BECC-107

STATISTICAL METHODS FOR ECONOMICS

School of Social Sciences
Indira Gandhi National Open University
Maidan Garhi, New Delhi-110068

cs and
e

EXPERT COMMITTEE

COURSE PREPARATION TEAM

Block/ Unit Title		**Unit Writer**
Block 1	**Descriptive Statistics**	
Unit 1	Basic Statistical Concepts and Data Collection Methods	Adapted from IGNOU study material of EEC 13: Elementary Statistical Methods and Survey Techniques, Units 1, 2 and 3 written by Avatar Singh with modifications by K. Barik
Unit 2	Tabulation and Graphical Representation of Data	
Unit 3	Summarisation of Univariate Data	Adapted from IGNOU study material of EEC 13: Elementary Statistical Methods and Survey Techniques, Units 4, 5 and 6 written by R S Bharadwaj with modifications by K. Barik
Unit 4	Measures of skewness and kurtosis	
Block 2	**Summarisation of Bivariate and Multivariate Data**	
Unit 5	Correlation and Regression	Kaustuva Barik, Indira Gandhi National Open University
Unit 6	Index numbers	Adapted from IGNOU study material of EEC 13: Elementary Statistical Methods and Survey Techniques, Unit 10 written by J Roy with modifications by K. Barik
Unit 7	Deterministic Time Series and Forecasting	Adapted from IGNOU study material of EEC 13: Elementary Statistical Methods and Survey Techniques, Unit 11 written by S Bandopadhyay with modifications by K. Barik
Unit 8	Vital Statistics	Adapted from IGNOU study material of EEC 13: Elementary Statistical Methods and Survey Techniques, Unit 12 written by C G Naidu with modifications by K. Barik
Block 3	**Probability Theory**	
Unit 9	Elementary Probability	Sri R A Chaudhury (Retd.), Kirorimal College, University of Delhi
Unit 10	Discrete Probability Distributions	Dr. Anup Chatterjee (Retd.), ARSD College, University of Delhi
Unit 11	Continuous Probability Distributions	
Block 4	**Sampling and Statistical Inference**	
Unit 12	Sampling Procedure	Prof. C G Naidu (Retd.), Indira Gandhi National Open University
Unit 13	Statistical Estimation	Kaustuva Barik, Indira Gandhi National Open University
Unit 14	Testing of Hypotheses	
Unit 15	Chi-squared Test	

Course Coordinator: Prof. Kaustuva Barik **Editor:** Prof. Kaustuva Barik

PRINT PRODUCTION

February, 2021

ISBN:

Further information on the Indira Gandhi National Open University courses may be obtained from the University's office at Maidan Garhi, New Delhi -110068 or visit our website: http://www.ignou.ac.in

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi, by Director, School of Social Sciences.

Laser Typeset: Mr. Mukesh Yadav **Cover Design:** Mr. Sandeep Maini **Printed at:**

CONTENTS

COURSE INTRODUCTION

The importance of statistics in economics does not need any elaboration. The purpose of the present course is to equip you with some of the important statistical tools, which would help you in analysis and interpretation of data. After going through the present course you would notice that our objective has been two-fold. One, we have tried to draw upon examples as far as possible from real life situations. Two, we have taken care to explain details of the derivations. Hope these will help you understand the subject matter and apply it to situations confronting you. The course is divided into four blocks comprising 15 Units.

Block 1 titled, **Descriptive Statistics**, comprises four units. Unit 1 begins with certain basic concepts frequently used in Statistics. Subsequently, it discusses the issues involved in collection of data. Unit 2 carries out a detailed discussion on presentation of data through tables and graphs. Unit 3 deals with summarization of univariate data by measures of central tendency and dispersion. In Unit 4, the measures of skewness and kurtosis are given.

Block 2 titled, **Summarization of Bivariate and Multivariate Data**, deals with topics such as correlation and regression, index number, deterministic time series, and vital statistics. In Unit 5 you are introduced to the tools of analyzing bivariate data such as correlation and regression. Methods of calculating index numbers, various types of index numbers, and their limitations are subject matter of Unit 6. Unit 7 brings out the components of time series while Unit 8 discusses various birth, death and fertility rates.

Block 3 deals with probability and probability distributions. It consists of three units. Unit 9 titled, Elementary Probability, highlights various laws of probability theory. Unit 10 presents discrete probability distributions such as binomial and Poisson. Unit 11 presents the salient features of normal, chi-squared, t and F distributions.

Block 4 titled, **Sampling and Statistical Inference**, comprises four units. Unit 12 begins with the concept of sampling and its types. The Unit also discusses the procedure of drawing a sample from a population. In Unit 13 we present the underlying ideas behind statistical estimation, including the concepts of sampling distribution of a statistic and standard error. In Unit 14 we discuss various methods of testing hypotheses. The final unit, Unit 15, deals with the application of chi-squared test in the context of contingency table.

UNIT 1 BASIC CONCEPTS AND DATA COLLECTION METHODS*

Structure

1.0 OBJECTIVES

After going through this Unit you will be able to

- distinguish between data and information;
- describe the types of variables and their measurement;
- identify the steps to be followed in collection of data;
- distinguish between primary and secondary data;
- describe the advantages and limitations of secondary data; and
- design a questionnaire.

1.1 INTRODUCTION

Now-a-days the term 'statistics' has becomes a household word, although different people comprehend it in different senses. Today an educated person has to be a person of statistics, broadly understanding its meaning and applying it to

[*] Adapted from IGNOU study material of EEC 13: Elementary Statistical Methods and Survey Techniques, Units 1, 2 and 3 written by Avatar Singh with modifications by K. Barik

his/her life in different ways. For example, every day we come across different types of quantitative information in print and electronic media on topics like population, exchange rate fluctuation, inflation rate, weather, etc. In order to improve our understanding of the world around us, it is necessary to

i) measure what is being said,
ii) express it numerically, and
iii) utilize quantitative information or expression to draw conclusions and suggest policy measures.

Let us try to define the term 'Statistics'. You will see a variety of definitions from various sources. Such variation arises because of the fact that Statistics as a discipline has witnessed rapid developments over time. Thus, topics that were considered to be the most important in statistics have been pushed to the periphery. For example, collection and description of data were of prime importance in the beginning stages. They continue to be important, but many other topics (such as, generalization of results from sample surveys) have gained importance. We define statistics as follows: Statistics is the branch of science that deals with collection, organization, presentation and interpretation of data.

There are two major branches of statistics: 'descriptive statistics' and 'inferential statistics'. As the name suggests, descriptive statistics is about description of various characteristics of the dataset. Thus descriptive statistics includes tabular and diagrammatic presentation of data, measures of central tendency (mean, median, mode, geometric mean and harmonic mean), measures of dispersion (range, variance, standard deviation), kurtosis, skewness, etc. In inferential statistics we collect data through a sample survey and generalize the sample results to the whole population. Inferential statistics makes extensive use of probability laws.

You would come across many applications of statistics, viz., consumer price index, BSE SENSEX of the Indian stock market, Six Sigma as a business management tool, etc. You may have watched on TV the exit poll results before elections – it forecasts the seats all political parties are likely to win after election. Here, the analysts take a random sample of voters, analyse their responses, and extend the sample results to the whole population. Further, many times we conduct experiments in laboratories or in the field. Finding out the effectiveness of a new medicine, estimation of yield of a new variety of seed, and testing for safety measures in a new car model are some examples experimental statistics. Proper design of experiments and drawing inferences from experiments are important steps in experimental statistics.

1.2 CERTAIN CONCEPTS

In this section we define certain concepts which are frequently used.

Data and Information: There is a basic difference between data and information. Raw data (for example, daily data of the number of students visiting

IGNOU central library during October 2019) may not be of much use. We need to process the raw data so that it provides some useful information (for example, on an average 65 students visited IGNOU central library everyday during October 2019). You would have come across data from various sources on a lot of themes such as labour, stock prices, agricultural production, births and deaths, etc. When we process these data we obtain some meaningful information. The meteorological department provides daily data on rainfall, wind speed, temperature, etc. These figures, on being processed further, provide us useful information in the form of weather forecast.

Population and Sample: In statistics, the term population is somewhat different from that in ordinary speech. In ordinary speech we may say ‘the population of Odisha according to 2011 census is 4.2 crores’. In this statement, population reflects human beings. In statistics, however, population is a much broader concept – it means a group of objects, observations, events or living beings that we plan to study. Thus, population is not necessarily a collection of human beings. It could be a collection of objects (such as electric bulbs manufactured by a company or equities traded in a stock exchange), a collection of living beings (such as new born babies in a city or tigers in a forest), a collection of events (accidents during 2020 on a specific road) or observations (monthly income of employees of a firm). Sample is a subset of the population. Inn sample surveys we draw a sample from a population. Thus, the size of the sample is always smaller than the size of the population.

Parameter and Statistic: Parameter is a characteristic of the population under study. Statistic, on the other hand, is a characteristic of the sample. For instance, population mean is a parameter whereas sample mean is a statistic. Usually we denote parameters by Greek letters such as μ (mu) for population mean and σ (sigma) for population standard deviation.

