KARNATAKA STATE OPEN UNIVERSITY

MUKTHAGANGOTHRI, MYSORE - 570 006.

DEPARTMENT OF STUDIES AND RESEARCH IN MANAGEMENT

M.B.A I Semester

COURSE - 5

STATISTICS AND OPTIMIZATION TECHNIQUES

BLOCK

1

INTRODUCTION TO STATISTICS

Course Design and Editorial Committee

Prof. M.G. Krishnan
Vice-Chancellor & Chairperson
Karanataka State Open University
Mukthagangothri, Mysore - 570006

Prof. S.N. Vikram Raj Urs
Dean (Academic) & Convenor
Karanataka State Open University
Mukthagangothri, Mysore - 570006

Editors and Subject Co-ordinators

Dr. C. Mahadevamurthy
Associate Professor and Chairman
Department of Management
Karanataka State Open University
Mukthagangothri, Mysore - 570006

Dr. H. Rajeshwari
Assistant Professor
Department of Management
Karanataka State Open University
Mukthagangothri, Mysore - 570006

Course Writers

Smt. Prathiba Jennifer **Block - 1** **(Units 1 to 4)**
Assistant Professor
Department of Management
Mahajana's PG Centre
Mysore.

Publisher

Registrar
Karanataka State Open University
Mukthagangothri, Mysore - 570006

Developed by Academic Section, KSOU, Mysore

Further information may be abtained from the University's office at Mukthagangothri, Mysore-6.

Printed and Published on behalf of Karanataka State Open University, Mysore-6.

MODULE -1 : INTRODUCTION TO STATISTICS

The science of statistics is indispensable for a clear appreciation of any problem affecting all the branches of human knowledge. It covers all the fields of enquiry in which a grasp of the significance of large numbers is looked for. It is applicable to all the disciplines.

This block gives you the fundamental aspects of statistics. Basic concepts of Central Tendency such as mean, median and mode are discussed here. This Module gives you an insight about data collection, tabulation and analysis of data. Methods of calculating disperssion is also taught in this module.

You are expected to understand these concepts and workout problems given at the end of each unit.

This block consists of four units.

Unit 1: Provides an introduction to business statistics i.e. Meaning, scope, importance and limitations, statistics in Business Management.

Unit 2 : Gives an idea about how to analyse of data i.e. Introduction, sources of data, collection, classification, tabulation and depiction of data.

Unit 3: Describes various Measures of Central tendency i.e Arithmetic, weighted, geometric mean, Harmonic mean, median and mode.

Unit 4: Explains the Measures of Dispersion i.e. Range, Quartile deviation, Mean deviation, Standard deviation, variance, Coefficient of variation, Skewness and Kurtosis.

UNIT -1: INTRODUCTION TO BUSINESS STATISTICS

STRUCTURE

1.0 OBJECTIVES

After studying this unit you should be able to :

- Define statistics and explain scope limitations and importance;
- Appreciate use of statistics in Business decisions and
- Analyze the strengths and weaknesses of statistics and the various terminologies of statistics.

1.1 INTRODUCTION

Statistics is not a new discipline. It is as old as this human society itself. The word statistics is derived from Latin word 'Status' or the Italian word "Statista' or the German word 'statistik', each of which means a political state. In the ancient times, the scope of statistics was limited to collection of the data related to age, gender wise population, property and wealth of the country for framing military and fiscal policies.

The development of statistics has been noticed from the time of Pharaohs of Egypt before 2000 years ago and the traces of application of it was seen in Kautilya's Arthashastra and during the time of Chandragupta Maurya. During 16^{th}, 17^{th}, and 18^{th} century systematic development took place in the field of statistics. During this period theory of Probability, theory of Games and chance, Regression and Correlation Analysis, Goodness of Fit tests like Chi Square, t – Test etc were developed. R A Fisher who is called as the Father of Statistics is the pioneer to apply statistics to genetics biometry, psychology, education, agriculture etc which made a remarkable step to introduce statistics to other fields. Thus statistics became a full fledged science. Today statistics has given solutions to various complicated fields like Economics, Business, Management, Accountancy, Social science, Industry, Biology and Medical Sciences.

1.2 DEFINITION AND MEANING OF STATISTICS

It has been defined differently by different writers. Statistics has been expressed as Numerical Data and Statistical methods.

Statistics as Numerical data:

"*Statistics are numerical statements of facts in any department of enquiry placed in relation to each other.*" – **Bowley**

"By statistics we mean quantitative data affected to a marked extent by multiplicity of causes." – **Yule and Kendall**

"Statistics are the classified facts representing the conditions of the people in a state, such as those facts which can be represented in tables as numbers in any classified arrangement" **- Webster**

Statistics as Statistical Methods:

"Statistics is the science of estimates and probabilities" **– Boddington**

"Statistics may be called as the science of counting or averages" – **Bowley**

"Statistics is a method of decision making in the face of uncertainty on the basis of numerical data and calculated risks" – **Prof. Ya-Lun-Chou**

Thus we can summarize that statistics is :

- Aggregate of Facts
- Affected by multiplicity of causes
- Numerically expressed
- Enumerated as estimated as a reasonable Standard of accuracy
- Collected in a systematic Manner
- Collected for a predetermined purpose and comparable
- It is both science and art
- Helps to arrive to a valid decision
- Evaluates the various alternatives

So Statistics is a subject which consists of the processes, techniques, and methods of collecting , organizing, presenting analyzing and interpreting data for decision making in uncertainty.

1.3 SCOPE OF STATISTICS

In olden days it was regarded as the science of statecraft. Today the scope has widened to various phenomenon from social aspects to economics. Today it is not only used to collect numerical data but also for their handling, analysis and drawing inferences from them. It embraces all fields of sciences and finding numerous solutions to various disciplines such as industry, business, biometry, economics planning, sociology, insurance etc. Today it has become indispensible for the life of a citizen.

1.4 IMPORTANCE OF STATISTICS

The importance of statistics can be understood by its wide applicability in the fields of

- ***Planning*** of Government level like five year plans, budgets etc.
- ***Statistics in state – craft*** like collecting e data relating to manpower , Military, crime, income and wealth etc to formulate suitable policies.
- ***Statistics in Economics*** – it gives solutions to various problems of pricing, production, supply, consumption, distribution wealth and income, savings, profits, investments,

expenditure etc. various laws of economy are developed through statistics. Powerful tools (trend Analysis, Time Series, Forecasting techniques) are used in analysis of economic data.

- ***Business and Management***- Business deals with uncertainty and frequently it lands in dilemmas. Statistics is the scientific ways to come out of the ambiguities of business and helps widely to take decisions.
- ***Accountancy and Auditing*** – Today Chartered Accountants and ICWAs have statistics as a vital subject in their syllabus since its usages in accounting has become inevitable. It is widely used in profit analysis, dividend decisions, assets and liability, analysis, etc. in auditing sampling techniques are used for test checking of voluminous data related to the business transactions.
- ***Statistics in Industry*** – it is used intensively in Quality Control in production process
- ***Statistics in Physical sciences*** – physical sciences like astronomy, geology, engineering, and meteorology expects accurate results in their applications. Statistics techniques like least square methods give solutions to this.
- ***Statistics in Social Science*** – to study the demography features, mortality, fertility, population growth, poverty etc requires statics to analyze data
- Statistics is also used in Biology, Medical sciences to study the cause and effects, and also it is used in Psychology and education like scaling of mental health, determining I Q levels etc.

1.5 LIMITATIONS OF STATISTICS

Statistics is indispensible for almost all the activities of human activities. But it has few limitations such as

1. It does not study Qualitative Phenomenon
2. It does not study individuals
3. Statistical laws are not exact
4. Statistics is liable to be misused
5. Expertise knowledge is required to analyze the data and interpret it.

1.6 STATISTICS IN BUSINESS AND MANAGEMENT

Today the development in business activities has variety of dimensions in size and competitions in the market. Every step in the business or management has become complicated. Personal evaluation based on observations is not enough. A team of specialized

managerial executives are inevitable for running up of today's business activities such as sales, purchases, production, control, finance, marketing etc. In this the statistical tools and theories such as forecasting techniques, estimation theory, sampling, probability, least square, game theory etc. play an indispensible role. *"Statistics is a method of decision making in the face of uncertainty on the basis of numerical data and calculated risks"* – **Prof. Ya – Lun - Chou and** According to **Wallis and Roberts** *" Statistics may be regarded as a body of methods for making wise decisions in the face of uncertainty."* These definitions reflect the application of statistics in today's modern business which has its roots in accuracy in estimation and forecasting regarding future demand for the product, market trends and so on. The statistical information related to the business is also acts as a guide to future economic events. The uses of statistics in some of the business decisions are:

- To estimate the probable trends in demand of the goods
- Ordering the right quantity of Materials
- Seasonal and cyclical movements of the business
- Relationship between supply and demand
- To know the purchasing power of money
- Statistical quality control of production to produce without waste
- To promote the new businesses
- To conduct the customer surveys and collect the demographic information to fix the target segments.
- To understand the consumer expectations and level of product awareness
- To launch the new products through sample surveys
- Optimization of Profits and Investments and minimizing the expenses
- To forecast the future and balance the uncertainties through probability and estimation theories

Thus uses of statistics have become indispensible in all the branches of business activities.

1.7 DESCRIPTIVE AND INFERENTIAL STATISTICS IN BUSINESS DECISIONS

There are two major divisions in statistics: - *Descriptive Statistics* and *Inferential Statistics*.

Descriptive Statistics: Descriptive statistics deals with collecting, summarizing and simplifying the data which are otherwise very voluminous. Through this, meaningful conclusion can be drawn readily from the data. Thus this method facilitates an understanding

of the data and systematic reporting which makes the data useful for further discussions, analysis and interpretations. A well thought out data classification facilitates easy descriptions and a variety of summary measures. These include Measures of Central Tendency, Dispersion, Skewness and Kurtosis which constitute the essential scope of descriptive statistics.

Inferential Statistics: this is also called as inductive statistics. It goes beyond describing a given problem situation by means of collecting summarizing and presenting related data. Instead it consists of the methods that are used for drawing inferences, making broad generalizations about total observations on the basis of a part from it. That is obtaining a particular value from the sample information and using it for drawing an inference about the entire population is inferential statistics.

In business, the decisions are to be taken in uncertainty and most of the time; total coverage of the information through Census method is not possible. And it may not be always feasible and practical for various reasons. In such situation it is the inferential statistics which is used in taking business decisions.

Risk evaluation and Statistics:

Inferential statistics helps to evaluate the risk involved in getting inferences or generalizations about an unknown population on the basis of sample information. There is always a risk of an inference about a population being incorrect when based on the knowledge limited to a sample. The rescue lies in evaluating the risk. The probability distributions help us for drawing statistical inferences and estimating the degree of reliability of these inferences.

1.8 STRENGTHS AND WEAKNESSES OF STATISTICS

Strengths: It develops statistical mode of thinking. Collects and compiles massive set of data which are systematically and carefully analyzed to seek useful insights and reach valid conclusions for sound decision making. Statistics creates a flexible mind with a judicious sense that helps to understand the dangers.

Weakness : The general feeling of distrust in it is the important weakness of statistics. Element of accuracy and reliability of the data is often questioned. The whole process from collection of data till interpreting the results is porous and allows number of errors in each step. The data can be manipulated easily by the collector himself. These flaws in statistics can be minimized but cannot be eliminated. Thus the inaccuracy, unreliability, manipulatin of data creates distrust in statistics.

1.9 LANGUAGE OF STATISTICS

Statistics uses some common Concepts or Notions while analyzing and interpreting the data. Notions are the shorthand expressions of concepts and statements which are also called language of statistics. Various concepts used are Variables, Observed data or values of the variables, samples, sample size, population, sample statistics etc.

1.10 SUMMARY

A layman knows that statistics is data. It is the numerical information expressed in quantitative terms. It can also be called as science of data. This unit has thrown light on the basics of statistics. Today statistics has grown into a separate subject since its importance in each and every aspect of human life and its relevancy in almost all disciplines. Today Business decisions are based on statistical inferences as the future is very uncertain. Though there are many advantages of using statistical tools in decision making, the distrust still prevails because of the inadequacy, inaccuracy, manipulation of data, lack of expertise skill and knowledge to interpret the results.

1.11 KEY WORDS

Statistics: The aggregate of facts which are numerically expressed collected and classified

Uncertainty: Insecurity about the future.

1.12 SELF ASSESSMENT QUESTIONS

a. Explain the utility of statistics in Business and Management in the current scenario.
b. Explain the importance of statistics in detail.
c. "Statistics is a method of decision making in the face of uncertainty on the basis of the numerical data and calculated risks." Explain with suitable illustrations.
d. Define statistics as a discipline. Also bring out its scope.
e. What are the causes of distrust in statistics?
f. Differentiate between descriptive and inferential statistics. How inferential statistics is useful in business decisions?

1.13 REFERENCES

1. Gupta S.*P. Business Statistics*, New Delhi: S Chand and Sons Publishers, 2000
2. Shahsi Kumar. *Quantitative Techniques and methods*, Mysuru: Chetana Book House, 2010
3. Vignanesh Prajapathi, *Big data Analysis With R and Hadoop*, Mumbai: Packt Publishing, 2013
4. SD Sharma, *Operation Research,* Delhi: Discovery Publishing House, 1997
5. Srinath L. S, *PERT and CPM*, Delhi: East West Press,2001
6. Kalavathy, *Operation Research* , New Delhi: Vikas Publishing House, 2010
7. Richard I. Levin. *Statistics for Management,* New Delhi: Pearson education India, 2008

UNIT 2 : ANALYSIS OF DATA

STRUCTURE

2.0 OBJECTIVES

After studying this unit you should be able to :

- Define statistical data;
- Explain types of data and the sources of collecting the data;
- Demenstrate to present a data and tabulation and
- Describe Graphic presentation of data.

