{45}{5} = 9$$

$$M_1' = \frac{\sum (X_i - 9)^1}{5} = 0$$

$$M_2' = \frac{\sum (X_i - 9)^2}{5} = \frac{128}{5} = 25.6$$

$$M_3' = \frac{\sum (X_i - 9)^3}{5} = \frac{486}{5} = 97.2$$

$$M_4' = \frac{\sum (X_i - 9)^4}{5} = \frac{7940}{5} = 1588$$

Exercise: Find the relationship between moments about any arbitrary constant A and moments about the mean.

4.5 Skewness

- Skewness is the degree of asymmetry or departure from symmetry of a distribution.
- A skewed frequency distribution is one that is not symmetrical.
- Skewness is concerned with the shape of the curve not size.
- If the frequency curve (smoothed frequency polygon) of a distribution has a longer tail to the right of the central maximum than to the left, the distribution is said to be skewed to the right or said to have positive skewness. If it has a longer tail to the left of the central maximum than to the right, it is said to be skewed to the left or said to have negative skewness.
- For moderately skewed distribution, the following relation holds among the three commonly used measures of central tendency.

$$Mean - \text{mod}e = 3(mean - median)$$

Measures of Skewness

-Denoted by $s_k \ or \ \alpha_3$

-There are various measures of skewness.

1. The Pearsonian coefficient of skewness

$$\alpha_3 = s_k = \frac{mean - \text{mod}e}{std.dev} = \frac{\bar{x} - \hat{x}}{s}$$

2. The moment coefficient of skewness

$$S_k = \frac{M_3}{(M_2)^{3/2}} = \frac{M_3}{(\sigma^2)^{3/2}} = \frac{M_3}{\sigma^3},$$

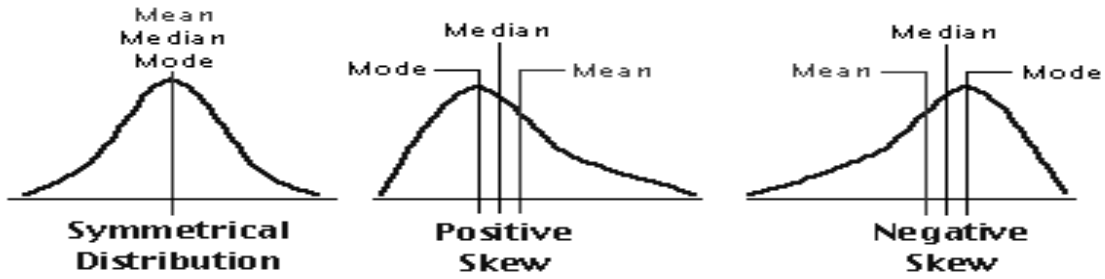
where σ the population standard deviation

The shape of the curve is determined by the value of S_k

- $s_k > 0$ then the *distribution is positively skewed*.
- S_k =0 then the *distribution is symmetric*
- $s_k < 0$ then the *distribution is positively skewed*

Remark:

- In a positively skewed distribution, smaller observations are more frequent than larger observations. i.e. the majority of the observations have a value below an average.
- In a negatively skewed distribution, smaller observations are less frequent than larger observations. i.e. the majority of the observations have a value above an average.



Examples:

1. Suppose the mean, the mode, and the standard deviation of a certain distribution are 32, 30.5 and 10 respectively. What is the shape of the curve representing the distribution?
2. Some characteristics of annually family income distribution (in Birr) in two regions is as follows:

Region	Mean	Median	Standard Deviation
A	6250	5100	960
B	6980	5500	940

a) Calculate coefficient of skewness for each region
b) For which region is, the income distribution more skewed. Give your interpretation for this Region
c) For which region is the income more consistent?

Solutions: (**exercise)**

3. For a moderately skewed frequency distribution, the mean is 10 and the median is 8.5. If the coefficient of variation is 20%, find the Pearsonian coefficient of skewness and the probable mode of the distribution. (**exercise)**
4. The sum of fifteen observations, whose mode is 8, was found to be 150 with coefficient of variation of 20%

a. Calculate the Pearson coefficient of skewness and give appropriate conclusion.
b. Are smaller values more or less frequent than bigger values for this distribution?
c. If a constant *k* was added on each observation, what will be the new Pearson coefficient of skewness? Show your steps. What do you conclude from this?

4.6 Kurtosis

Kurtosis is the degree of peakdness of a distribution, usually taken relative to a normal distribution. A distribution having relatively high peak is called leptokurtic. If a curve representing a distribution is flat topped, it is called platykurtic. The normal distribution which is not very high peaked or flat topped is called mesokurtic.

Measures of kurtosis

The moment coefficient of kurtosis:

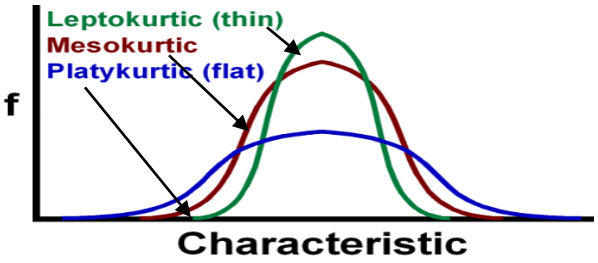
• Denoted by α_4 and given by

$$\alpha_4 = \frac{M_4}{(M_2)^2} = \frac{M_4}{\sigma^4}$$

$$where\ M_4\ the\ fourth moment about the mean.$$
$$M_3\ the third moment about the mean.$$
$$\sigma\ is\ the\ population the\ s\tan dard\ deviation$$

The peakdness depends on the value of α_4 .

- If $\alpha_4 > 3$ then the curve leptokuric
- *If* $\alpha_4 = 3 then the curve is mesokurtic$
- *If* $\alpha_4 < 3 then the curve is platykurtic$



Examples:

1. If the first four central moments of a distribution are:
   M_1=0, M_2=16, M_3=-60, M_4=162

a) Compute a measure of skewness

b) Compute a measure of kurtosis and give your interpretation.

Solutions:

a)
$$S_k = \frac{M_3}{(M_2)^{3/2}} = \frac{-60}{16^{\frac{3}{2}}} = -0.94 < 0$$
$$\Rightarrow The distribution is negatively skewed$$

b)

$$\alpha_4 = \frac{M_4}{(M_2)^2} = \frac{162}{16^2} = 0.6 < 3$$
$$\Rightarrow The curve is leptokurtic$$

2. The median and the mode of a mesokurtic distribution are 32 and 34 respectively. The 4th moment about the mean is 243. Compute the Pearson coefficient of skewness and identify the type of skewness. Assume (n-1 = n).

3. If the standard deviation of a symmetric distribution is 10, what should be the value of the fourth moment so that the distribution is mesokurtic?

Solutions (**exercise)**.