1.3 TYPES OF VARIABLES AND THEIR MEASUREMENT

Variable is an attribute that describes a statistical unit (i.e., the unit of observation). Which assume different values? Thus, example of a variable could be certain characteristic of a person, place, object or event. For example, age of a person is a variable; so are gender, educational qualification, income, height and weight. You should note that value of a variable changes from one unit to another. You can think of several objects and events

‘Random variable’ as a concept is discussed in probability theory – it is a variable that has certain probability attached to each value it takes. For example, if you toss a coin you have two possibilities – Head and Tail. Each one has a probability of occurrence of 0.5. Thus tossing of a coin is a random variable. You can think of several examples of random variables – if you roll a dice, the probability of occurrence of each face is $\frac{1}{6}$.

1.3.1 Discrete Variable and Continuous Variable

Discrete variables assume numerical values that are countable. For example, the number of children in a family, the shoe size of a person, the number of customer complaints in a firm, etc. You should note that the values that the above variables take are whole numbers. For example, the number of children in a family can be complete number such as 2 or 3 – there is nothing in between 2 and 3.

Continuous variables, on the other hand, assume numeric values; but the number of values between any two points is infinite. For example, the time taken to run 100 metres – the world record for 100 metre race, as of January 2021, is 9.58 seconds. May be in the coming years someone will break this record by a fraction of a second! You can think of several continuous variables – height, weight and age of persons. In the case of weight, for instance, the weight of a person could assume any value – it could be in fractions of any measurement unit.

1.3.2 Measurement Scales

As pointed out earlier, variable is an attribute, which can be measured. Different attributes, however, require different measurement scales. There are four types of measurement scales, viz., nominal, ordinal, interval and ratio.

(a) **Nominal Scale:** In this case we put a population or sample characteristic under certain number of categories. For example, gender (male or female), place of residence (rural or urban), payment method (cash, cheque, credit card, debit card or net banking), etc. You should note that, in the case of nominal variables, all categories are considered to be equal.

(b) **Ordinal Scale:** In this case also a population characteristic is put under several categories; but the categories are hierarchical. In other words, we can rank the categories. For example, we can divide a group of persons into four categories according to their educational qualification, viz., illiterate, secondary, senior secondary, and graduation. We can also say that graduation is higher in qualification than senior secondary, senior secondary is higher than secondary, and secondary is higher than illiterate. Thus there is certain 'order' across the categories in the case of ordinal scale. But the difference between secondary and senior secondary is not equal to the difference between senior secondary and graduation. Thus, the differences across categories are not equal in an ordinal variable.

(c) **Interval Scale:** Interval scale has two important properties. One, there is no absolute zero. Two, the difference between two numbers is equal. Let us take a concrete example - measurement of temperature by degree Celsius (^{0}C). The freezing point of water is 0^0C while the boiling point is 100^0C. Here, 0^0C does not mean absence of temperature; similarly, 100^0C does not mean the maximum temperature of an object (it can be much higher than 100^0C!). The difference between 97^0C and 98^0C however is equal to the difference between 98^0C and 99^0C.

(d) **Ratio Scale:** When a scale has all the attributes of an interval scale and in addition, has a true zero as its origin, then it is called ratio scale. For example, weight of an object. It can be measured in grams. Further, there is an absolute zero – it means there is no weight at all.

1.4 COLLECTION OF DATA

Collection of relevant and reliable statistical information is a pre-requisite of any statistical inquiry. This and the subsequent Sections of this Unit are devoted to data collection techniques.

Collection of reliable and sufficient data requires a careful planning and execution of a statistical survey. If this is not so then the result obtained may be misleading or incomplete and hence useless. They may even do more harm than good. In the following Section an attempt is made to explain the planning aspect of a statistical enquiry.

Statistical data can be collected either by a survey or by performing an experiment. Surveys are more popular in social sciences like economics and business studies. In natural/ physical sciences experimentation is commonly used.

1.4.1 Statistical Inquiry – Planning and Conduct

Data collected by observing various individuals or items, included in a survey, are affected by a large number of uncontrollable factors. For example, wages in a country are affected by a lot of factors like skill, education and sex of worker; training and experience; and in some countries even on race to which a worker belongs. In India low caste and historically underprivileged people like sweepers are the least paid workers for social reasons also.

It is interesting to note that even the data obtained through experiments in physical sciences are affected by a large number of uncontrollable factors in spite of the fact that such experiments are conducted under controlled conditions. The uncontrollable factors, in this case, may arise due to the bias of the person(s) conducting the experiment, nature and accuracy of measuring instrument, etc.

Any statistical survey consists of two stages:

I Planning Stage

II Executing Stage

1.4.2 Planning Stage – Requisites of a Statistical Inquiry

Before collecting data through primary or secondary source, the investigator has to complete the following preliminaries.

a) *What is the objective / aim and scope of the inquiry*?

Unless the investigator answers this question most satisfactorily, he cannot proceed in the right direction and can go astray. Both money and efforts will be lost if data, not relevant to inquiry, are collected.

Not only this, one must also be clear about how much data are required and hence ensure that only the necessary data get collected. For example, if we want to collect data on pattern of wheat production in a particular state, we need to collect data on the type of land, agricultural inputs, educational levels of farmers involved, presence or absence of defects of land tenure system, availability and cost of agricultural finance, nature of marketing, etc.

b) *What shall be the source of information*?

The investigator has to make a choice between primary source, where he himself collects the data, or secondary source, where he lays his hand on already collected data.

c) *What shall be the nature of inquiry*?

That is, the investigator has to make a choice between:

1. *Census* or *Sample* inquiry. In census method (s)he examines each and every item / individual of the population whereas in sample method he examines only the item / individual included in the sample. For example, in census method (s)he examines each and every person in a village, but in sample method, he examines only a limited number of persons.
2. *Direct* or *Indirect* inquiry. In a direct inquiry the observations can be directly obtained in quantitative terms as for example, sales of T.V. sets and the advertisement cost in rupees. On the other hand, in an indirect inquiry, like intelligence of a group of students, marks secured by them are used to judge their intelligence.
3. *Original* or *Repetitive* inquiry. An inquiry conducted for the first time is original but if it is undertaken over and over again, it is repetitive. For example, population census in India is conducted every 10 years. All these inquiries must be related.
4. *Open* or *Confidential* inquiry. In open inquiry the results are made public, as for example, the population and national income data. On the other hand, the results of many government inquiries are kept confidential for reasons of national security, as for example, data on defence, atomic energy, space research and development, etc.

d) *What shall be the statistical units of investigation or counting?*

A statistical unit is an attribute or a set of attributes conventionally chosen so that individuals or objects possessing them may be counted or measured for the purpose of enquiry. Thus a statistical unit is a characteristics or a set of characteristic of an individual or item that are observed to collect information. For example, various characteristics of a person may be his height, weight, income, etc. The definition of a statistical unit means the specification of the characteristics of an individual or item on which data are to be collected.

It must be pointed out that the result of observation of a statistical unit may be a number which is obtained either by counting or by measurement. If the number is obtained by measurement, it is also necessary to specify the units of measurement. The specification of statistical units and the units of measurements is very necessary for the maintenance of uniformity in the collected data.

e) *What shall be the degree of accuracy*?

In various economic and business studies, absolute accuracy is neither necessary nor possible. In population data, accuracy till the last man is not required. For example, population of India according to 2011 census is 1,210,193,422, which is approximated as 1.21 billion. The degree of accuracy required will determine the choice between different methods of collecting data. Further, the degree of accuracy, once decided, must be maintained throughout the survey.

1.4.3 Execution Stage

This stage comes after the planning stage, where the plan is put in operation. It includes:

1) Setting up the *central administrative machinery* which prepares a format of questions relating to the inquiry, called a *questionnaire* or a *question schedule*. It decides the setting up of branch offices to cover large geographical areas, depending upon the type and size of inquiry.

2) *Selection* and *Training* of field staff called interviewers or investigators or research staff or enumerators. They will approach the respondents in different ways as explained in Section 1.4. These people should be properly trained, should be honest and hard working. Any error at this stage will jeopardize the whole process of investigation giving misleading results. To obtain the best possible results from a survey, it is desirable to have the field staff who are familiar with the language of the respondents and have patience and tact of dealing with them.

3) *Supervision* of field staff is a must to ensure that information is actually obtained from the respondents rather than that the questionnaires are fictitiously filled up in the hotel rooms. Further, there must be some experts to make clarifications on problems faced by the investigators in the field work.

   While conducting field surveys the problem of *Non-response is* common. This includes:

   a. Non-availability of the listed respondent. Here in no case this respondent be replaced by another because it may spoil the random character of sample and the results of investigation are likely to become biased.

   b. Due to non-response, a part or certain questions of the questionnaire may remain unanswered or partly answered. These should not be replaced or tempered with by the investigator.