2.1 INTRODUCTION

Data is the information that is collected. Statistical data are the basic raw material of statistics. Data may relate to an activity of our interest, a problem, or a phenomenon or a situation under study. The data is the result of the process of measuring, counting or observing. Therefore Statistical data refer to those aspects of a problem situation that can be measured, quantified, counted or classified. In any statistical investigation, the collection of the numerical data is the first and the most important matter to be attended. Often a person investigating, will have to collect the data from the actual field of inquiry. For this he may issue suitable questionnaires to get necessary information or he may take actual interviews; personal interviews are more effective than questionnaires, which may not evoke an adequate response. Another method of collecting data may be available in publications of Government bodies or other public or private organizations. Sometimes the data may be available in publications of Government bodies or other public or private organizations. Such data, however, is often so numerous that one's mind can hardly comprehend its significance in the form that it is shown. Therefore it becomes, very necessary to tabulate and summarize the data to an easily manageable form. In doing so we may overlook its details. But this is not a serious loss because Statistics is not interested in an individual but in the properties of aggregates. For a layman, presentation of the raw data in the form of tables or diagrams is always more effective.

Prerequisites of statistical data:

a. It should be unambiguous
b. It should be specific and as per the objectives and scope of the study
c. It should be stable
d. It should be appropriate to the enquiry
e. It should be uniform
f. Degree of accuracy should be aimed.
g. It should be apt and reliable.

2.2 TYPES OF DATA

A. Based on the characteristics, measured data can be classified into two broad categories.

a. Quantitative data

b. Qualitative data

Quantitative data: The data that can be quantified in definite units of measurement are called as quantitative data. That is the successive measurements yield quantifiable observations. Depending on the nature of the variable observed or measurement this can be further categorized as ***continuous data and discrete data***. Continuous data represent the numerical values of a continuous variable. A continuous variable is the one that can assume any value between any two points on a line segment and thus it represents an interval of values. For example: temperature, thickness, velocity, height, weights etc. Discrete data are the values assumed by discrete variables. That one whose outcomes are measured in fixed numbers. Ex: Number of customers visiting a store every day, the number of trains arriving at the station, number of defects in one consignment etc.

Qualitative data: The data related to the qualitative characteristics of the subject or object is qualitative data. This data is the data gathered through the attributes. These data are classified as ***Nominal Data and Rank Data.*** The count data obtained from classification is called as Nominal data. For example: classification of students according to gender, division of workers as per skills, education etc. Rank data is the result of assigning the ranks. For example: ranking as per the performance in the interview, exams, performance as per quality etc.

B. Based on the sources of data the data can be categorized as :

a. Primary data

b. Secondary data

Primary data: these are the data that do not exist in any form, and thus have to be collected for the first time from the primary sources. Since they are collected for the first time, they are the fresh data. And they are the data collected from the sample drawn from the whole population.

Secondary data: these data already exist in some form published or unpublished in an

identifiable secondary source.

2.3 DATA COLLECTION AND SOURCES OF PRIMARY AND SECONDARY DATA

Data collection is the act of assembling and gathering the needed numerical information for a research. The process of counting or measuring together with the systematic recording of results is called collection of statistical data. Collection can be from primary or secondary source.

Preliminaries of Data Collection:

- Objectives and scope of enquiry
- Statistical units to be used
- Sources of information
- Method of data collection
- Degree of accuracy aimed at in the final results
- Type of enquiry

Primary data do not exist in any form the only source from where they can be collected is of field surveys from the population or the sample from the population. Primary data sources can be of internal and external.

Methods of collecting primary data are:

1. Personal Interviews
2. Direct personal Investigation
3. Indirect oral interviews
4. Information received through local agencies
5. Mailed questionnaire method
6. Schedules through enumerators

Secondary Data sources may be broadly classified as two groups:

(i) Published sources : Official Publications of Central Government, Publications of semi government Statistical Organizations, publications of research institutions, Commercial and Financial Institutions, Reports of Various Committees and Commissions appointed by the Government, News papers and Periodicals, International Publications are the sources of secondary data in published form

(ii) Unpublished sources : Records maintained by the private firms, the research carried

out by individuals, records of business concerns etc. The other forms of sources are Internal Sources and External sources. This depends on the type of the user of the data whether he is an insider or the outsider. As caution the secondary data should be carefully examined before use to see that they suit the objectives of the research or the study. But these data are more convenient to use especially when used as supportive evidences. The other advantage of secondary data is its easy availability, convenient to reach and access and not much effort is needed to classify and tabulate the data. The precautions to be taken are to check whether the data is reliable, suitable and adequate.

2.4 CLASSIFICATION AND TABULATION OF DATA

Classification "Classified and arranged facts speak of themselves, and narrated they are as dead as mutton" This quote is given by J.R. Hicks.The process of dividing the data into different groups (viz. classes) which are homogeneous within but heterogeneous between themselves, is called a classification.It helps in understanding the salient features of the data and also the comparison with similar data. For a final analysis it is the best friend of a statistician.

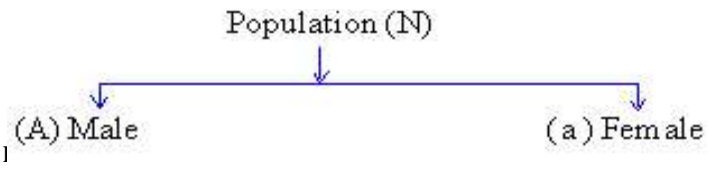
Methods Of Classification

The data is classified in the following ways:

1. According to attributes or qualities this is divided into two parts :
   (A) Simple classification
   (B) Multiple classifications.
2. According to variable or quantity or classification according to class intervals. -

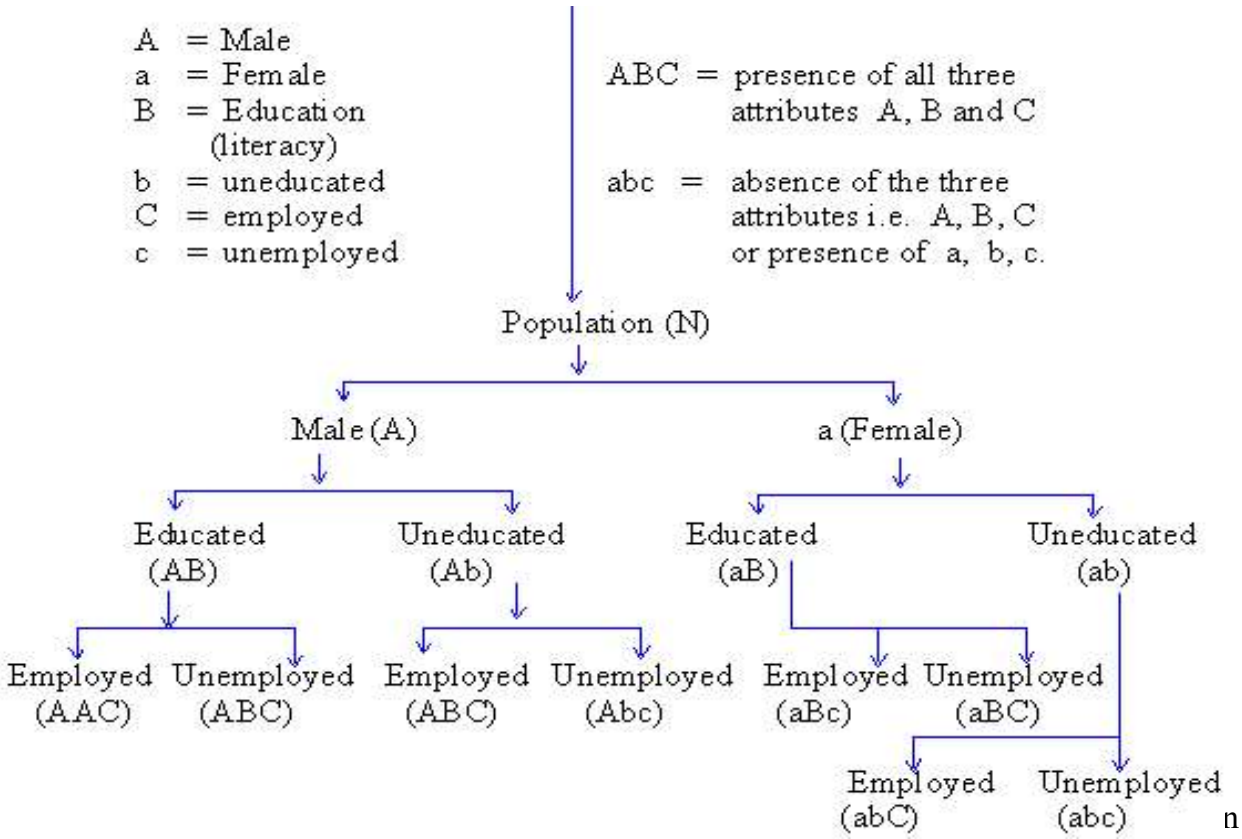
Qualitative Classification : When facts are grouped according to the qualities (attributes) like religion, literacy, business etc., the classification is called as qualitative classification.

(A) Simple Classification : It is also known as classification according to Dichotomy. When data (facts) are divided into groups according to their qualities, the classification is called as 'Simple Classification'. Qualities are denoted by capital letters (A, B, C, D .....) while the absence of these qualities are denoted by lower case letters (a, b, c, d, ... etc.) For example,



(B) Manifold or ssified using one or

more qualities. First, the data is divided into two groups (classes) using one of the qualities. Then using the remaining qualities, the data is divided into different subgroups. For example, the population of a country is classified using three attributes: sex, literacy and business as,



numbers (quantitative data), is classified according to class-intervals. While forming class-intervals one should bear in mind that each and every item must be covered. After finding the least value of an item and the highest value of an item, classify these items into different class-intervals. For example if in any data the age of 100 persons ranging from 2 years to 47 years, is given, then the classification of this data can be done in this way:**Table - 1**

Age in years	No. of persons
0 - 10	5
10 - 20	9
20 - 30	32
30 - 40	34
40 - 50	20

According to the clas rms are used :

Class-limits: A class is formed within the two values. These values are known as the class-limits of that class. **Magnitude of the class-intervals :** The difference between the upper and lower limits of a class is called the magnitude or length or width of a class and is denoted by ‘ i ‘ or ‘ c **Mid-value or class-mark :** The arithmetical average of the two class limits (i.e. the lower limit and the upper limit) is called the mid-value or the class mark of that class-interval. **Class frequency :** The units of the data belong to any one of the groups or classes. The total number of these units is known as the frequency of that class and is denoted by fi or simply f

Classification is of two types according to the class-intervals - (i) Exclusive Method (ii) Inclusive Method.

Exclusive Method : In this method the upper limit of a class becomes the lower limit of the next class. It is called ‘ Exclusive ‘ as we do not put any item that is equal to the upper limit of a class in the same class; we put it in the next class, i.e. the upper limits of classes are excluded from them. See table 1. For example, a person of age 20 years will not be included in the class-interval (10 - 20) but taken in the next class (20 - 30), since in the class interval (10 - 20) only units ranging from 10 - 19 are included.

Inclusive Method : In this method the upper limit of any class interval is kept in the same class-interval. In this method the upper limit of a previous class is less by 1 from the lower limit of the next class interval. In short this method allows a class-interval to include both its lower and upper limits within it. For example :

Table - 2

Inclusive method		Inclusive method	
Class	Frequency	Class	Frequency
0 - 4	5	0 - 4.9	5
5 - 9	7	5 - 9.9	7
10 - 14	9	10 - 14.9	9
15 - 19	12	15 - 19.9	12
20 - 24	11	20 - 24.9	11
25 - 29	14	25 - 29.9	14

Open-end Css-interval

or the upper limit of the last class-interval, are not given then subtract the class length of the next immediate class-interval from the upper limit. This will give us the lower limit of the first class-interval. Similarly add the same class length to the lower limit of the last class-interval. But always notice that the lower limit of the first class (i.e. the lowest class) must not be negative or less than 0. For example :

Table - 3

With open ends	Completed classes	With open ends	Completed classes
Below 10	0 - 10	Below 10	0 - 10
10 - 20	10 - 20	10 - 25	10 - 25
20 - 30	20 - 30	25 - 40	25 - 40
30 - 40	30 - 40	40 - 70	40 - 70
40 - 50	40 - 50	above 70	70 - 100
above 50	50 - 60		

Tabulation

It is the process of condensation of the data for convenience, in statistical processing, presentation and interpretation of the information. A good table is one which has the following requirements:

1. It should present the data clearly, highlighting important details.
2. It should save space but attractively designed.
3. The table number and title of the table should be given.+
4. Row and column headings must explain the figures therein.
5. Averages or percentages should be close to the data.
6. Units of the measurement should be clearly stated along the titles or headings.
7. Abbreviations and symbols should be avoided as far as possible.
8. Sources of the data should be given at the bottom of the data.
9. In case irregularities creep in table or any feature is not sufficiently explained, references and foot notes must be given.
10. The rounding of figures should be unbiased.

Types of Tables: **The important types of statistical table are as follows:**

1. Single column or Single Row Tables
2. Multiple column or multiple row tables
3. Reference and Summary tables

Components of a Table: **the structure or the components of the table should have:**

1. Table Number
2. Title of the table
3. Head Notes
4. Stub and Stub Heads
5. Box Head and Sub Heads
6. Body of the table
7. Footnote
8. Source

2.5 SUMMARIZING THE DATA – FREQUENCY DISTRIBUTION

The frequency distribution is the outcome of a process of classification of individual observations of a set of data into an appropriate number of classes. It is also called as grouped data. The frequency distribution can be constructed through

a. Tally Method and b. Entry form Method

Relative Frequency: The relative frequency of a class is the frequency of the class divided by the total number of frequencies of the class and is generally expresses as a percentage.