4) After the data have been arranged, the next job is to analyse the same. The methods of doing this are fully described in later Units. Now-a-days computers are available to do this job.

5) After analysis of data, now is the turn for writing a detailed report mentioning the main findings of the survey/statistical inquiry. The main conclusions drawn and policy recommendations are duly recorded at the end of this report.

1.4.4 Primary and Secondary Data

A pertinent question that arises now is how and from where to get data? Data are obtained through two types of investigations, namely,

1) *Direct Investigation* which implies that the investigator collects information by observing the items of the problem under investigation. As explained above, it is the primary source of getting data or the source of getting primary data, and can be done through observation or through inquiry. In the former we watch an event happening, as for example, number and type of vehicles passing through Vijay Chowk in New Delhi during different hours of the day and night. In the latter we ask questions from the respondents through questionnaire (personally or through mail). It is costly method in terms of money, time and efforts.

2) *Investigation through Secondary Source* which means obtaining data from the already collected data. Secondary data are the other people's statistics, where other people includes governments at all levels, international bodies or institutions like IMF, IBRD, etc., or other countries, private and government research organisations, Reserve Bank of India and other banks, research scholars of repute, etc. Broadly speaking we can divide the sources of secondary data into two categories published sources and unpublished sources.

A. Published Sources

1. Official publications of the government at all levels – Central, State, Union Territories and Councils.
2. Official publications of foreign countries.
3. Official publications of international bodies like IMF, UNESCO, WHO, etc.
4. Newspapers and Journals of repute, both local and international.
5. Official publications of RBI, and other Banks, LIC, Trade Unions, Stock Exchange, Chambers of Commerce, etc.
6. Reports submitted by reputed economists, research scholars, universities, commissions of inquiry, if made public.

Some important sources of published data in India are

I. **Central Statistical Office** (C.S.O.): It publishes data on national income, savings, capital formation, etc. in a publication called National Accounts Statistics.

II. **National Sample Survey Office** (N.S.S.O.): Under the Ministry of Finance, this organisation provides us data on all aspects of national economy, such as agriculture, industry, labour and consumption expenditure.

III. **Reserve Bank of India** (R.B.I.): It publishes a number reports and databases. Its publications are Handbook of Statistics on Indian Economy, Handbook of Statistics on Indian States, Report on Currency and Finance, RBI Bulletin, Statistical Tables Relating to Banks in India, etc.

IV. **Labour Bureau**: Its publications are Indian Labour Statistics, Indian Labour Year Book, Indian Labour Journal, etc.

V. **Population Census**: Undertaken by the office of the Registrar General India, Ministry of Home Affairs. It provides us different types of statistics about population.

B. Un-published Sources

1. Unpublished findings of certain inquiry committees.
2. Research workers' findings.
3. Unpublished material found with Trade Associations, Labour Organisations and Chambers of Commerce routine.

Check Your Progress 1

1. Explain the following terms :

   (Answers should not exceed three sentences each.)

   a) Variable b) Population

   c) Non-response d) Statistical unit

   e) Statistical inquiry f) Question Schedule

   ....

   ....

   ....

   ....

2. Distinguish between the following terms :

   (Answers should not exceed four sentences each.)

   a) Nominal scale and Ordinal scale

   b) Interval scale and Ratio scale

   c) Census and Sample Survey

   d) Planning and Execution of Statistical Inquiry

   e) Primary and Secondary Data

   f) Survey and Experiment

....

....

....

....

....

....

3. What are the different sources of data?

....

....

....

....

1.5 COLLECTION OF PRIMARY DATA — SURVEY TECHNIQUES

After the investigator is convinced that the gain from primary data outweighs the money cost, effort and time, she/he can go in for this. She/he can use any of the following methods to collect primary data:

a) Direct Personal Investigation

b) Indirect Oral Investigation

c) Use of Local Reports

d) Questionnaire Method

a) Direct Personal Investigation

Here the investigator collects information personally from the respondents. She/he meets them personally to collect information. This method requires much from the investigator such as:

1. She/he should be polite, unbiased and tactful.
2. She/he should know the local conditions, customs and traditions so that she/he is able to identify herself/himself as one of the respondents.
3. She/he should be intelligent possessing good observation power.
4. She/he should use simple, easy and meaningful questions to extract information.

This method is suitable only for intensive investigations. It is a costly method in terms of money, effort and time. Further, the personal bias of the investigator cannot be ruled out and it can do a lot of harm to the investigation.

The method is a complete flop if the investigator does not possess the above mentioned qualities.

b) Indirect Oral Investigation Method

This method is generally used when the respondents are reluctant to part with the information due to various reasons. Here, the information is collected from a witness or from a third party who are directly or indirectly related to the problem and possess sufficient knowledge. The person(s) who is/are selected as informants must possess the following qualities:

1. They should possess full knowledge about the issue.
2. They must be willing to reveal it faithfully and honestly.
3. They should not be biased and prejudiced.
4. They must be capable of expressing themselves to the true spirit of the inquiry.

c) Use of Local Reports

This method involves the use of local newspaper, magazines and journals by the investigators. The information is collected by local press correspondents and not by the investigators. Needless to say, this method does not yield sufficient and reliable data. The method is less costly but should not be adopted where high degree of accuracy or precision is required.

d) Questionnaire Method

It is the most important and systematic method of collecting primary data, especially when the inquiry is quite extensive. It involves preparation of a list of questions relevant to the inquiry and presenting them in the form of a booklet, often called a questionnaire. The questionnaire is divided into two parts:

1. General introductory part which contains questions regarding the identity of the respondent and contains information such as name, address, telephone number, qualification, profession, etc.
2. Main question part containing questions connected with the inquiry. These questions differ from inquiry to inquiry.

Preparation of the questionnaire is a highly specialized job and is perfected with experience. Therefore, some experienced persons should be associated with it. The following few important points should be kept in mind while drafting a questionnaire:

I. The task of soliciting information from people in desired form and with sufficient accuracy is the most difficult problem. By their nature people are not willing to reveal any information because of certain fears.

Many a times they provide incomplete and faulty information. Therefore, it is necessary that the respondents be taken into confidence. They should be assured that their individual information will be kept confidential and no part of it will be revealed to tax and other government investigative agencies. This is very essential indeed.

II. Where providing information is not legally binding, the informant has to be induced through appeals or by using clever arguments. They must be explained and convinced that the results of the survey will help the authorities to frame policies which will ultimately benefit them. It is obvious that some element of good salesmanship is also required in the investigation.

III. Always avoid personal questions which may embarrass the respondents. For example, questions like ‘Do you evade income tax?’ or ‘Are you engaged in smuggling or black marketing?’ should not be asked.

IV. Questions hurting the sentiments of respondent should not be asked. These include questions on his gambling habits, sex habits, indebtedness, etc.

V. Questions involving lengthy and complex calculations should be avoided because they require tedious extra work in which the respondent may lack both interest as well as capabilities. In such cases it would be better to

i) either get documents like balance sheet, profit and loss account and inventory record from the respondent from where the investigator can get or calculate the required information himself, or

ii) ask indirect and simple questions which, with some calculation later on, can help us to acquire the required information.

VI. Ask questions which enable to cross check the correctness of the information supplied by the respondent. For example, questions on total wage bill of a factory can be cross checked if the other questions seek information on different types of workers working in administrative, production, store and marketing departments. Similarly information on saving of a household can be cross checked by getting information on different sources of income and its expenditure on different heads.

VII. As far as possible questions should be of Yes/No type. These are precise and simple to understand, and take very little time to answer. Later on they are easy to tabulate. For example,

Are you married? Yes/No

Tick (√) the right answer.

VIII. Questions should be short and clear. That is, they should not be ambiguous and confusing.

As far as possible, attempt should be made to suggest the possible answers to a question and the respondent may be asked to simply tick the answer/s he/she thinks is/are correct.

Since the list of answers may not be exhaustive, therefore, a line of "others, if any" should also be inserted leaving sufficient blank space for the answer.

Following is an example of a question:

Why do people not exercise their right to vote?

Tick (√) the right answer:

A) They are illiterate and do not understand the value of the vote.

B) They think, it does not matter if their one vote is not cast out of lakhs.

C) The polling booths are far from their residence.

D) They are afraid of the local goons and violence.

E) They are not happy with the government and do not vote out of protest.

F) They do not vote unless some money is not offered to them.

G) Any other reason, please state.

This form of questions and answers also helps in arranging and tabulating the data.

IX. A very large number of questions should be avoided because it leads to the feeling of monotony. Many respondents will hesitate to answer a long list of questions, for want of time and interest.

A sample questionnaire on family planning is presented below.