Cumulative Frequency: Many a times the frequencies of different classes are not given. Only their cumulative frequencies are given. The total frequency of all values less than or equal to the upper class boundary of a given class-interval is called the cumulative frequency up to and including that class interval. These cumulative frequencies are called less than or more than cumulative frequencies. For example,

Class – interval	0-10	10-20	20-30	30-40	40-50
Frequency	4	9	5	12	15

Table -5

Less than cumulative frequency		More than cumulative frequency	
(Upper limits)	(cum. freq.)	(Lower limits)	(cum. freq.)
Less than 10	4	More than 0	45 = 41 + 4
Less than 20	4 + 9 = 13	More than 10	41 = 32 + 9
Less than 30	13 + 5 = 18	More than 20	32 = 27 + 5
Less than 40	18 + 12 = 30	More than 30	27 = 15 + 12
Less than 50	30 + 15 = 45	More than 40	15

Preparation of Frequency Distribution

Consider the data collected by one of the surveyors, interviewing about 50 people. This is as follows :

Size of the shoes : 2, 5, 6, 8, 2, 5, 6, 7, 6, 8, 7, 4, 3, .. This is called the raw data. Here some values repeat themselves. For instance the size 5 is repeated 10 times in 50 people. We say that the value of 5 of the variate has the frequency of 10. ***Frequency means the number of times a value of the variate or an attribute, as the case may be, is repeated in the data.*** A table which shows each value of the characteristic with its corresponding frequency, is known as a **Frequency Distribution**. The procedure of preparing such a table is explained as below:

Discrete variate : Consider the raw data which gives the size of shoes of 30 persons

2, 5, 6, 4, 5, 7, 4, 4, 6, 2
3, 5, 5, 4, 5, 6, 5, 4, 3, 2
4, 4, 5, 4, 5, 5, 3, 2, 4, 4

The least value is 2 and the highest is 7. All sizes are integers between 2 and 7 (both inclusive). We can prepare a frequency distribution table as follows :

Table - 6

Sizes of shoes	Tally Marks	Frequency
2	||||	4
3	|||	3
4	𝍸 𝍸	10
5	𝍸 𝍸	9
6	|||	3
7	|	1
Total		30

In this example the size difference from 2 to 7 is very small. If the range of a variate is very large, it is inconvenient to prepare a frequency distribution for each value of the variate. In such a case we divide the variate into convenient groups and prepare a table showing the groups and their corresponding frequencies. Such a table is called a **grouped frequency distribution**.

Consider the marks (out of 100) of 50 students as below :

40,	39,	43,	62,	30,	47,	33,	31,	17,	28
36,	29,	40,	32,	39,	24,	57,	42,	15,	30
50,	52,	47,	65,	31,	07,	37,	47,	17,	20
25,	53,	65,	85,	89,	56,	55,	41,	43,	10
44,	40,	69,	22,	40,	65,	39,	36,	71,	12

The range of the variate (marks) is very large. Also we are eager to know the performance of the students. The passing limit is 35 and above. Marks between 35 and 44 form the third class (or grade). Marks ranging between 45 - 59 are considered as second class and 60 - 100 form the first class. Thus we have a grouped frequency distribution as:

Table

Marks	Tally Marks	Frequency
0 - 34	𝍸 𝍸 𝍸 |	16
35 - 44	𝍸 𝍸 𝍸 |||	18
45 - 59	𝍸 ||||	9
60 - 100	𝍸 ||	7
Total		50

Concerns in constructing frequency distribution:

The factors or issues that have to be kept in mind before constructing a frequency distribution are:

1. Number of Classes
2. Width of the class Intervals
3. Establishing the initial class
4. Stated and real Class Limits

2.6 DIAGRAMMATIC AND GRAPHIC REPRESENTATION OF DATA

It is not always easy for a layman to understand figures, nor is it is interesting for him. Apart from that too many figures are often confusing. One of the most convincing and appealing ways in which statistical results may be represented is through graphs and diagrams. It is for this reason that diagrams are often used by businessmen, newspapers, magazines, journals, government agencies and also for advertising and educating people.The various graphic presentation of data can be done through**:**

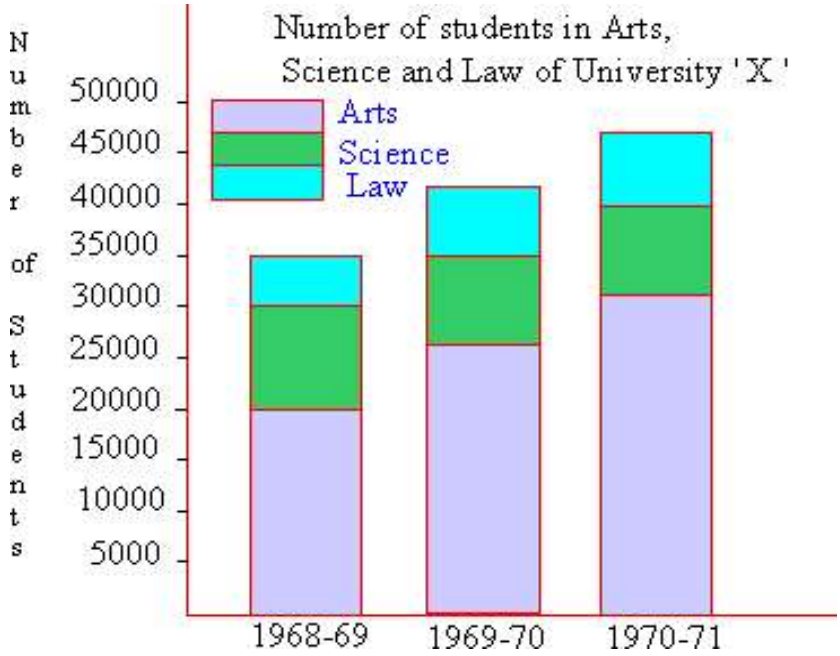
1. Bar Diagrams

1) Simple 'Bar diagram':- It represents only one variable. For example sales, production, population figures etc. for various years may be shown by simple bar charts. Since these are of the same width and vary only in heights (or lengths), it becomes very easy for readers to study the relationship. Simple bar diagrams are very popular in practice. A bar chart can be either vertical or horizontal; vertical bars are more popular.

2) Sub - divided Bar Diagram:- While constructing such a diagram, the various components in each bar should be kept in the same order. A common and helpful arrangement is that of presenting each bar in the order of magnitude with the largest component at the bottom and the smallest at the top. The components are shown with different shades or colors with a proper index.

Illustration:- During 1968 - 71, the number of students in University ' X ' are as follows. Represent the data by a similar diagram.

Year	Arts	Science	Law	Total
1968 - 69	20,000	10,000	5,000	35,000
1969 - 70	26,000	9,000	7,000	42,000
1970 - 71	31,000	9,500	7,500	48,000

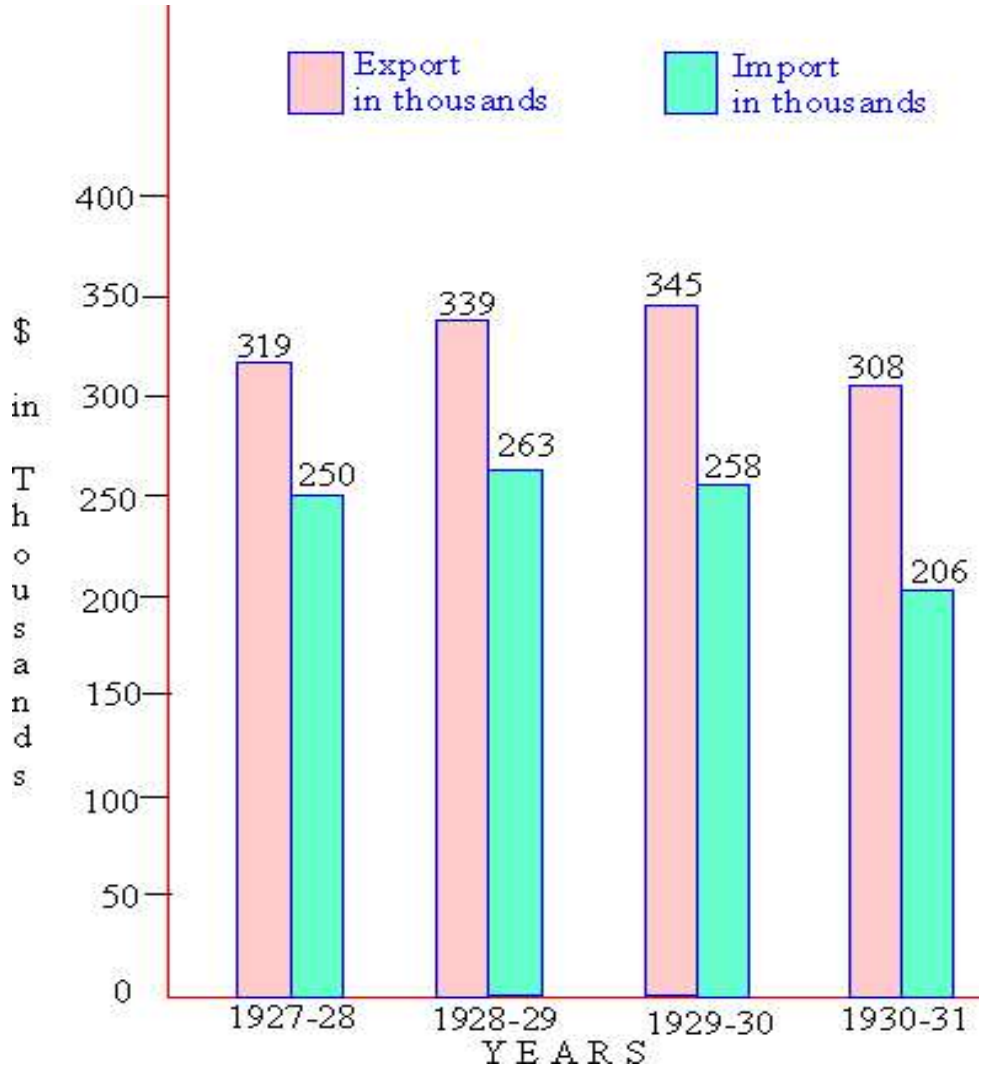


3) Multiple Bar Diagram:- This method can be used for data which is made up of two or more components. In this method the components are shown as separate adjoining bars. The height of each bar represents the actual value of the component. The components are shown by different shades or colors. Where changes in actual values of component figures only are required, multiple bar charts are used.

Illustration:- The table below gives data relating to the exports and imports of a certain country X (in thousands of dollars) during the four years ending in 1930 - 31.

Year	Export	Import
1927 - 28	319	250
1928 - 29	339	263
1929 - 30	345	258
1930 - 31	308	206

Represent the data by a suitable diagram



2. Pie Chart

Geometrically it can be seen that the area of a sector of a circle taken radically, is proportional to the angle at its center. It is therefore sufficient to draw angles at the center, proportional to the original figures. This will make the areas of the sector proportional to the basic figures.

For example, let the total be 1000 and one of the component be 200, then the angle will be

$$\left(\frac{200}{1000}\right) \times 360^0 = 72^0$$

iii) As an example consider the yearly expenditure of a Mr. Ted, a college undergraduate.

Tuition fees	$ 6000
Books and lab.	$ 2000
Clothes / cleaning	$ 2000
Room and boarding	$ 12000
Transportation	$ 3000
Insurance	$ 1000
Sundry expenses	$ 4000
Total expenditure =	$ 30000

Now as explained above, we calculate the angles corresponding to various items (components).

Tuition fees $= \frac{6000}{30000} \times 360^0 = 72^0$

Book and lab $= \frac{2000}{30000} \times 360^0 = 24^0$

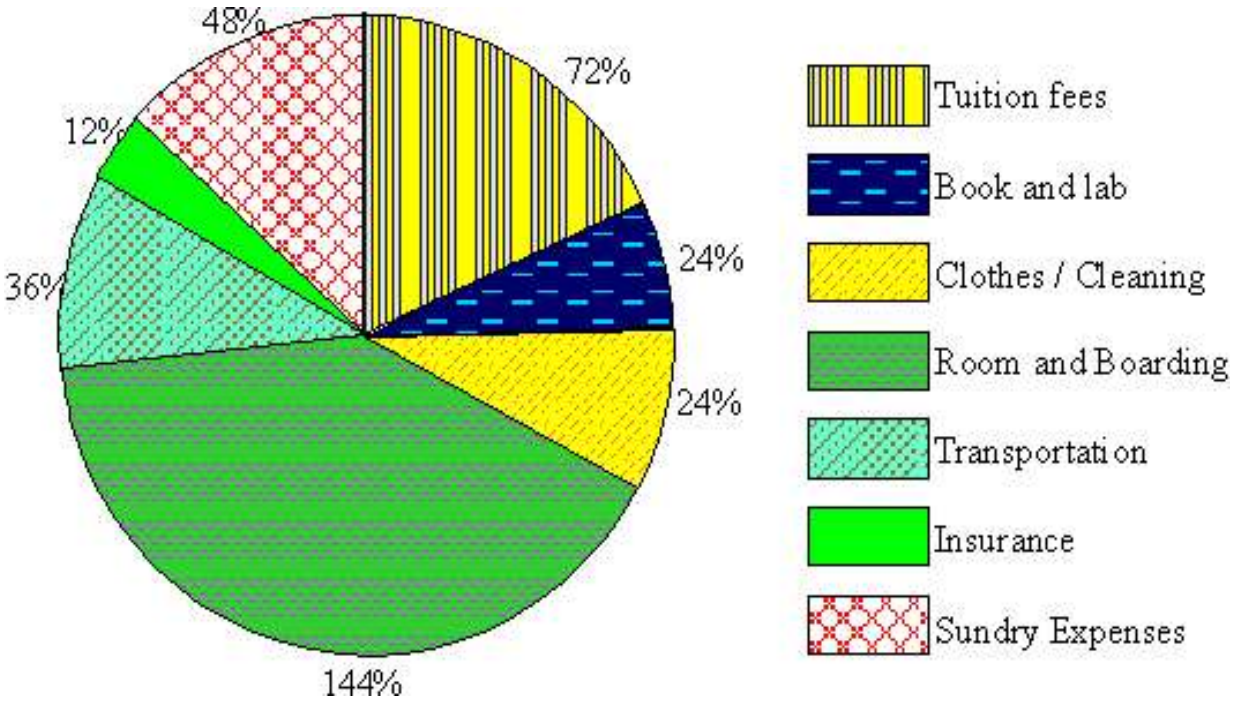
Clothes / cleaning $= \frac{2000}{30000} \times 360^0 = 24^0$

Room and boarding $= \frac{12000}{30000} \times 360^0 = 144^0$

Transportation $= \frac{3000}{30000} \times 360^0 = 36^0$

Insurance $= \frac{1000}{30000} \times 360^0 = 12^0$

Sundry expenses $= \frac{4000}{30000} \times 360^0 = 48^0$



Uses:- A pie diagram is useful when we want to show relative positions (proportions) of the figures which make the total. It is also useful when the components are many in number.

3. Graphs

A graph is a visual representation of data by a continuous curve on a squared (graph) paper. Like diagrams, graphs are also attractive, and eye-catching, giving a bird's eye-view of data and revealing their inner pattern.

Graphs of Frequency Distributions:-

The methods used to represent a grouped data are :-

1. Histogram
2. Frequency Polygon
3. Frequency Curve
4. Ogive or Cumulative Frequency Curve

1. **Histogram :-** It is defined as a pictorial representation of a grouped frequency distribution by means of adjacent rectangles, whose areas are proportional to the

frequencies.

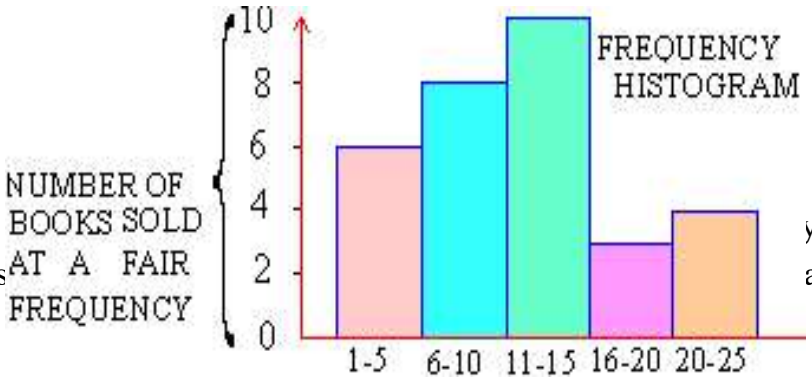
For example, in a book sale, you want to determine which books were most popular, the high priced books, the low priced books, books most neglected etc. Let us say you sold a total 31 books at this book-fair at the following prices.

Rs....2, Rs 1, Rs 2, Rs 2, Rs 3, Rs. 5, Rs. 6, Rs. 17, Rs.17, Rs.7, Rs.15, Rs.7, Rs. 7, Rs.18, Rs. 8, Rs.10, Rs. 10, Rs. 9, Rs. 13, Rs.11, Rs 12, Rs. 12, Rs. 12, Rs. 14, Rs.16, Rs. 18, Rs. 20, Rs. 24, Rs.21, Rs. 22, Rs. 25.

The books are ranging from $1 to $25. Divide this range into number of groups, class intervals. Typically, there should not be fewer than 5 and more than 20 class-intervals are best for a frequency Histogram.Therefore now we have distribution of books at a book-fair

Class-interval	Frequency
$ 1- $ 5	6
$6 - $10	8
$11 - $15	10
$16 - $20	3
$21 - $25	4
Total	$n = \Sigma f_i = 31$

Note that each class-interval is of equal width i.e. $5 inclusive. Now we draw the frequency Histogram as under.



Relative y difference is that it compares ad of the frequency of

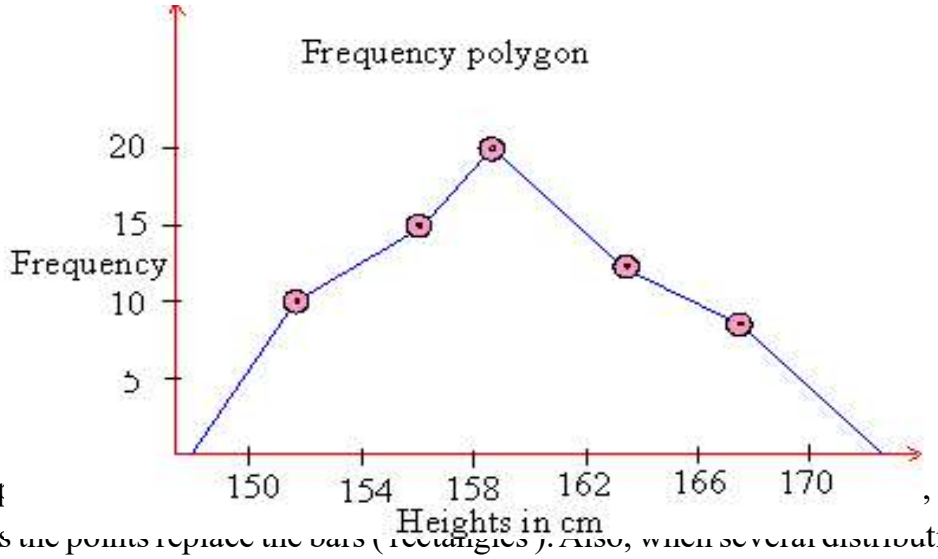
each class-interval, their relative frequencies are used. Naturally the vertical axis (i.e. y-axis) uses the relative frequencies in places of frequencies.

2 Frequency Polygon:- Here the frequencies are plotted against the mid-points of the class-intervals and the points thus obtained are joined by line segments.

Example : -

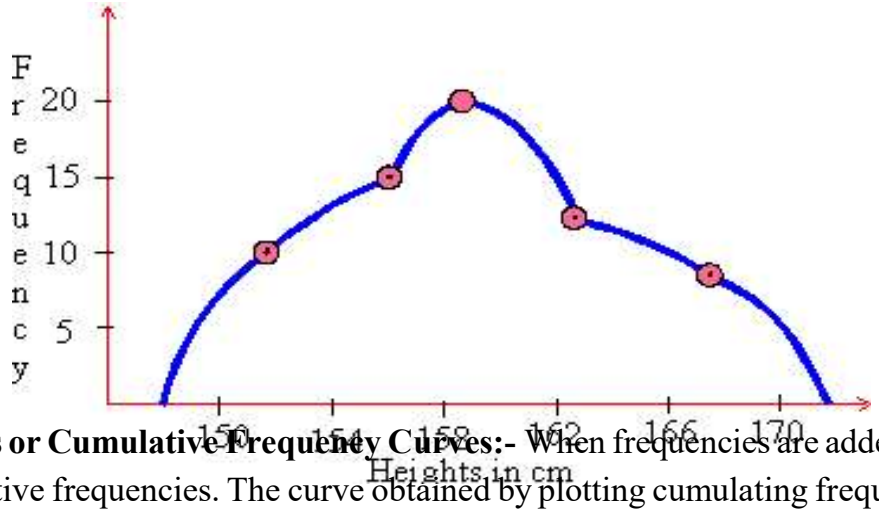
Height in cm.	150 - 154	154 - 158	158 - 162	162 - 166	166 - 170
No. of children	10	15	20	12	8

The polygon is closed at the base by extending it on both its sides (ends) to the midpoints of two hypothetical classes, at the extremes of the distribution, with zero frequencies.



On comp , in frequency polygons the points replace the bars (rectangles). Also, when several distributions are to be compared on the same graph paper, frequency polygons are better than Histograms.

3) Frequency Distribution (Curve):- Frequency distribution curves are like frequency polygons. In frequency distribution, instead of using straight line segments, a smooth curve is used to connect the points. The frequency curve for the above data is shown as:



4. gives or Cumulative Frequency Curves:- When frequencies are added, they are called cumulative frequencies. The curve obtained by plotting cumulating frequencies is called a cumulative frequency curve or an ogive (pronounced ojive).

To construct an Ogive:-

1) Add up the progressive totals of frequencies, class by class, to get the cumulative frequencies.

2) Plot classes on the horizontal (x-axis) and cumulative frequencies on the vertical (y-axis).

3) Join the points by a smooth curve. Note that Ogives start at (i) zero on the vertical axis, and (ii) outside class limit of the last class. In most of the cases it looks like 'S'.Note that cumulative frequencies are plotted against the 'limits' of the classes to which they refer.

(A) **Less than Ogive:-** To plot a less than ogive, the data is arranged in ascending order of magnitude and the frequencies are cumulated starting from the top. It starts from zero on the y-axis and the lower limit of the lowest class interval on the x-axis.

(B) **Greater than Ogive**:- To plot this ogive, the data are arranged in the ascending order of magnitude and frequencies are cumulated from the bottom. This curve ends at zero on the the y-axis and the upper limit of the highest class interval on the x-axis.

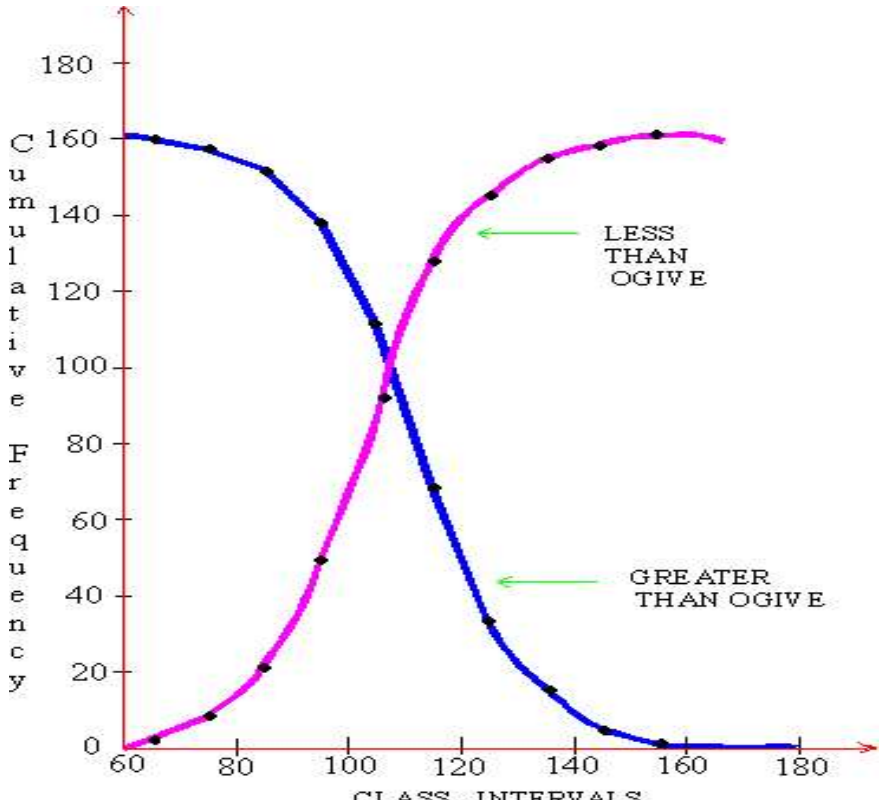
Illustrations:- On a graph paper, draw the two ogives for the data given below of the I.Q. of 160 students.

Class -intervals :60 - 70 70 - 80 80 - 90 90 - 100 100 - 110

No. of students : 2 7 12 28 42

110 - 120 120 - 130 130 - 140 140 - 150 150 - 160
36 18 10 4 1

Solution classes	f	c.f. less than	c.f. greater tha
60 - 70	2	2	159+1 = 160
70 - 80	7	2+7 = 9	151+7 = 158
80 - 90	12	9+12 = 21	139+12 = 151
90 -100	28	21+28 = 49	111+28 = 139
100 -110	42	49+4 = 91	69+42 = 111
110 -120	36	91+36 = 127	33+36 = 69
120 -130	18	127+18 = 145	15+18 = 33
130 -140	10	145+10 = 155	5+10 = 15
140 -150	4	155+4 = 159	1+4 = 5
150 -160	1	159+1 = 160	1
	Σf=160		



Uses : - Certain values like median, quartiles, deciles, quartile deviation, coefficient of

skewness etc. can be located using Ogives. it can be used to find the percentage of items having values less than or greater than certain value. Ogives are helpful in the comparison of the two distributions.

2.7 SUMMARY

Data is the information that is collected and it is the raw material for statistics. This data has to be collected in asystematic manner from the right source like Primary or secondary. Thus collected data should be classified and tabulated for further process. This statistical data can be presented through frequency distributions or through Graphs to understand them and process them easily.

2.8 KEY WORDS

Data	–	the information collected and compiled
Population	-	the totalityofobservations made
Sample	–	the part of totality actually observed to collect the data and analysis and togeneralise about the population is a sample.
Reference Tables	-	the ones which preasent extensive information on any subject
Ordered Array	-	a convinent order to arrasnge the data. It can bei ascending order or in decending order.
class interval	-	the width of the class

2.9 SELF ASSESSMENT QUESTIONS

1. Why should a given set of data be presented in an organised form? Explain.
2. List the advantages of converting the data to a frequency distribution.
3. What are the sources of data ? Explain in detail
4. What are merits and demerits of using secondary data?
5. What problems do unequal class intervals create ? Explain
6. what do you understand by classification and Tabulation of data ? Discuss the modes of classification.
7. Prepare a frequency distribution from the following figures relating to bonus paid to workers

BONUS IN (Rs.)