Survey on Family Planning

1. Name____

2. Father's / Husband's Name ______________________________

3. Residential address_____________________________________

4. Place of Work____

5. Age ____________________ 6. Male/Female___________

7. Religion ________________ 8. Telephone No. __________

9. Profession: a) Self _________b) Spouse ________________

10. Annual Income of the family from all sources _______________

11. Educational Qualifications: (Tick (√) the right answer)

a) Illiterate b) Primary level
c) Middle level d) Secondary
e) Sr. Secondary f) Graduate
g) Post graduate

12. Educational Qualifications of spouse: (Tick (√) the right answer)

a) Illiterate b) Primary level
c) Middle level d) Secondary
e) Sr. Secondary f) Graduate g) Post graduate

13. Number of years of married life ______________________

14. Number of children born: Girls________ Boys__________

15. Number of surviving children: Girls_____ Boys__________

16. State the gap in years between the children

  a) Between marriage and first child : _____________

  b) Between first and second child : _____________

  c) Between second and third child : _____________

  d) Between third and fourth child : _____________

  e) Between fourth and fifth child : _____________

  f) Between fifth and sixth child : _____________

17. Do you favour family planning? (Yes/No)

18. If no, what are the reasons?

  a) Children are natural gift: (Yes/No)

  b) Family planning is against my religion: (Yes/No)

  c) Family planning means murdering an unborn child: (Yes/No)

  d) Number of children is the part of my fate: (Yes/No)

  e) Any other reason, please state:

  ____

19. If you favour family planning, state the reasons

  a) Small family is a happy family: (Yes/No)

  b) Two children can be controlled easily: (Yes/No)

  c) Two children can be properly educated and fed: (Yes/No)

  d) There are fewer complications in life: (Yes/No)

  e) The health of the mother is not adversely affected: (Yes/No)

  f) Any other reason, please state:_________________________

20. State Age, Educational Level and Health Condition of your children.

  Sl.No. Name Age Educational Level Health condition*

  1.___________________ ___ _________ ___________

  2.___________________ ___ _________ ___________

  3.___________________ ___ __________ ___________

  4.___________________ ___ __________ ___________

(* State whether Poor, Below Normal, or Excellent)

How to approach the Respondent with a Questionnaire?

There are three methods available to us:

I. Send the questionnaires by email to the respondents with a forwarding letter highlighting the importance of the survey to them as well as to the community or nation, and requesting cooperation in filling it and then you can sit back and wait for the response. It is often seen that the response is generally poor.

II. Send the questionnaire through investigators, who will interview the respondents and record the information personally. This method, though costly, is better. It helps the respondents to understand questions properly. The response is certainly better because the scope of laziness and irresponsibility is reduced. A clever and intelligent investigator with tact and initiative is able to get better response.

III. Send the questionnaire by post/e-mail followed by the visit of the investigator. This in fact is the best method as it combines the benefits of both the methods. It, no doubt, is a costly method. It is very useful for extensive studies. Being expensive, it can and is normally used by Government who has financial resources at its command.

1.6 COLLECTION OF SECONDARY DATA

As pointed out in Section 3.3.4 that the direct investigation, though desirable, is costly in terms of money, time and efforts. Alternatively, information can also be obtained through a secondary source. It means drawing or collecting data from the already collected data of some other agency. Technically, the data so collected are called secondary data.

Limitations of Secondary Data

Although the secondary source is cheap in terms of money, time and effort, utmost care should be taken in their use. It is desirable that such data should be vast and reliable; and the terms and definitions must match the terms and definitions of the current inquiry. The suitability of the data may be judged by comparing the nature and scope of the present inquiry with that of original inquiry. Secondary data will be reliable if these were collected by unbiased, intelligent and trained investigators. The time period to which these data belong, should also be properly scrutinized. Corner has rightly remarked, "*Statistics, especially other people's statistics are full of pitfalls for the user*". Needless to say, before using secondary data, the investigator must weigh the advantage in terms of saving of money, time and effort with the disadvantage of reaching misleading conclusions. Whether secondary data are safe or not should be judged from its *adequacy*, *suitability* and *reliability*.

Thus, before the use of secondary data, i.e., other persons' data, we must properly scrutinize and edit them to find whether these data are:

1. Reliable,
2. Suitable, and
3. Adequate.

Reliability of data has to be the obvious requirement of any data, and more so of secondary data. The user must make himself sure about it. For this he must check whether data were collected by reliable, trained and unbiased investigators from dependable sources or not. Second, we should see whether data belong to almost the same type of class of people or not. Third, he should make sure that due to the lapse of time, the conditions prevailing then are not much different from the conditions of today in respect of habits, customs, fashion, etc. Of course we cannot hope to get exactly the same conditions.

Suitability of data is another requirement. The research worker must ensure that the secondary data he plans to use suits his inquiry. He must match class of people, geographical area, definitions of concepts, unit of measurement, time and other such parameters of the source he wants to use with those of his inquiry. Not only this, the aim and objectives should also be matched for suitability.

Secondary data should not only be reliable and suitable, but also *adequate* for the present inquiry.

It is always desirable that the available data be much more than required by the inquiry. For example, data on, say, consumption pattern of a state cannot be derived from the data on its major cities and towns.

Check Your Progress 2

1. State, with reasons, whether the following statements are correct or incorrect?

   a) Secondary data are better than primary data.

   b) Data obtained from Population Census of India 2011 are Primary Source of Data.

   c) Secondary Data should not be accepted without scrutiny.

   d) A Questionnaire with a very long list of questions is justified.

   e) Of all the Survey Techniques, the Questionnaire method is the best.

   ....

   ....

   ....

   ....

2. State two most important characteristics of an investigator when Direct Personal Investigation Method is being used for collection of Primary data.

   ....

   ....

   ....

   ....

3. State and explain various Survey Techniques.

   ....

   ....

   ....

   ....

4. Comment on the statement: "We should always use secondary data."

   ....

   ....

   ....

   ....

1.7 LET US SUM UP

When data is processed we obtain some useful information. Data can be collected through a census or a sample survey. Measurement of a variable

For conducting an inquiry, we need data which can be collected afresh or from a secondary source. Both require Statistical Survey which has a planning stage and an executing stage. In the Planning Stage, the investigator should decide whether to use Primary or Secondary source, Census or Sample inquiry, nature of the statistical units and the units of measurement, degree of accuracy desired and so on.

In the Execution stage, the chief investigator has to set up administration, select and train field staff and supervise the entire process of data collection.

Care has to be taken in using the Secondary data, derived from published or unpublished source, as they contain various pitfalls.

Of all the Survey techniques, the Questionnaire method is very important. A Questionnaire contains a set of relevant questions which should be simple, unambiguous, Yes/No type with suggestive answers. Their list should not be very long. Personal and embarrassing questions should be avoided.

1.8 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1. For (a) and (b) go through Section 1.2

   For (c), (d), (e) and (f) go through Section 1.4
2. For (a) and (b) go through Section 1.2

   For (c), (d), (e) and (f) go through Section 1.4
3. Go through Sub-Section 1.3.4

Check Your Progress 2

1. (a), (b), (d) False

   (c), (e) True
2. Go through Section 1.4.
3. Go through Section 1.4.
4. Go through Section 1.5

UNIT 2 TABULATION AND GRAPHICAL REPRESENTATION OF DATA*

Structure

2.0 OBJECTIVES

After going through this Unit, you will be able to familiarize yourself with

- stages of statistical inquiry after data have been collected;
- methods of organizing (classification and arrangement) and condensing statistical data;
- concepts of frequency distribution and its various types; and
- different methods of presentation of statistical data such as tables, graphs, diagrams, pictograms, etc.

* Adapted from IGNOU study material of EEC 13: Elementary Statistical Methods and Survey Techniques, Units 1, 2 and 3 written by Avatar Singh with modifications by Kaustuva Barik.

2.1 INTRODUCTION

In the preceding Unit, we discussed the methods of collection of data either by a statistical survey (or inquiry) or from some secondary source. Data collected either from census or sample inquiry, that is from primary source, are always hotchpotch and in rudimentary form. To start with, they are contained in hundreds and thousands of questionnaires. To make a head and tail out of them, they must be organised (i.e., classified and arranged), and condensed or summarised. For this purpose we can use various methods like preparing master sheets in which various information are recorded directly from the questionnaires. From these sheets small summary tables can be prepared manually. Now-a-days computers can be used for organisation and condensation of data more swiftly, efficiently and in much less time. Some computer softwares are available which help us to construct various types of graphs and diagrams.

Data can be summarized numerically also. Here we use summary measures like measures of central tendency of first order or degree (like Arithmetic, Geometric and Harmonic Means, Mode and Median); measures of central tendency of second order or degree, also called dispersion measures (like Range, Quartile Deviation, Mean Deviation, and Standard Deviation); measures of association in bivariate analysis (like Correlation and Regression), Index Numbers, etc. In this Unit we plan to discuss how data can be summarized using tables and graphs. Numerical summarization will be discussed subsequently in Units 3 and 4). It must be born in mind that a good summarization and presentation of data is not undertaken for its own sake. It is not an end in itself. In fact it sets the stage for useful analysis and interpretation of data. Again, a good presentation helps us to highlight significant facts and their comparisons. Figures can be made to speak out thereby making possible their intelligent use.

In this Unit we plan to concentrate on organising and condensing data in the form of simple array (ascending and descending order), frequency array and continuous frequency distributions, etc.; and presentation of statistical data in the form of tables and graphs.