86	62	58	73	101	90	84	90	76	61	84	63	56	88
72	92	60	83	102	76	99	54	64	87	103	61	88	55

Take a class interval of 5

8. What are the merits and demerits of diagrammic representation of the data?
9. Represent the following in a Pie chart

   Tea – 3260 tons, Cocoa – 1850 tons, Coffee – 900 tons

   Total - 6010tons
10. For the frequency distribution give below obtain Less than and More than Cumulative frequencies. Also draw Ogives for each.

L1 -L2	03-05	06-08	09-11	12-14	15-17	18-20
Frequency	5	8	11	15	7	4

2.10 REFERENCES

1. Gupta S.*P. Business Statistics*, New Delhi: S Chand and Sons Publishers, 2000
2. Shahsi Kumar. *Quantitative Techniques and methods*, Mysuru: Chetana Book House, 2010
3. Vignanesh Prajapathi, *Big data Analysis With R and Hadoop*, Mumbai: Packt Publishing, 2013
4. SD Sharma, *Operation Research,* Delhi: Discovery Publishing House, 1997
5. Srinath L. S, *PERT and CPM*, Delhi: East West Press,2001
6. Kalavathy, *Operation Research* , New Delhi: Vikas Publishing House, 2010
7. Richard I. Levin. *Statistics for Management,* New Delhi: Pearson education India, 2008

UNIT -3 : MEASURES OF CENTRAL TENDENCY

STRUCTURE

3.0 OBJECTIVES

After studying this unit you should be able to :

- Explain the measures of central tendency;
- Compute Mean , Median and Mode;
- Identify the merits and demerits of computing the Mean, Median and Mode and
- Analyze the relationship between the three measures

3.1 INTRODUCTION

In the previous unit, we have studied how to collect raw data, its classification and tabulation in a useful form, which contributes in solving many problems of statistical concern. Yet, this is not sufficient, for in practical purposes, there is need for further condensation, particularly when we want to compare two or more different distributions. We may reduce the entire distribution to one number which represents the distribution.

A single value which can be considered as typical or representative of a set of observations and around which the observations can be considered as Centered is called an 'Average' (or average value) or a Center of location. Since such typical values tends to lie centrally within a set of observations when arranged according to magnitudes, averages are called measures of central tendency.In fact the distribution have a typical value (average) about which, the observations are more or less symmetrically distributed. This is of great importance, both theoretically and practically. Dr. *A.L. Bowley correctly stated, "Statistics may rightly be called the science of averages."*The word average is commonly used in day-to-day conversations. For example, we may say that Abert is an average boy of my class; we may talk of an average American, average income, etc. When it is said, "Abert is an average student," it means is that he is neither very good nor very bad, but a mediocre student. However, in statistics the term average has a different meaning.

There is a peculiar tendency of the data to cluster or centre around a specific value. On the whole they tend to be closer to one particular value than others. This peculiar tendency of the data is called as central tendency. Thus a measure of central tendency of a set of data lies in obtaining this central value.

The fundamental measures of tendencies are:

(1) Arithmetic mean

(2) Median

(3) Mode

(4) Geometric mean

(5) Harmonic mean

(6) Weighted averages

However the most common measures of central tendencies or Locations are ***Arithmetic mean, median and mode.***

3.2 ARITHMETIC MEAN AND ITS COMPUTATION

This is the most commonly used measure of central tendency popularly called as Average or mean.

Horace Sacrist : "Arithmetic mean is the amount secured by dividing the sum of values of the items in a series by their number".

W.I. King : "The arithmetic average may be defined as the sum of aggregate of a series of items divided by their number".

Thus, the students should add all observations (values of all items) together and divide this sum by the number of observations (or items).

Ungrouped Data

Suppose, we have 'n' observations (or measures) x_1 , x_2 , x_3, ..., x_n then the Arithmetic mean is

obviously $$\frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

This method is known as the **''Direct Method".**

Example :A variable takes the values as given below. Calculate the arithmetic mean of 110, 117, 129, 195, 95, 100, 100, 175, 250 and 750.

Solution: Arithmetic mean = $\frac{\sum x_i}{n}$

$\sum x_i$ = 110 + 117 + 129 +195 + 95 +100 +100 +175 +250 + 750 = 2021

and n = 10, **therefore A M = 2021/10 = 202.1**

Indirect Method (Assumed Mean Method)

$\bar{u} = \frac{\sum u_i}{n}$ where $u = x_i - A$

A = Assumed Mean Let A = 175 then

Σu_i = -65, -58, -46, +20, -80, -75,-75, +0, + 75, +575= 670 - 399

= 271/10 = 27.1

$\therefore \quad \bar{x} = A + \bar{u}$

= 175 + 27.1= 202.1

Discrete Series :

Arithmetic mean $(\bar{x}) = \frac{\sum f_i x_i}{\sum f_i}$ i.e. $\frac{f_1 x_1 + f_2 x_2 + f_3 x_3 + \ldots\ldots + f_n x_n}{f_1 + f_2 + f_3 + \ldots\ldots + f_n}$

The formulae for Arithmetic mean by direct method and by the short-cut methods are as follows:

Direct method **Short-cut method**

$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$ $\bar{x} = A + \bar{u}$ where $\bar{u} = \frac{\sum f_i u_i}{\sum f_i}$

and $u = x_i - A$

Therefore, $\bar{x} = A + \frac{\sum f_i u_i}{\sum f_i}$

Example Find the mean of the following 50 observations.

19, 19, 20, 20, 20, 19, 20, 18, 21, 19,

20, 20, 19, 19, 20, 19, 21, 19, 19, 21,

18, 20, 18, 18, 17, 20, 20, 22, 20, 20,

20, 20, 20, 21, 20, 17, 23, 18, 17, 21,

20, 21, 20, 20, 20, 18, 21, 19, 21, 19

Solution: We may tabulate the given observations as follows.

Observations (x_i)	Frequency (f_i)	fx_i
17	3	$17 \times 3 = 51$
18	6	$18 \times 6 = 108$
19	11	$19 \times 11 = 209$
20	20	$20 \times 20 = 400$
21	8	$21 \times 8 = 168$
22	1	$22 \times 1 = 22$
23	1	$23 \times 1 = 23$
Total	$\sum f_i = 50$	$\sum f_i x_i = 981$

The arithmetic mean is $\bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{981}{50} = 19.62$

Mean for Grouped data

Continuous series: The procedure of finding the arithmetic mean in this series, is the same as we have used in the discrete series. The only difference is that in this series, we are given class-intervals, whose mid-values (class-marks) are to be calculated first.

Formula, Arithmetic mean $(\bar{x}) = \frac{\Sigma f_i x_i}{\Sigma f_i}$

where x = mid-value

Example The weights (in gms) of 30 articles are given below :

14, 16, 16, 14, 22, 13, 15, 24, 23, 14, 20, 17, 21, 18, 18, 19, 20, 17, 16, 15, 11, 22, 21, 20, 17, 18, 19, 12, 23,11.

Form a grouped frequency table, by dividing the variate range into intervals of equal width, one class being 11-13 and then compute the arithmetic mean.

Solution:

Weights	mid-values x_i	frequency f_i	Direct method $f_i x_i$	Short-cut method A = 18	
				$u_i = x_i - A$	$f_i u_i$
11 - 13	12	3	36	-6	-18
13 - 15	14	4	56	-4	-16
15 - 17	16	5	80	-2	-10
17 - 19	18	6	108	0	0
19 - 21	20	5	100	+2	+10
21 - 23	22	4	88	+4	+16
23 - 25	24	3	72	+6	+18
Total		$\Sigma f_i = 30$	$\Sigma f_i x_i = 540$		$\Sigma f_i u_i = 0$

Direct Method

Arithmetic mean

$$(\bar{x}) = \frac{\Sigma f_i x_i}{\Sigma f_i}$$

$$\bar{x} = \frac{540}{30}$$

$$\bar{x} = 18$$

Short-cut method

Arithmetic mean

$$(\bar{x}) = A + \frac{\Sigma f_i u_i}{\Sigma f_i}$$

$$\bar{x} = 18 + \frac{0}{30}$$

$$\bar{x} = 18$$

Mean weight = 18 gms

Example Find the arithmetic mean for the following :

Properties Of Arithmetic Mean

1. The sum of the deviations, of all the values of x, from their arithmetic mean, is zero.
2. The product of the arithmetic mean and the number of items gives the total of all items.
3. If $\bar{x}_1$ and $\bar{x}_2$ are the arithmetic mean of two samples of sizes n_1 and n_2 respectively then, the arithmetic mean $\bar{x}$ of the distribution combining the two can be calculated as

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

Merits of Arithmetic Mean : -

1. It is rigidly defined. Its value is always definite.
2. It is easy to calculate and easy to understand. Hence it is very popular.
3. It is based on all the observations; so that it becomes a good representative.
4. It can be easily used for comparison.
5. It is capable of further algebraic treatment such as finding the sum of the values of the observations, if the mean and the total number of the observations are given; finding the combined arithmetic mean when different groups are given etc.
6. It is not affected much by sampling fluctuations.

Demerits of Arithmetic Mean : -

1. It is affected by outliers or extreme values. In such a case A. mean is not a good representative of the given data.
2. It is a value which may not be present in the given data.
3. Many a times it gives absurd results like 4.4 children per family.
4. It is not possible to take out the averages of ratios and percentages.
5. We cannot calculate it when open-end class intervals are present in the data.

3.3 WEIGHTED ARITHMETIC MEAN AND ITS COMPUTATION

When individual observations vary in importance, they are assigned weights according to the level of importance of each in the computation of their mean. The arithmetic mean of asset of observations computed by taking into account of their corresponding weights is known as weighted arithmetic mean or average.

Weighted A M = A +(wi xi/wi)

3.4 GEOMETRIC MEAN AND ITS COMPUTATION

The geometric mean of a set of 'n' sample observation is the nth root of their product. That is

GM = $\sqrt[2]{x1\ X\ x2\ X\ \ldots..X\ Xn}$

When n is more than 2 ,

GM = antilog (∑ log Xi / n)

Weighted Geometric Mean = **GM = antilog (∑ wi log xi / ∑ wi)**

GM is particularly used in averaging ratios and percentages and rates of change in one period over the other.

3.5 HARMONIC MEAN AND ITS COMPUTATION

It is defined as the reciprocal of the arithmetic mean of the reciprocals of a given set of observations then harmonic mean is denoted as

H M = $\frac{n}{\Sigma\frac{1}{xi}}$ **Weighted HM =** $\frac{\Sigma wi}{\Sigma\frac{wi}{xi}}$

Harmonic mean is particularly useful in averaging rates and ratios. It is the appropriate average where the unit of observation such as per hour, per day etc. Remains the same and the act being performed that is covering distance is constant.

3.6 MEDIAN

It is the value of the size of the central item of the arranged data (data arranged in the ascending or the descending order). Thus, it is the value of the middle item and divides the series in to equal parts.In ***Connor's words*** - "The median is that value of the variable which divides the group into two equal parts, one part comprising all values greater and the other all values lesser than the median." For example, the daily wages of 7 workers are 5, 7, 9, 11, 12, 14 and 15 dollars. This series contains 7 terms. The fourth term i.e. $11 is the median.

Median In Individual Series (ungrouped Data)

1. Set the individual series either in the ascending (increasing) or in the descending (decreasing) order, of the size of its items or observations.
2. If the total number of observations be 'n' then

A. If 'n' is odd,The median = size of $\left(\frac{n+1}{2}\right)^{th}$ observation

B. If 'n' is even, the median

$$= \frac{1}{2}\left[\begin{array}{l}\text{size of } \left(\frac{n}{2}\right)^{th} \text{ observations} \\ + \text{size of } \left(\frac{n+2}{2}\right)^{th} \text{ observations}\end{array}\right]$$

Example The following figures represent the number of books issued at the counter of a Statistics library on 11 different days. 96, 180, 98, 75, 270, 80, 102, 100, 94, 75 and 200. Calculate the median.

Solution:

Arrange the data in the ascending order as 75, 75, 80, 94, 96, 98, 100, 102,180, 200, 270.

Now the total number of items 'n'= 11 (odd)

Therefore, the median = size of $\left(\frac{n+1}{2}\right)^{th}$ item

$$= \text{size of } \left(\frac{11+1}{2}\right)^{th} \text{ item}$$

$$= \text{size of } 6^{th} \text{ item}$$

$$= 98 \text{ books per day}$$

Example The population (in thousands) of 36 metropolitan cities are as follows :
2488, 591, 437, 20, 131, 143, 1490, 407, 384, 176, 263, 193, 181, 777, 387, 302, 213, 204, 153, 733, 391, 176 178, 142, 522, 360, 65, 260, 193, 92, 672, 258, 239, 160, 147, 151. Calculate the median.

Solution:

Arranging the terms in the ascending order as :

20, 65, 92, 131, 142, 143, 147, 151, 153, 160, 169, 176, 178, 181, 193, 204, (213, 239), 258, 263, 260, 384, 302, 360, 387, 391, 407, 437, 522, 591, 672, 733, 777, 1490, 2488.

Since total number of items n = 36 (Even).

the median

$$= \frac{1}{2}\left[\text{size of } \left(\frac{n}{2}\right)^{th} \text{ item} + \text{size of} \left(\frac{n+2}{2}\right)^{th} \text{ item}\right]$$

$$= \frac{1}{2}\left[\text{size of } 18^{th} \text{ item} + \text{size of } 19^{th} \text{ item}\right]$$

$$= \frac{1}{2}\left[213+239\right]$$

$$= 276 \text{ Thousands}$$

Median In Discrete Series :Steps :

1. Arrange the cumulative frequencies.
2. Find the cumulative frequencies.
3. Apply the formula :

A. If 'n' = Σf_i (odd) then,

$$\text{Median} = \text{size of } \left(\frac{n+1}{2}\right)^{th} \text{ item}$$

B. If 'n' = Σf_i (even) then,

$$\text{Median} = \frac{1}{2}\left[\text{size of } (n/2)^{th} \text{ item} + \text{size of} \left(\frac{n+2}{2}\right)^{th} \text{ item}\right]$$

Example Locate the median in the following distribution.

Size	:	8	10	12	14	16	18	20
Frequency	:	7	7	12	28	10	9	6

Solution:

Size (x_i)	Frequency f_i	Cumulative frequency (c.f)
8	3	3
10	7	3 + 7 = 10
12	12	12 + 10 = 22
14	28	28 + 22 = 50
16	10	50 + 10 = 60
18	9	60 + 9= 69
20	6	69 + 6 = 75
	$n = \Sigma f_i = 75$ (odd)	

Therefore, the median = $\text{size of } \left(\frac{n+1}{2}\right)^{th} \text{ item}$

$= \text{size of } \left(\frac{75+1}{2}\right)^{th} \text{ item} = \text{size of } 38^{th} \text{ item}$

In the order of the cumulative frequency, the 38th term is present in the 50th cumulative frequency, whose size is 14.

Therefore, the median = 14

Median In Continuous Series (grouped Data)

Steps :

1. Determine the particular class in which the value of the median lies. Use $n/2$ as the rank of the median and not $\left(\frac{n+1}{2}\right)$

2. After ascertaining the class in which median lies, the following formula is used for determining the exact value of the median.

Median = $\ell_1 + \left[\frac{N/2 - c.f}{f}\right](\ell_2 - \ell_1)$

where, ℓ_1 = lower limit of the median class, the class in which the middle item of the distribution lies.

ℓ_2 = upper limit of the median classc.f = cumulative frequency of the class preceding the median classf = sample frequency of the median class

It should be noted that while interpolating the median value of frequency distribution it is assumed that the variable is continuous and that there is an orderly and even distribution of items within each class.

Example Calculate the median for the following and verify it graphically.

Age (years) : 20-25 25-30 30-35 35-40 40-45

No. of person : 70 80 180 150 20

Solution:

Age (years) C.I.	No. of persons f_i	Cumulative frequency c.f.
20 - 25	70	70
25 - 30	80	70 + 80 = 150
30 - 35	180	150 + 180 = 330
35 - 40	150	330 + 150 = 480
40 - 45	20	480 + 20 = 500
	$n = \Sigma f_i = 500$	

Now median = size of $(n/2)^{th}$ item

$$= \text{size of } \left(\frac{500}{2}\right)^{th}$$

$$= \text{size of } 250^{th} \text{ item which}$$

lies in $(30 - 35)$ class interval

$$\text{Median} = \ell_1 + \left[\frac{N/2 - \text{c.f}}{f}\right](\ell_2 - \ell_1)$$

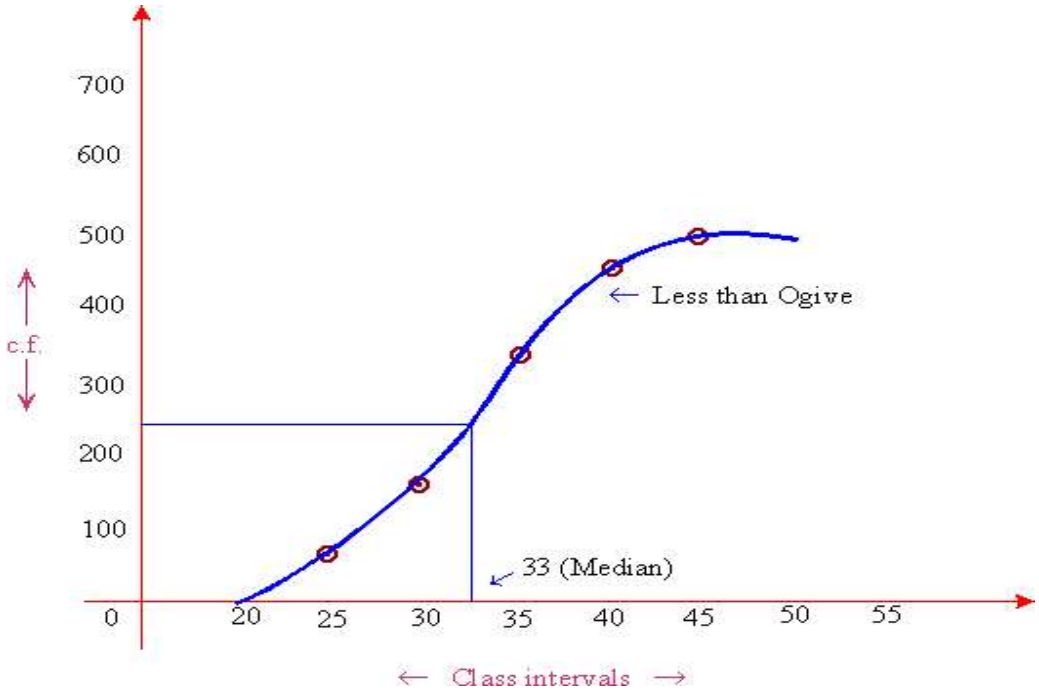
Here $\ell_1 = 30$, $\ell_2 = 35$, $n/2 = 250$, c.f. = 150 and f = 180

Therefore, Median

$$= 30 + \left[\frac{250 - 150}{180}\right](35 - 30)$$

$$= 30 + \frac{100}{180} \times 5$$

$$= 32.78 \text{ years}$$



Note that, while calculating the median of a series, it must be put in the 'exclusive class-interval' form. If the original series is in inclusive type, first convert it into the exclusive type and then find its median.

Merits Of Median

1. It is rigidly defined.
2. And it is easy to calculate and understand.
3. It is not affected by extreme values like the arithmetic mean.
4. It can be found by mere inspection.
5. It is fully representative and can be computed easily.
6. It can be used for qualitative studies.
7. Even if the extreme values are unknown, median can be calculated if one knows the number of items.
8. It can be obtained graphically.

Demerits of Median

1. It may not be representative if the distribution is irregular and abnormal.
2. It is not capable of further algebraic treatment.
3. It is not based on all observations.
4. It is affected by sample fluctuations.

5. The arrangement of the data in the order of magnitude is absolutely necessary.

3.7 MODE

It is the size of that item which possesses the maximum frequency. According to Professor Kenney and Keeping, the value of the variable which occurs most frequently in a distribution is called the mode. It is the most common value. It is the point of maximum density.

Ungrouped Data

Individual series: The mode of this series can be obtained by mere inspection. The number which occurs most often is the mode.

Example Locate mode in the data 7, 12, 8, 5, 9, 6, 10, 9, 4, 9, 9

Solution : On inspection, it is observed that the number 9 has maximum frequency. Therefore 9 is the mode.

Grouped Data: **Steps :**

1. Determine the modal class which as the maximum frequency.
2. By interpolation the value of the mode can be calculated as -

$$\text{Mode} = \ell_1 + \left[\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right](\ell_2 - \ell_1)$$

where

ℓ_1 = lower limit of the modal class

ℓ_2 = upper limit of the modal class

f_1 = frequency of the modal class

f_0 = frequency of the class preceding to the modal class

f_2 = frequency of the class succeding the modal class

Example Calculate the modal wages.

Daily wages in $:	20 -25	25-30	30-35	35-40	40-45	45-50
No. of workers :	1	3	8	12	7	5

Verify it graphically.

Solution:

Here the maximum frequency is 12, corresponding to the class interval (35 - 40) which is the modal class.

Therefore $\ell_1 = 35,\ \ell_2 = 40,\ f_0 = 8,\ f_1 = 12,\ f_2 = 7$

By interpolation

$$\text{Mode} = \ell_1 + \left[\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right](\ell_2 - \ell_1)$$

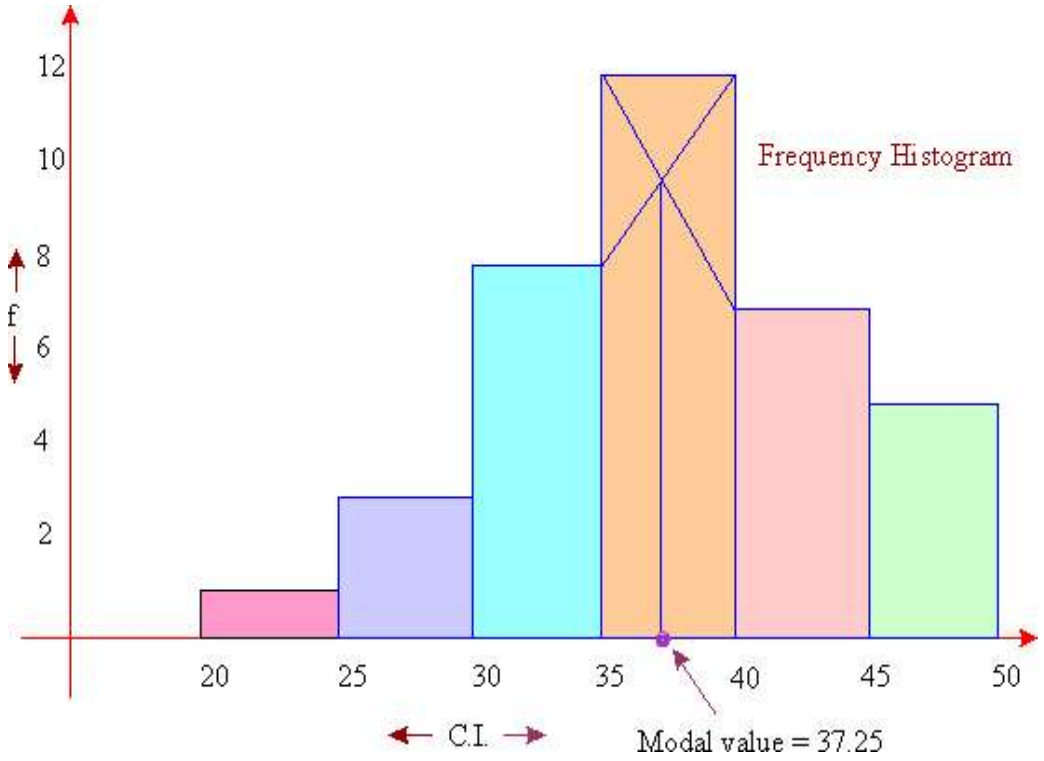
$$= 35 + \left[\frac{12-8}{24-8-7}\right](40-35)$$

$$= 35 + \left[\frac{4}{9}\right] \times 5$$

$$= 35 + 2.22$$

$$= 37.22$$

Modal wages is $37.22



Merits of mode

1. It is simple to calculate.
2. In individual or discrete distribution it can be located by mere inspection.
3. It is easy to understand. Everyone is used to the idea of average size of a garment, an average American etc.
4. It is not isolated like the median as it is the most common item.
5. Like the Average mean, it is not a value which cannot be found in the series.
6. It is not necessary to know all the items. What we need the point of maximum density frequency.
7. It is not affected by sampling fluctuations.

Demerits

1. It is ill defined.
2. It is not based on all observations.
3. It is not capable of further algebraic treatment.
4. It is not a good representative of the data.
5. Sometimes there are more than one values of mode.

3.8 RELATIONSHIP BETWEEN MEAN, MEDIAN AND MODE

In the symmetrical distribution Mean, Median and Mode have the same value. The relationship is **AM = GM = HM.**

But usually the frequency distribution tends to deviate and get skewed. Then this will affect the mean than median and mode and the following situation will arise.

Positively skewed : the distribution skewed to the right. As a result mean gains highest value, followed by descending order of Median and Mode

Negatively Skewed: the distribution skewed to the left. As a result value of mean tends to be lowest followed in ascending order by median and mode.

All the three are empirically related as:

Mean – Mode = 3(Mean – Median)

Mode = mean - 3(Mean – Median)

Mean - Median = 1/3(Mean – Mode)

Mode = 3 Median - 2 Mean

Comparison of the three measures:

1. Mean is the most familiar and widely used measure of central tendency as it takes into account all observations in its computation. The presence of extreme values affect Mean more than the Median and the Mode. Mean is used more in symmetric distributions.
2. Median is easier to understand and compute when the data is relatively small. The extreme values do not affect median more as such as mean and therefore it is frequently used as a best measure of central tendency in asymmetric distributions.
3. Mode is the least used measure of central tendency. It is very easy to compute. It can be used for both quantitative and qualitative data. A little care should be taken in computing Mode because, every distribution may not have mode and there may be two modes present in one distribution.

3.9 SUMMARY

Measures of central tendency are the basics of statistics. It is an attempt to find out the central value of a given set of data. The idea behind determining such a typical value is to use it as representative of the entire data. There are three measures of central tendency: Mean, Median and Mode. The other measures are Geometric Mean and harmonic Mean. Mean is also called as Average. Median is the location average of the middle value of an ordered array of set of observations. Mode is also a location average and it is that value which appears the maximum number of times.

3.10 KEY WORDS

Weighted Average	-	A mean or average value calculated to take into account the importance of each value to the overall total
Bimodal Distribution	-	A distribution of 2 observations occurring more frequently than the others in a set of values.
Mean	–	the Arithmetic Average of the given set of observations

Median Class – A frequency distribution class interval denoting the median value of the observations

3.11 SELF ASSESSMENT QUESTIONS

1. What are the various Measures of Central Tendency? Explain each in detail.
2. Calculate the average value of age for a class of 10 students with their ages as under 11,12, 13, 13, 10, 13, 12, 11, 10, 12.
3. From the following calculate the average level of marks of the class.