2.3 ARRANGEMENT OF DATA

The mass of collected data is often voluminous, unintelligible and boring. It seems totally uninteresting and is not easily interpretable. For example, if you are provided with monthly income figures of 1000 families in a village it is difficult for you to infer anything. But if you are told that the average monthly income of the village is Rs. 2540, it is quite interesting and you are in a position to compare it with other figures.

The first step in the analysis and interpretation of data is its classification and tabulation. The process of arranging data into groups according to their common characteristics is known as its classification.

On the other hand tabulation implies a systematic presentation of data in rows and columns according to some salient features or characteristics.

In Unit 1, a questionnaire was prepared on Family Planning. Suppose this questionnaire was used to collect information from 50 families of C-III block of XYZ Colony, New Delhi. Let us assume that it produced the following types of information as given in Tables 2.1 and 2.2. *Can we make any head or tail out of it*?

Table 2.1
Number of Children per family in C-III block, XYZ Colony, New Delhi

2	0	1	5	3	1	2	1	0	2
4	3	2	2	3	4	1	0	2	3
1	4	2	3	1	2	5	4	1	3
2	1	3	2	3	4	1	2	3	1
4	5	2	1	1	0	3	2	0	2

Table 2.2
Monthly Income of 50 families of C-III block, XYZ Colony, New Delhi

547	622	691	684	567	586	680	578	583	578
708	544	528	540	730	541	720	698	763	633
640	637	598	631	618	692	600	650	604	640
646	654	689	736	731	844	798	712	772	820
678	663	800	692	700	781	658	798	709	720

As pointed out earlier, to make any head or tail out of the mass of raw data, such as presented above, we have to classify and arrange it. This can be done either by forming a simple array or a frequency array (discrete frequency distribution) or a continuous frequency distribution. Sections 2.3.1, 2.3.2 and 2.3.3 attempt to explain this aspect.

2.2.1 Simple Array

It is an arrangement of given raw (univariate) data in ascending or descending order. In the ascending order, the observations are arranged in increasing order of magnitude. For example, numbers 3,5,7,8,9,10 are arranged in ascending order. In descending order, it is the reverse. For example, the numbers 10,9,8,7,6,5,3 are in descending order.

We can prepare both types of simple arrays from Table 2.1. In the following table, the figures have been arranged in ascending order. From the arrangement, it is clear that the lowest value is 0 and the highest one is 5.

Table 2.3
Number of Children per Family in C-III Block of XYZ Colony, New Delhi
Simple Array -- Ascending Order

0	0	0	0	0	1	1	1	1	1
1	1	1	1	1	1	1	2	2	2
2	2	2	2	2	2	2	2	2	2
2	3	3	3	3	3	3	3	3	3
3	4	4	4	4	4	4	5	5	5

After this arrangement the same figures make some sense.

The possible conclusions that can be drawn from this arrangement of data are that five families are issueless, twelve families have one child each, fourteen have two children each, ten families posses three children each, six families four children each and three families have five children each.

2.2.2 Frequency Array or Discrete Frequency Distribution

Here different observations are not repeatedly written as in simple array like 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, etc. We count the number of times (i.e., frequency) an observation repeats itself. For example, in Table 2.3 the observation 4 is repeated 6 times. Thus the frequency of 4 above is 6. The frequency array, for the simple array given in Table 2.3, will look like as given below in Table 2.4.

A frequency array is a statistical table in which various observations are arranged in order of their magnitude along with their respective frequencies.

Table 2.4

Number of Children:	0	1	2	3	4	5	Total
Number of Families:	5	12	14	10	6	3	50

When the number of observations is large enough, the counting process is often undertaken by the use of *tally bars*. In this method, all possible values of the variable are written in a column. For every observation, a tally bar denoted by (|) is noted against its corresponding values. Every fifth repetition is marked by crossing the previous four bars as (𝍸). In this way, we get blocks of five which simplify counting at the end. Thus a number or an observation repeated fourteen times will be marked as (𝍸 𝍸 IIII). Note that after representing each observation by a bar on the *tally sheet*, the same will be ticked (√) or crossed (×) so that it is not duplicated. The data of Table 2.1 is rewritten in the form of frequency distribution as shown in Table 2.5 below.

Table 2.5
Frequency Distribution of Number of Children per Family

No. of Children	Tally Sheet	Frequency
0	𝍸	5
1	𝍸 𝍸 𝍷𝍷	12
2	𝍸 𝍸 𝍷𝍷𝍷𝍷	14
3	𝍸 𝍸	10
4	𝍸 𝍷	6
5	𝍷𝍷𝍷	3
Total		**50**

2.3.3 Continuous or Grouped Frequency Distribution

Numbers like 1, 2, 3, 4, 5, 20, 40, etc. are discrete numbers and are used where no value between the two consecutive numbers is possible. As in the case of the number of children, it will be impossible as well as funny to say that a particular family has 2.083 or 2.1 or 2.75 number of children. The family can have either 2 or 3 children and not a fraction of a child. Out of the two examples of raw data mentioned in Section 2.3, the number of children (Table 2.1) is an example of discrete data while monthly income (Table 2.2) is an example of continuous variable giving continuous data.

In this Section we propose to illustrate the construction of continuous or grouped frequency distribution from the raw data of Table 2.2 on monthly income of the 50 families.

To construct a grouped frequency distribution, the range of the given data, i.e., the difference of the highest and the lowest observations, is divided into various mutually exclusive and exhaustive sub-intervals, also known as class-intervals. The frequency of each class interval is then counted and written against it.

Table 2.6
Frequency Distribution of Monthly Income of Families

Monthly Income (Rs.)	Tally Sheet	No. of Families (Frequency)
500 - 550	𝍸	5
550 - 600	𝍸 𝍷	6
600 - 650	𝍸 𝍸	10
650 - 700	𝍸 𝍸 𝍷𝍷	12
700 - 750	𝍸 𝍷𝍷𝍷𝍷	9
750 - 800	𝍸	5
800 - 850	𝍷𝍷𝍷	3
Total		**50**

We have just completed an exercise where the variable "*income of the family*" has been grouped in order to reduce it to a manageable form called grouped data or *Continuous Frequency Distribution*. However, prior to the construction of any grouped frequency distribution, it is very important to find answers to the following questions:

1. What should be the number of class intervals?
2. What should be the width of each class interval?
3. How will the class limits be designated?

1. What should be the number of class intervals?

Though there is no hard and fast rule regarding the number of classes to be formed, yet their number should be neither too small nor too large. If the number of classes is too small, i.e., width of each class is large, there is likelihood of greater loss of information due to grouping. On the other hand, if the number classes is very large, the distribution may appear to be too fragmented and may not reveal any pattern of behaviour of the variable. Based on experience, it has been observed that the minimum number of classes should not be less than 5 or 6 and in any case, there should not be more than 20 classes.

Mr. Sturges has given a formula to determine the number of classes, that is,

Number of classes = $1 + 3.322 \times \log_{10} N$,

where N is the total number of observations.

In our example of raw data on incomes of 50 families, the number of classes can be calculated as under:

$$\text{Number of classes} = 1 + 3.322 \times \log_{10} 50 = 1 + 3.322 \times 1.6990$$
$$= 1 + 5.644 = 6.644 \approx 7.$$

2. What should be the width of each class interval?

As far as possible, all the class intervals should be of equal width. However, when a frequency distribution, based on equal class intervals, does not reveal a regular pattern of behaviour of observations, it might become necessary to re-group the observations into class intervals of unequal width. By a regular pattern of behaviour we mean that there are no classes, with possible exclusion of extreme classes, where there are nil or very few observations while there is concentration of observations in their adjoining classes.

$$Width\ of\ a\ Class = \frac{L\arg est\ Observation - Smallest\ Observation}{Number\ of\ Class\ Intervals}$$

The approximate width of a class can be determined by the following formula:

However, the final decision, regarding width of class intervals, should also take into account the following points.

i) As far as possible, the width should be a multiple of 5, because it is easy to grasp numbers like 5, 10, 15, ... etc.

ii) It should be convenient to find the mid-value of a class.

iii) The observations in a class should be uniformly distributed.

3. How will the class limits be designated?

The smallest and the largest observations of a class interval are known as class limits. These are also termed as the lower and upper limits of a class, respectively. Since the mid-value of a class, which is used to compute mean, standard deviation, etc., is obtained from the class limits, it is very necessary to define these limits in an unambiguous manner. The following points should be kept in mind while defining class limits:

a) It is not necessary that the lower limit of the first class be exactly equal to the smallest observation of the data.
In fact it can be less than or equal to the smallest observation. Similarly, the upper limit of the last class may be greater than or equal to the largest observation of the data.

b) It is convenient to have the lower limit of a class either equal to zero or some multiple of 5 or 10.

c) The chosen class limits should be such that the observations in a class are uniformly distributed.