Marks:	0	2	3	4	5	6	7	8	9
Number of students:	11	10	9	21	12	17	8	22	15

4. Given below is the distribution of marks obtained by 60 students in final exams. Compute a. Mean, b. Median, c. Mode

Marks:	20	30	40	50	60	70
Number of students:	8	12	20	10	6	4

5. From the frequency distribution given below find Mean, Median and Mode

Class intervals :	50-52	53-55	56-58	59-61	62-64
Frequencies:	5	10	21	8	6

6. The average sales of a product for a particular week excluding Sunday wre 150units. Sunday there was a rush of sales which inflated the Average sales for the entire week to 210 units. Find the sales for Sunday.
7. Find the GM of 5 sample observations: 28, 45, 50, 65, and 90.
8. Obtain the HM of 5 samples: 4, 20, 12, 10 and 15
9. Calculate the Arithmetic Mean by Step Deviation Method for the following data:

Class – Intervals	Mid Points (X)	Frequency (f)
0-10	5	7
10-20	15	9
20-30	25	15
30-40	35	11
40-50	45	27
50-60	55	18
60-70	65	5

10. Given the following distribution calculate Mean Median and Mode and also show their empirical relationship.

Pay Scale (Rs)	No. of employees
Less than 2000	14
Less than 3000	19
Less than 4000	26
Less than 5000	35
5000 and above	42

3.12 REFERENCES

1. Gupta S.*P. Business Statistics*, New Delhi: S Chand and Sons Publishers, 2000
2. Shahsi Kumar. *Quantitative Techniques and methods*, Mysuru: Chetana Book House, 2010
3. Vignanesh Prajapathi, *Big data Analysis With R and Hadoop*, Mumbai: Packt Publishing, 2013
4. SD Sharma, *Operation Research,* Delhi: Discovery Publishing House, 1997
5. Srinath L. S, *PERT and CPM*, Delhi: East West Press,2001
6. Kalavathy, *Operation Research* , New Delhi: Vikas Publishing House, 2010
7. Richard I. Levin. *Statistics for Management,* New Delhi: Pearson education India, 2008

UNIT 4 : MEASURES OF DISPERSION

STRUCTURE

4.0 OBJECTIVES

After studying this unit you should be able to :

- Explain measures of Dispersion ;
- Analyse compute each measure of dispersion and
- Analyse the relationship and significance of each of the measures

4.1 INTRODUCTION

The measures of central tendencies i.e. means indicate the general magnitude of the data and locate only the center of a distribution of measures. They do not establish the degree of variability or the spread out or scatter of the individual items and their deviation from (or the difference with) the mean.

i) **According to Nciswanger,** "Two distributions of statistical data may be symmetrical and have common means, medians and modes and identical frequencies in the modal class. Yet with these points in common they may differ widely in the scatter or in their values about the measures of central tendencies."

ii) **Simpson and Kafka said,** "An average alone does not tell the full story. It is hardly fully representative of a mass, unless we know the manner in which the individual item. Scatter around it ....a further description of a series is necessary, if we are to gauge how representative the average is."

From this discussion we now focus our attention on the scatter or variability which is known as **dispersion**. Let us take the following three sets.

Students	Group X	Group Y	Group Z
1	50	45	30
2	50	50	45
3	50	55	75
Mean- $\bar{x}$	50	50	50

Thus, the three groups have same mean i.e. 50. In fact the median of group X and Y are also equal. Now if one would say that the students from the three groups are of equal capabilities, it is totally a wrong conclusion then. Close examination reveals that in group X students have equal marks as the mean, students from group Y are very close to the mean but in the third group Z, the marks are widely scattered. It is thus clear that the measures of the central tendency is alone not sufficient to describe the data.

Definition of dispersion : The arithmetic mean of the deviations of the values of the individual items from the measure of a particular central tendency used. Thus the 'dispersion' is also known as the "**average of the second degree.**" **Prof. Griffin and Dr. Bowley** said the same about the dispersion.

In simple terms Dispersion is the variability among individual observations comprising a set of data. It describes the spread characteristics of the data. A measure of dispersion lies in quantifying the variability among individual observations and their scatter around the central value.

Characteristics of ideal measure of dispersion:

1. It should be rigidly defined
2. It should be easy to calculate
3. It should be based on all the observations
4. It should be amenable for further mathematical treatment
5. It should be affected as little as possible by fluctuations of sampling and by extreme observations

Methods of Computing Dispersion: the various measures of dispersions are

1. Range
2. Quartile Deviation
3. Mean Deviation
4. Variance
5. Standard Deviation

In measuring dispersion, it is imperative to know the amount of variation (absolute measure) and the degree of variation (relative measure). In the former case we consider the range, mean deviation, standard deviation etc. In the latter case we consider the coefficient of range, the coefficient mean deviation, the coefficient of variation etc.

(I) Method of limits:

(1) The range (2) Inter-quatrile range (3) Percentile range

(II) Method of Averages:

(1) Quartile deviation (2) Mean deviation

(3) Standard Deviation and (4) Other measures.

4.2 RANGE

In any statistical series, the difference between the largest and the smallest values is called as the range.

Coefficient of Range : The relative measure of the range. It is used in the comparative study of the dispersion co-efficient of Range =

$$\text{Thus Range (R)} = L - S \begin{cases} L = \text{Largest value of the series.} \\ S = \text{Smallest value of the series.} \end{cases}$$

Example (Individual series) Find the range and the co-efficient of the range of the following items : $\frac{L-S}{L+S}$

110, 117, 129, 197, 190, 100, 100, 178, 255, 790.

Solution: R = L - S = 790 - 100 = 690

Example (Continuous series) Find the range and its co-efficient from the following data.

$$\text{Co-efficient of Range} = \frac{L-S}{L+S} = \frac{790-100}{790+100} = \frac{690}{890} = 0.78$$

Size:	10 - 20	20 - 30	30 - 40	40 - 50	50 - 10
Frequency :	2	3	5	4	2

Solution: R = L - S = 100 - 10 = 90

Co-efficient of range = $\frac{L-S}{L+S} = \frac{100-10}{100+10} = \frac{90}{110} = 0.82$

Merits and Demerits of range:

It is the simplest but crude method of dispersion. It is rigidly defined and readily comprehensible and easiest to compute. But it is not based on all the observations and cannot be used for further mathematical treatment. It is based on only two extreme values and not on entire set of data. It is affected by the fluctuations of sampling. Range cannot be used if we are dealing with open end class. It is very sensitive to the size of the sample. Therefore Range is regarded as too indefinite to be used as a practical measure of dispersion.

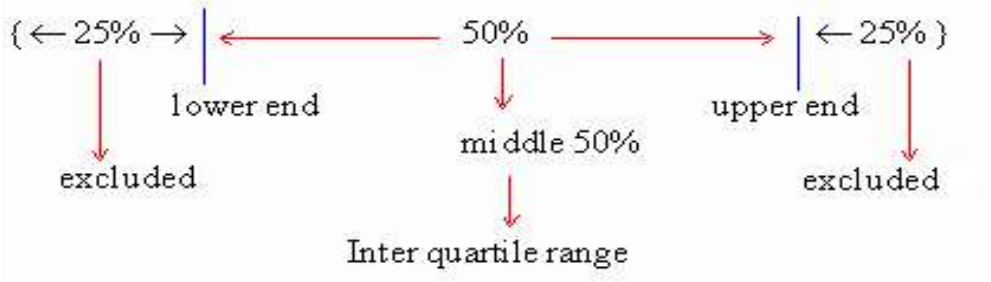
4.3 QUARTILE DEVIATION AND COMPUTATIONS

Quartiles and Interquartile Range

It is the measure of dispersion based on the upper quartile Q3 and Lower Quartile Q1. Quartile deviation is obtained from interquartile range on dividing by 2 . Hence it is also called as Semi Inter Quartile Range. Therefore Q. D. (SI QR) =

$$\frac{Q_3 - Q_1}{2}$$

If we concentrate on two extreme values (as in the case of range), we don't get any idea about the scatter of the data within the range (i.e. the two extreme values). If we discard these two values the limited range thus available might be more informative. For this reason the concept of interquartile range is developed. It is the range which includes middle 50% of the distribution. Here 1/4 (one quarter of the lower end and 1/4 (one quarter) of the upper end of the observations are excluded.



Now the lower quartile (Q_1) is the 25th percentile and the upper quartile (Q_3) is the 75th percentile. It is interesting to note that the 50th percentile is the middle quartile (Q_2) which is in fact what you have studied under the title ' Median ". Thus symbolically

Inter quartile range = $Q_3 - Q_1$

If we divide ($Q_3 - Q_1$) by 2 we get what is known as Semi-Iinter quartile range.

i.e. $\frac{Q_3 - Q_1}{2}$. It is known as Quartile deviation (Q. D or SI QR).

Therefore Q. D. (SI QR) = $\frac{Q_3 - Q_1}{2}$

Coefficient of QD = ($Q_3 - Q_1$***) /(***$Q_3 + Q_1$***)***

Example: Find the following from the distribution.

1. Interquartile range
2. Quartile deviations
3. Coefficient of Q D

Class Interval	F	Less than C F
0-15	8	8
15-30	26	34
30-45	30	64
45-60	45	109
60-75	20	129
75-90	17	146
90-105	4	150
Total	N =150	

$Q_1 =$ N/4 = 150/4 = 37.5.

The CF greater than 37.5 is 64. Therefore Q_1 lies in corresponding class of 30-45

$Q_1 = l + \frac{h}{f}$ (N/4 – C)

$Q_1 = 30+15/30(37.5- 34) = 31.75$

$Q_3 = 3N/4 = 3(150)/4 = 112.5$

The CF greater than 112.5 is 129. Therefore Q_3 lies in corresponding class of 60-75

$$Q_3 = l + \frac{h}{f} (3N/4 - C)$$

$Q_3 = 60 + 15/20 (112.5- 109) = 62.625$

a. Inter quartile Range = $Q_3 - Q_1 = 62.625 – 31.75 = 30.875$
b. Quartile deviation =$(Q_3 - Q_1)/ 2 = 30.875/2 =15.44$
c. Coefficient of QD = $(Q_3 - Q_1) / (Q_3 + Q_1) = (62.625 – 31.75) / (62.625 + 31.75) = 0.33$

Merits and Demerits of QD:

It easy to understand and calculate. It used 50% of the data and thus a better measure than range. It is not affected by extreme values as it excludes 25% of the data from the beginning and 25% from the top. It can also be calculated from the open end class and it is the only measure of dispersion to deal with open end class.

It is not based on all observations since it ignores 25% in the beginning and 25% in the end. It gets affected by fluctuations of sampling and not suitable for further mathematical treatment.

4.4 MEAN DEVIATION

Average deviations (mean deviation) is the average amount of variations (scatter) of the items in a distribution from either the mean or the median or the mode, ignoring the signs of these deviations.

Individual SeriesSteps to calculate MD :

1. Find the mean or AM of the distribution by usual methods.
2. Take the deviation d=X-A of each observation from the average.
3. Ignore the negative signs of the deviations taking all the deviations to be positive to obtain the absolute deviations l d l = l X - A l
4. Obtain the sum of the absolute deviations obtained in step 3.
5. Divide the total obtained in step 4 by **n.** (the number of observations).

The result gives the value of mean deviation about the average A.

In case of frequency distribution MD is obtained as :

M D (about the Average A) $= \frac{1}{N}\sum f\,(d)$

M D (about Mean) $= \frac{1}{N}\sum f\,(X - M)$

M D (about Median) $= \frac{1}{N}\sum f\,(X - Md)$

M D (about Mode) $= \frac{1}{N}\sum f\,(X - Mo)$

Example (Continuous series) Calculate the mean deviation and the coefficient of mean deviation from the following data using the mean.Difference in ages between boys and girls of a class.

Diff. in years:	No.of students:
0 - 5	449
5 - 10	705
10 - 15	507
15 - 20	281
20 - 25	109
25 - 30	52
30 - 35	16
35 - 40	4

Solution :

Diff. in age	Mid-values (x_i)	frequency (f_i)	$f_i x_i$	$\lvert x_i - \bar{x} \rvert$	$f_i \lvert x_i - \bar{x} \rvert$
0 - 5	2.5	449	1122.5	8	3592
5 - 10	7.5	705	5287.5	3	2115
10 - 15	12.5	507	6337.5	2	1014
15 - 20	17.5	281	4917.5	7	1967
20 - 25	22.5	109	2452.5	12	1308
25 - 30	27.5	52	1430.0	17	884
30 - 35	32.5	16	520.0	22	352
35 - 40	37.5	4	150.0	27	108
		n = 2123	$\Sigma f_i x_i$ = 22217.5		$\Sigma f_i \lvert x_i - \bar{x} \rvert$ = 11440

Calculation:

1) $\bar{X} = \frac{\Sigma fi\, xi}{n} = \frac{22217.5}{2123} = 10.5$ (approx.)

2) M. D. $= \frac{\Sigma f_i \left| x_i - \bar{x} \right|}{n} = \frac{11440}{2123} = 5.4$

3) co –efficient of M. D. $= \frac{M.D.}{\bar{x}} = \frac{5.4}{10.5} = 0.514$

Merits and demerits of Mean deviations:

M D is rigidly defined and is easy to understand and calculate. It is based on all the observations. MD removes the irregularities in the distribution and provides accurate and true measure of dispersion. It is less affected by extreme observations.

The major demerit is we take the absolute values and neglect the signs of the deviations which mathematically unsound and illogical. This makes it useless for further mathematical treatment.it is not satisfactory measure when taken about Mode or Median.it cannot be computed with open end class. And it is tend to increase in size as the size of the sample increases.