The class limits can be defined in either of the following methods:

i) *Exclusive Method*, and ii) *Inclusive Method*.

i) Exclusive Method: In this method, the upper limit of a class is taken to be equal to the lower limit of the following class. In order to keep various class intervals as mutually exclusive, it is decided that the observations with magnitude greater than or equal to lower limit but less than the upper limit of a class are included in it. For example, the class 500 - 550 shall include all observations with magnitude greater than or equal to 500 but less than 550. An observation with magnitude equal to 550 will be included in the next class, i.e., the class 550 - 600.

The major benefit of exclusive class intervals is that it ensures continuity of data because the upper limit of one class is the lower limit of the next class. In our example on monthly income (Table 2.6), there are 5 families whose income lies between Rs. 500 to Rs. 550, i.e., Rs. 500 to 549 and 6 families whose income lies between Rs. 550 to Rs. 600, i.e., Rs. 550 to 599, and so on. Based on this presumption we can rewrite this frequency distribution in the form of Table 2.7 also.

Table 2.7
Exclusive Class Intervals

Monthly Income (Rs.)	*Number of Families* (Frequency)
500 but less than 550	5
550 but less than 600	6
600 but less than 650	10
650 but less than 700	12
700 but less than 750	9
750 but less than 800	5
800 but less than 850	3
Total	**50**

ii) **Inclusive Method:** In this method, all the observations with magnitude greater than or equal to the lower limit but less than or equal to the upper limit of a class is included in it. Now observe Table 2.8. Income of Rs. 549 is included in the class 500 to 549 so that an income of Rs. 550 automatically goes to the next class of 550 to 599. Since the upper limit of one class is not equal to the lower limit of the following class, this saves us from the confusion whether Rs. 550 goes to (500 to 549) or (550 to 599) class.

Table 2.8
Inclusive Class Intervals

Monthly Income (Rs.)	Number of Families (Frequency)
500 - 549	5
550 - 599	6
600 - 649	10
650 - 699	12
700 - 749	9
750 - 799	5
800 - 849	3
Total	**50**

The *choice between exclusive and inclusive methods* depends upon whether we are dealing with continuous variable like income, heights, weights, etc. or a discrete variable like number of children in a family. For a continuous variable it is desirable to construct frequency distribution by the exclusive method because, as we have seen earlier, it ensures continuity. For a discrete variable like number of children in a family or number of students getting first division, the frequency distributions should be constructed by using inclusive type of class intervals.

Mid-Value of a Class

In exclusive type of class intervals, the mid-value or class mark of a class is defined as the arithmetic mean of its lower and upper limits. However, in case of inclusive class intervals, there is a gap between the upper limit of a class and the lower limit of the following class. This gap is eliminated by adding half of the gap to the upper limit and subtracting half of the gap from the lower limit. The new class limits, thus obtained, are known as *class boundaries*. The class boundaries of the inclusive class intervals in Table 2.8 are given in Table 2.9.

Table 2.9

Monthly Income (Rs.)	No. of Families (Frequency)
499.5 - 549.5	5
549.5 - 599.5	6
599.5 - 649.5	10
649.5 - 699.5	12
699.5 - 749.5	9
749.5 - 799.5	5
799.5 - 849.5	3
Total	**50**

2.2.4 Various Forms of Frequency Distributions

Here we propose to introduce the meaning of the following frequency distributions:

a. Open Ended Frequency Distribution
b. A Frequency Distribution with Unequal Class Width
c. Cumulative Frequency Distribution
d. Relative Frequency Distribution

a) **Open End Frequency Distribution**

Open-end frequency distribution is one which has at least one of its ends open. Either the lower limit of the first class or upper limit of the last class or both are not specified. The words "below" or "less than" and "above" or "more than" are used. In the former the value extends to – ∞ and in the latter to + ∞. Example of such a frequency distribution is given in Table 2.10.

Table 2.10

Open-end Class Frequency

Class	Frequency
Below 25	1
25 - 30	3
30 - 40	5
40 - 50	2
50 and above	1
Total	**12**

Table 2.11

Unequal Class Frequency

Class	Frequency
20 - 25	1
25 - 30	3
30 - 40	5
40 - 55	2
55 - 60	1
Total	**12**

b) **A Frequency Distribution with Unequal Class Width**

The classes of a frequency distribution may or may not be of equal width. A frequency distribution with unequal class width is reproduced in Table 2.11. Here, the width of 1st, 2nd and 5th classes is 5, while that of 3rd is 10 and that of 4th is 15. As we will see in Unit 5, *mode* is not representative value in such types of series and hence not defined.

c) **Cumulative Frequency Distribution**

Suppose that, with reference to data given in Table 2.6, we ask the following questions:

i) How many families have their monthly income less than or equal to Rs. 700?
ii) How many families have their monthly income greater than or equal to Rs. 600?

The answers to the above questions can be easily obtained by forming an appropriate cumulative frequency distribution. To answer the first question, we need to form a "less than type" cumulative frequency distribution while a "greater than type" cumulative frequency distribution is required for answering the second question. These distributions are given in Tables 2.12 and 2.13 respectively.

Table 2.12

"Less-than type" Cumulative Frequency Distribution

Monthly Income (Rs.)	**Frequencies**		
	Simple		Cumulative
Less than 550	5		5
Less than 600	6	5+6	11
Less than 650	10	5+6+10	21
Less than 700	12	5+6+10+12	33
Less than 750	9	5+6+10+12+9	42
Less than 800	5	5+6+10+12+9+5	47
Less than 850	3	5+6+10+12+9+5+3	50

Table 2.13

"More-than type" Cumulative Frequency Distribution

Monthly Income (Rs.)	**Frequencies**		
	Simple		Cumulative
More than 500	5	3+5+9+12+10+6+5	50
More than 550	6	3+5+9+12+10+6	45
More than 600	10	3+5+9+12+10	39
More than 650	12	3+5+9+12	29
More than 700	9	3+5+9	17
More than 750	5	3+5	8
More than 800	3		3

d) Relative Frequency Distribution

So far we have expressed the frequency of a value or that of a class as the number of times an observation is repeated. We can also express these frequencies as a *fraction* or a *percentage* of the total number of observations. Such frequencies are known as *the relative frequencies*. Table 2.14 demonstrates the construction of relative frequency distribution.

Table 2.14

Relative Frequency Distribution of Monthly Income of 50 Families

Class	Frequency	Relative Frequency	
		As a fraction	As a percentage
500 - 549	5	5 ÷ 50 = 0.10	0.10 × 100 = 10
550 - 599	6	6 ÷ 50 = 0.12	0.12 × 100 = 12
600 - 649	10	10 ÷ 50 = 0.20	0.20 × 100 = 20
650 - 699	12	12 ÷ 50 = 0.24	0.24 × 100 = 24
700 - 749	9	9 ÷ 50 = 0.18	0.18 × 100 = 18
750 - 799	5	5 ÷ 50 = 0.10	0.10 × 100 = 10
800 - 849	3	3 ÷ 50 = 0.06	0.06 × 100 = 6
Total	**50**	**1.00**	**100**

From Table 2014 it is clear that the sum of relative frequencies should be either 1 (in the case of fraction) or 100 (in the case of percentages).

Check Your Progress 1

1. Distinguish between the following, giving at least two points of distinction.

   a) Discrete and continuous frequency distributions
   b) Simple and cumulative frequency distributions
   c) Exclusive and inclusive class intervals
   d) Simple and frequency array

   ....

   ....

   ....

   ....

   ....

   ....

2. Explain the following terms giving examples:

   a) Ungrouped data
   b) Class mark
   c) Open end classes
   d) Class limits
   e) Class boundaries
   f) Class frequencies
   g) Tally bar
   h) Relative frequencies

   ....

   ....

   ....

   ....

   ....

   ....

3. Build a hypothetical frequency distribution on monthly pocket money of 20 students belonging to the lower middle class of a college. Prepare a relative frequency distribution from it.

   ....

   ....

   ....

   ....

   ....

4. What points are to be kept in mind while taking decisions for preparing a frequency distribution in respect of:

   a) The number of classes, and
   b) Width of the class interval?

   ....

   ....

   ....

   ....

5. Construct less than and more than type cumulative frequency distributions from the following data:

Class:	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70
Frequency:	5	8	10	12	8	7

   ....

   ....

   ....

   ....

   ....

6. Construct a relative frequency distribution for the data given in question 5.

   ....

   ....

   ....

   ....

2.3 TABULATION OF DATA

Good presentation of data is as important as their satisfactory collection and arrangement. In fact, satisfactory collection and arrangement of data must be followed by good presentation. However, good presentation is not an end in itself. It may be necessary for satisfactory analysis and interpretation. A satisfactory presentation helps us in more than one ways. *Firstly,* it helps to highlight significant facts contained in the data. *Secondly*, it facilitates the comparison of data. *Finally*, it facilitates the easy understanding and an intelligent use of statistical information.

We will discuss presentation of statistical data under three heads.

i) Formal tables

ii) Graphic methods which will include line graphs, histograms, frequency polygon and curves, and cumulative frequency curves.

iii) Geometric forms, pictures and statistical maps, which will include pie diagrams, bar diagrams, area and volume diagrams, etc.

In this Section we concentrate on tabular forms of presentation.

2.3.1 Meaning and Types of Tables

A table or a statistical table is a systematic arrangement of related statistical data in columns and rows, with a given predetermined and a well decided objective. A row of a table represents a horizontal while a column represents a vertical arrangement of data. To explain the nature of information given in a table, its rows and columns are designated by appropriate stubs and captions (or headings or sub-headings) respectively. Presentation of data in a tabular form should be simple, planned, unambiguous and logical.

Table 2.15 is based on hypothetical figures of exports and imports of country X with countries A, B, C, and D for three years 1995, 1996 and 1997.

Table 2.15

Imports and Exports of X with A, B, C and D during 1995 - 1997

(in Crore of Rupees)

Country	1995		1996		1997@	
	Imports	Exports	Imports	Exports	Imports	Exports
A	60	70	65	75	70	65
B	50	60	60	65	65	60
C	40	30	40	40	42	50
D	45	42	60	55	63	55
Total	**195**	**202**	**225**	**235**	**240**	**230**

Note : @ Figures are quick estimates.
Source : Trade Bulletin, 1998, Ministry of Foreign Trade of X.

In this table it is clear that the purpose is to show the imports and exports of country X vis-à-vis the rest of the world. Note that a particular entry of the table refers to a column and a row. For example, an entry at the intersection of second row and third column indicates that in 1996 country X imported good and services worth Rs. 60 crores from country B. This figure then can be compared with other import and export figures to seek important interpretations.

Types of Tables

Basically, we have two types of tables:

i) Reference tables or general purpose tables

ii) Text tables or special purpose tables.

i) *Reference Tables* are a general purpose tables and are a store of information with the aim of presenting detailed statistical information. From these tables, we can derive our information (i.e., secondary source). Tables presented by different government departments, ministries, Reserve Bank of India, Economic Surveys, etc. are reference tables and are a routine work of these departments. Another important example is the Population Census tables prepared by the Registrar General of India giving detailed information on the demographic features of India.

ii) Students are advised to consult the latest issue of "Economic Survey" which is issued every year along with the union budget of India. Prepare from it a table on exports and imports of India to USA, UK, Russia, Canada and Germany for three or four years.

iii) *Text Tables* are the special type of tables. They are smaller in size and are prepared from the reference tables. Their aim is to analyse only a particular aspect to bring out a specific point or to answer a particular question. For example from the Population Census tables we may pick out information on the number of people in Bombay and Delhi who speak different languages (mother tongue), profess different religions and come from different states of India. Similarly from various publications of Reserve Bank of India, we may be able to extract information, in tabular form, on money supply, rate of interest and bank rate for the last ten years or so.

Tables can be simple and one way, like the tables given in Section 2.2, where we deal with only one variable, say, income. Alternatively, it is called a univariate frequency distribution. In addition to this, we can have two-way or multi-way tables where we deal with two or more related characteristics (for example, Table 2.15).

2.3.2 Parts of a Table

Parts or *elements* of a table vary from table to table depending upon the nature of data and purpose of tabulation. Yet some points are common. These are:

1. **Table number** is required for the identification of a table particularly when there are more than one tables in a particular analysis. Table number is always mentioned in the centre at the top.

2. **Title of the table** gives the indication of the type of information contained in the body of the table. It is said that the *title is to the table what heading is to an essay*. Next to the table number, we mention the title of the table. Its purpose is to answer the questions like:
   a) *What* is in the table?
   b) *Where* is it in the table?
   c) *When* did a particular information occur?
   d) *How* has a particular information been arranged?

In respect of a sample of a table on exports and imports, (Table 2.15), these questions will be answered as below:

a) The table contains values of exports and imports of country X.

b) Information contained in the body of the table shows exports (sales to) and imports (purchases from) four countries A, B, C and D.

c) These exports and imports occurred in 1995, 1996 and 1997.

d) Information on exports and imports has been arranged according to year and countries.

Dos and Don'ts of the Title

Do not opt for long sentences. Title should be brief and to the point. Present the title in bold letters and/or in capital letters. Expressions used should not convey more than one meaning.

Avoid the expressions like 'Table Presents ....' or 'A Detailed Comparison of Data Relating to ....', etc. It should be like a telegraphic message.

3. **Head note**, also called prefactory note, is written just below the title. It shows contents and unit of measurement like (rupees crore) or (lakh tonnes) or (thousand bales). It should be written in brackets and should appear on right side top just below the title. However, every table does not need a head note, like number of students in each class.

4. **Stubs** are used to designate rows. They appear on the left hand column of the table. Stubs consist of two parts:

   a) *Stub head* describes the nature of stub entry.

   b) *Stub entry* is the description of row entries.

5. **Captions**, also called box heads, designate the data presented in the columns of the table. It may contain more than one column heads, and each column head may be sub-divided in more than one sub-head. For example, we can divide the students of a college into hostelers and non-hostlers and then again into males and females. This will help us to know the number of male hostelers in, say, first year, second year and third year.

6. **Main body of the table**, also called *field* of the table, is its most important and bulky part. It contains the relevant numerical information about which a hint is already contained in the title of the table. In our example of Table 2.15 the title amply suggests that the body of the table contains numerical information on exports and imports of country X for a period of three years.

7. **Foot Note** is a qualifying statement put just below the table (at the bottom). Its purpose is to caution about the limitations of the data or certain omissions. For example in Table 2.15, the foot note reads that "figures are quick estimates" implying that these figures are not final. Similarly in the latest population census data the foot note may be "Excluding the State of Jammu and Kashmir".

8. **Source of data** may be the last part of a table, yet it is important one. It speaks about the authenticity of the data quoted. It also offers opportunity to the reader to check the data if he so desires and get more of it.

Taking all these points into consideration, the format of a hypothetical table is presented below:

Table No. 2.16

(---------------TITLE-----------------)

Head note:

(in rupees crore)

Stub Head	←--------------- Caption ---------------→			
	Column Head I		Column Head II	
	Sub-head	Sub-head	Sub-head	Sub-head
Stub Entries	MAIN	BODY	OF THE	TABLE
Totals				

Foot note(s): Source:

2.3.3 Importance of Tables

Numerical information arranged in tabular form has distinct advantage over other forms of presentation. First, tabulated data are easy to understand and interpret. Secondly, one can make quick comparison between different characteristics, for example, 'Are imports greater than exports over all the three years?' or 'Are exports increasing?' Thirdly, it opens doors for further investigations. Fourthly, they have a more lasting impression on human mind than the textual statements. Needless to say, that the statistical tables are used extensively in almost all fields of human inquiry.

Check Your Progress 2

1. Distinguish between

   a) Caption, stub head and stub entries

   b) One-way and two-way tables

   c) Reference tables and text tables

   d) Column entry and row entry

   e) Head note and foot note

   ....

   ....

   ....

   ....

   ....

2. Comment on the statement: "Title is to the table what heading is to an essay".

   ....

   ....

   ....

   ....

   ....

3. Enumerate the various parts of a Statistical table.

   ....

   ....

   ....

   ....

4. Make a sketch of a two-way table to show the following information:
   a) Division of college according to Ist Year, 2nd Year and 3rd Year students
   b) Hosteler and non-hostelers
   c) Male and female students

   Take hypothetical information.

   ....

   ....

   ....

   ....

2.4 GRAPHICAL PRESENTATION OF DATA

Besides formal tables, statistical data can also be presented in the form of various types of graphs. Graphs are a useful way of conveying information very quickly and briefly. With the same ease and efficiency, they help in comparing data over time and space. They are visual aids and have a powerful impact on the people. It is often said, "*a picture is worth a thousand words*". They attract a reader's attention to what they are supposed to convey about the data. Further, they may help us to estimate some values at a glance, and serve as a pictorial check on the accuracy of our solutions.

However, graphical presentation of data, although useful in different ways mentioned above, is only one method of describing data. This cannot and is not a substitute for other forms of presentation as well as further statistical analysis. In the following, we discuss some of the graphical methods of presentation.

2.4.1 Line Graphs

Although there are four quadrants on a plane, in economics we usually draw our diagrams only in the first quadrant where both the quantities measured on X-axis and Y-axis are positive. Economic quantities like price, quantity demanded and supplied, national income, consumption, production and host of other such variable are non-negative (≥ 0).

Let us take a demand schedule and plot it on the graph. The resultant curve on joining different points, assuming continuity, will give us line graph expressing relation between price and quantity demanded. Such a line graph in Economics is called a *demand curve*. Note that price is measured on Y-axis and quantity demanded on X-axis. The demand curve for data given in Table 2.17 is given in Fig. 2.1.