4.5 VARIANCE

The term variance was used to describe the square of the standard deviation R.A. Fisher in 1913. The concept of variance is of great importance in advanced work where it is possible to split the total into several parts, each attributable to one of the factors causing variations in their original series. Variance is defined as follows:

$$\text{Variance} = \frac{\Sigma(x_i - \bar{x})^2}{n}$$

4.6 STANDARD DEVIATION (S. D.)

It is the square root of the arithmetic mean of the square deviations of various values from their arithmetic mean. it is denoted by s.d. or σ.

$$\text{Thus, s.d.}(\sigma x) = \sqrt{\frac{\Sigma\left(x_i - \bar{x}\right)^2}{n}} \quad \text{for the ungrouped data}$$

$$= \sqrt{\frac{\Sigma f_i\left(x_i - \bar{x}\right)^2}{n}} \quad \text{for the grouped data}$$

where n = Σ fi

Merits :

a. It is rigidly defined and based on all observations.
b. It is amenable to further algebraic treatment.
c. It is not affected by sampling fluctuations.
d. It is less erratic.
e. It is the most widely used measure of dispersion

Demerits :

a. It is difficult to understand and calculate.
b. It gives greater weight to extreme values.

Note that **variance V(x)** = $\frac{\Sigma\left(x_i - \bar{x}\right)^2}{n}$ and

and**s. d. (σ x)** = $\sqrt{\frac{\Sigma\left(x_i - \bar{x}\right)^2}{n}}$ and $\sqrt{\frac{\Sigma f_i\left(x_i - \bar{x}\right)^2}{n}}$

Then **V (x)** = σ_x^2

4.7 CO-EFFICIENT OF VARIATION (C. V.)

To compare the variations (dispersion) of two different series, relative measures of standard deviation must be calculated. This is known as co-efficient of variation or the co-efficient of s. d. Its formula is CV = (S.D / Mean)100

$$\frac{\sigma x}{\bar{x}} \times 100$$

Thus it is defined as the ratio s. d. to its mean.

Remark: It is given as a percentage and is used to compare the consistency or variability of two more series. The higher the C. V. , the higher the variability and lower the C. V., the higher is the consistency of the data.

Example Calculate the standard deviation and its co-efficient from the following data.

A	B	C	D	E	F	G	H	I	J
10	12	16	8	25	30	14	11	13	11

Solution :

No	x_i	$(x_i - x)$	$(x_i - x)^2$
A	10	-5	25
B	12	-3	9
C	16	+1	1
D	8	-7	49
E	25	+10	100
F	30	+15	225
G	14	-1	1
H	11	-5	16
I	13	-2	4
J	11	-4	16
n= 10	xi = 150		|xi - x |2= 446

Calculations :

i) $$\bar{x} = \frac{\Sigma x_i}{n} = \frac{150}{10} = 15$$

ii) $$s.d.(\sigma x) = \sqrt{\frac{\Sigma\left(x_i - \bar{x}\right)^2}{n}} = \sqrt{\frac{446}{10}} = 6.7$$

iii) $$co-efficient\,of\,s.d. = \frac{\sigma x}{\bar{x}} = \frac{6.7}{15} = 0.45$$

Example Calculate s.d. of the marks of 100 students.

Marks	No. of students (f_i)	Mid-values (x_i)	$f_i x_i$	$f_i x_i^2$
0-2	10	1	10	10
2-4	20	3	60	180
4-6	35	5	175	875
6-8	30	7	210	1470
8-10	5	9	45	405
	n = 100		$\Sigma f_i x_i$ = 500	$\Sigma f_i x_i^2$ = 2940

Solution

1) $$\bar{x} = \frac{\Sigma f_i x_i}{n} = \frac{500}{1000} = 5$$

2) $$\text{s.d.}(\sigma x) = \sqrt{\frac{\Sigma f_i x_i^2}{n} - (xy)^2}$$

$$= \sqrt{\frac{2940}{100} - (5)^2} = \sqrt{29.40 - 25} = \sqrt{4.40}$$

$$x \quad 2.21$$

Combined Standard deviation : If two sets containing n_1 and n_2 items having means x_1 and x_2 and standard deviations σ_1 and σ_2 respectively are taken together then,

(1) Mean of the combined data is $\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$

(2) s.d. of the combined set is $\sigma = \sqrt{\frac{n_1\left(\sigma_1^2 + d_1^2\right) + n_2\left(\sigma_2^2 + d_2^2\right)}{n_1 + n_2}}$

Where $d_1 = \bar{x}_1 - \bar{x}$ and $d_2 = \bar{x}_2 - \bar{x}$

Example The score of two teams A and B in 10 matches are as:

A	40	32	0	40	30	7	13	25	14	5
B	21	14	29	13	5	12	10	13	30	0

Find the variance for both the series. Which team is more consistent ?

A			B		
x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$
40	19	361	21	4	16
32	11	121	14	-3	9
0	-21	441	14	12	144
40	19	361	30	13	169
30	9	81	5	-12	144
7	-8	196	12	-5	25
13	-8	64	10	-7	49
25	4	16	13	-4	16
14	-7	49	30	13	169
9	-12	144	6	-11	121
$\Sigma x_i = 210$		$\Sigma(x_i - \bar{x})^2 = 1834$	$\Sigma x_i = 110$		$\Sigma(x_i - \bar{y})^2 = 862$

Solution

1) $\bar{x} = \frac{\Sigma xi}{n} = \frac{210}{10}$ and $\bar{y} = \frac{\Sigma yi}{n} = \frac{170}{10} = 17$

$\therefore \sigma x = \sqrt{\frac{\Sigma(xi - \bar{x})^2}{n}} = \sqrt{\frac{1834}{10}} = \sqrt{183.4} = 13.54$

Also $\sigma y = \sqrt{\frac{\Sigma(yi - \bar{y})^2}{n}} = \sqrt{\frac{862}{10}} = \sqrt{86.2} = 9.28$

Therefore, $(C.V.)_A = \frac{13.54}{21} \times 100 = 64.47\%$

and $(C.V.)_B = \frac{9.28}{17} \times 100 = 54.69\%$

Since $(C.V.)_A < (C.V.)_B$

$\therefore$ The base ball team A is more consistent.

4.8 SKEWNESS

We study Skewness to have an idea about the shape of the curves which we can draw with the help of the given frequency distribution. It helps us to understand the nature of the concentration of observations towards higher and lower values of the variable. A distribution is said to be skewed if :

1. The frequency curve of the distribution is not a symmetric bell shaped curve but it is stretched more to one side than the other. If it has a longer tail towards the right it is said to be positively skewed. And if the tail is longer towards the left then it is negatively skewed.
2. The values of Mean, median and Mode fall at different points.
3. Quartiles Q_1 and Q_3 are not equidistant from the median.

Measures of Skewness:

1. Sk = Mean – Median
2. Sk = Mean – Mode
3. Sk = $(Q_3 – Md) – (Md - Q_1) = Q_3 + Q_1 – 2Md$

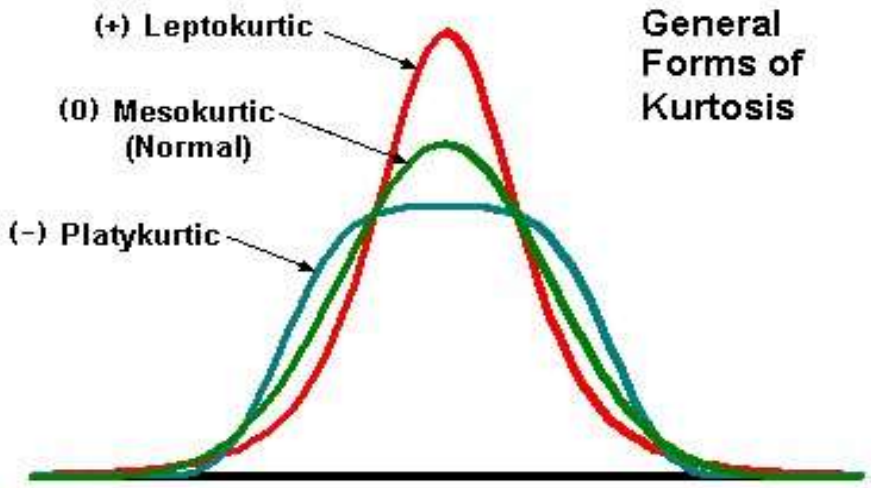
Karl Pearson's Coefficient of Skewness:

Sk = (Mean – Mode) / S.D

Or where the Mode is ill defined then

Sk = 3(Mean – Mode) / S.D

4.9 KURTOSIS



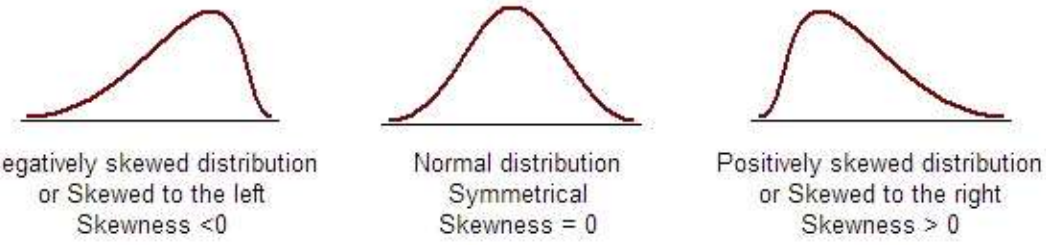
To know more about the distribution variability, **Prof. Karl Pearson** called it as Convexity of the curve or the Kurtosis. Kurtosis enables us to have an idea about the shape and nature of the hump (middle Part) of a frequency distribution.Therefore Kurtosis is concerned with the flatness or peachiness of the frequency curve. The normal curve is called as Mesokurtic. The curves which are more peaked than the normal curve are called as Leptokurtic and lack kurtosis and have negative Kurtosis. The curves which are flatter than the normal curve are platykurtic curves and have kurtosis in excess and called as positive kurtosis.

As a measure of Kurtosis Karl Pearson described coefficient β_2 as

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

Skewness

The coefficient of Skewness is a measure for the degree of symmetry in the variable distribution.



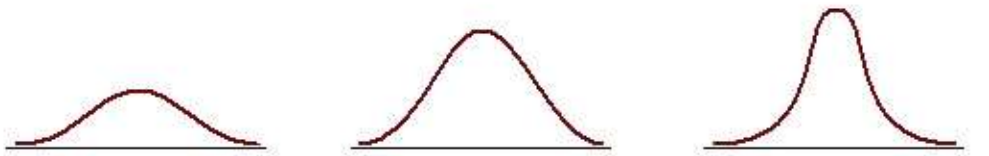
Negatively skewed distribution
or Skewed to the left
Skewness <0

Normal distribution
Symmetrical
Skewness = 0

Positively skewed distribution
or Skewed to the right
Skewness > 0

Kurtosis

The coefficient of Kurtosis is a measure for the degree of peakedness/flatness in the variable distribution.



Platykurtic distribution
Low degree of peakedness
Kurtosis <0

Normal distribution
Mesokurtic distribution
Kurtosis = 0

Leptokurtic distribution
High degree of peakedness
Kurtosis > 0

4.10 SUMMARY

Choice about various measure discussed above is based on their merits and demerits. Range is simplest of all but it is based on two extreme values. Quartile deviation is also not adequately representative as it uses only 50% of the data and it suits well with openend classes. The Variance and Standard Deviations are two most objective measures of dispersion as they cover all the set of observations of the data. SD is the a widely used measure of variability.

4.11 KEY WORDS

Dispersion	–	Variability among the observations made or deviations from the expected one.
MAD	–	Mean Absolute Deviation
Skewness	–	Lack of symmetry
Percentile range	-	this is a measure of dispersion based on the difference between certain percentiles.
Lorenz Curve	–	it is a graphic measure of studying the dispersion. This curve is used in business to study the disparities of the distribution of wages, profits, turnover, production, population etc.

4.12 SELF-ASSESSMENT QUESTIONS

1. Explain the validity of the statement "An Average when published should be accompanied by a measure of dispersion for significant interpretation".
2. What is dispersion? Explain each measure in detail.
3. The Standard Deviation is a best measure of dispersion.' Why?
4. Standard Deviation can never be negative – comment
5. Differentiate SD and MD
6. Calculate the mean deviation from the following:

X:	5	15	25	35	45	55	65
f:	8	12	10	8	3	2	7

7. Find the Median and Mean deviation from the following data;

size	Frequency
0-10	7
10-20	12
20-30	18
30-40	25
40-50	16
50-60	14
60-70	8

8. Find mean deviation from Mean and Median for the following:

Score	No. of Students
140-150	4
150-160	6
160-170	10
170-180	10
180-190	9
190-200	3

9. Find out Mean and Standard Deviation from the following:

Age:	10	20	30	40	50	60	70	80
Death :	15	30	53	75	100	110	115	125

10. Explain relative measure of dispersions:
11. The coefficient of variance is 60% and the SD is 12. Find its Mean.
12. During 10 weeks of a session, the marks are as follows.

Ramesh:	58	59	60	54	65	66	52	75	69	52
Suresh:	87	89	78	71	73	84	65	66	56	46

a. Who is better scorer?
b. Who is better consistent?

13. Write short notes on
    a. Skewness
    b. Kurtosis
    c. Percentile
    d. Lorenz Curve

4.13 REFERENCES

1. Gupta S.*P. Business Statistics*, New Delhi: S Chand and Sons Publishers, 2000
2. Shahsi Kumar. *Quantitative Techniques and methods*, Mysuru: Chetana Book House, 2010
3. Vignanesh Prajapathi, *Big data Analysis With R and Hadoop*, Mumbai: Packt Publishing, 2013
4. SD Sharma, *Operation Research,* Delhi: Discovery Publishing House, 1997
5. Srinath L. S, *PERT and CPM*, Delhi: East West Press,2001
6. Kalavathy, *Operation Research* , New Delhi: Vikas Publishing House, 2010
7. Richard I. Levin. *Statistics for Management,* New Delhi: Pearson education India, 2008