Table 2.17

Demand Schedule

Price of X (Rs.)	Quantity of X demanded
5	16
10	12
15	8
20	4
25	2
30	1

Table 2.18

Time Series Data

Year	Production of Steel(tons)
1990	10
1991	25
1992	20
1993	40
1994	50
1995	45
1996	60

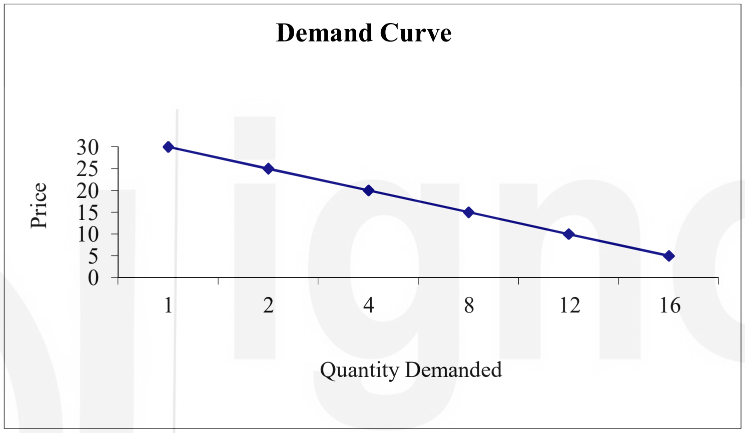


Fig. 2.1

A line graph may be used to show changes in some economic variable, say, steel production over time. In other words, if out of the two variables, one happens to be time (months, years, etc.), we get a line graph over time or simply *time series graph* or *historigram*. A time series expresses behaviour of an economic variable over time. An example of time series data is given in Table 2.18. Measuring years on X-axis and steel production on Y-axis, we can plot time series data on a graph, as shown in Fig.2.2.

Historigram of Production of Steel

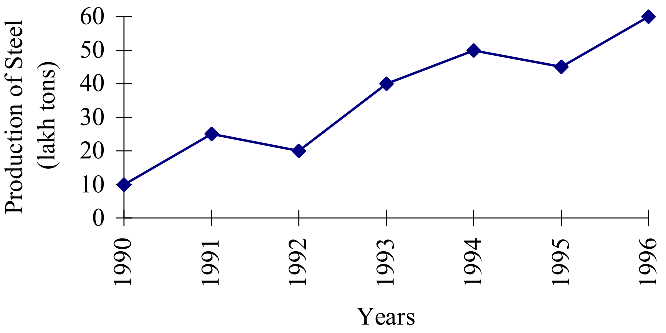


Fig. 2.2

2.4.2 Histogram, Frequency Polygon and Frequency Curve

Histogram (do not confuse with historigram discussed earlier) is a very common type of graph for displaying classified data. It is a set of rectangles erected vertically. It has the following features:

a) It is a rectangular diagram.

b) Since the rectangles are drawn with specified width and height, histogram is a two dimensional diagram. The width of a rectangle equals the class interval and height

$$= \frac{\text{Class frequency} \times \text{Width of the shortest class interval in the data}}{\text{Width of the class interval}}$$

c) The area of each rectangle is proportional to the frequency of the respective class.

Construction of Histogram

To plot a histogram of the frequency distribution given in Table 2.6 on a graph paper, we mark off class intervals like 500 - 550, 550 - 600, etc. on the horizontal axis. Similarly, we mark off frequencies on the vertical axis. Since all the classes are of equal width, the height of each rectangle is taken to be equal to the frequency of the respective class. The histogram is shown in Fig. 2.3.

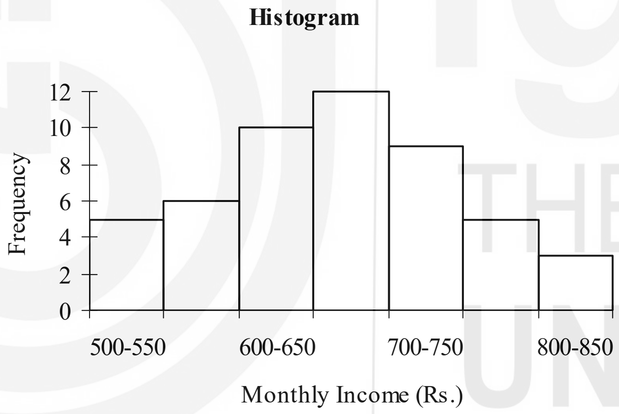


Fig. 2.3

Advantages of histogram are:

1. The width of various rectangles show the nature of classes in the distribution, i.e., whether of equal width or not.
2. Area of a rectangle shows the proportion of the class frequency in the total.

Frequency Polygon

Frequency Polygon has been derived from the word "polygon" which means many sides. In statistics, it means a graph of a frequency distribution. A frequency polygon is obtained from a histogram by joining the mid-points of the top of various rectangles with the help of straight lines, as shown in Fig. 2.4. In order that total area under the polygon remains equal to the area under histogram, two arbitrary classes, each with zero frequency, are added on both ends, as shown below.

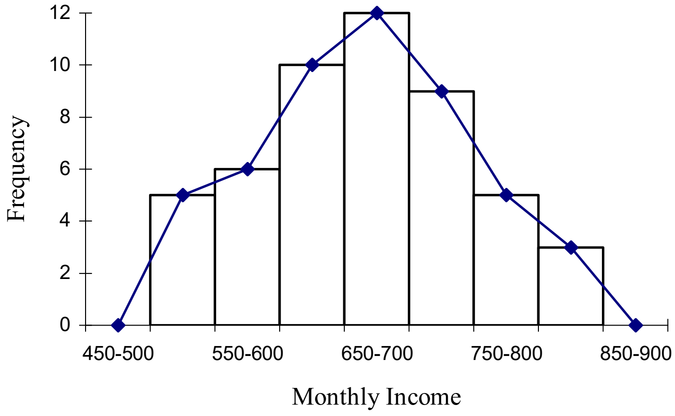


Fig. 2.4: Frequency Polygon

Frequency Curve

If the points, obtained in case of frequency polygon are joined with the help of a smooth curve, we get a frequency curve as shown in Fig. 2.5.

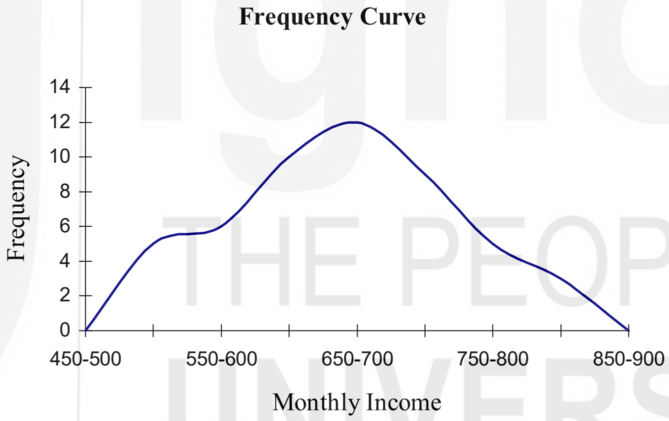


Fig. 2.5

2.4.3 Cumulative Frequency Curve – Ogive

The graph of a cumulative frequency distribution is known as cumulative frequency curve or ogive. Since a cumulative frequency distribution can be of ‘less than’ or ‘greater than’ type, accordingly, we can have ‘less than’ or ‘greater than’ type of ogives.

Ogives can be used to locate, graphically, certain partition values. We can also determine the percentage of observations lying between given limits. The ogives for the cumulative frequency distributions given in Tables 2.12 and 2.13 are drawn in Fig. 2.6.

Note that to draw a less than type ogive, we add a class interval of ‘less than 500’ with frequency equal to zero. Similarly, we add a class interval of ‘more than 900’ with frequency zero for the construction of a greater than type ogive.

'Less than' and 'More than' type Ogives

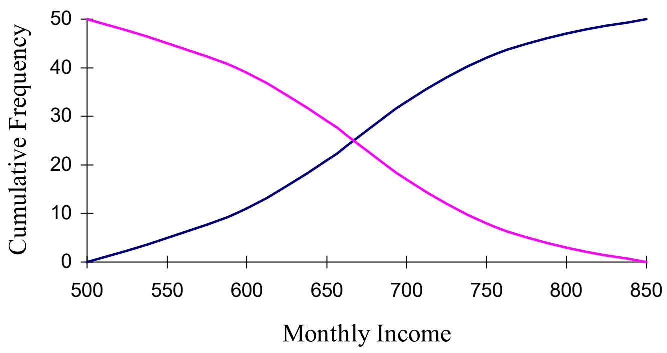


Fig 2.6

2.5 DIAGRAMMATIC PRESENTATION OF DATA

A diagram is a visual form for the presentation of statistical data. Diagram refers to bars, squares, circles, maps, pictorials, cartograms, etc. Diagrams are different from graphs as the former are used only for presentation while the later can be used for analysis in addition to the presentation of data.

2.5.1 One Dimensional Diagrams

These are also known as *bar diagrams*. A *bar* is defined as a *thick line*, often made thicker to attract the attention of a reader. The height of the bar highlights the value of the variable with *width presenting nothing*. Therefore, it has nothing to do with the area of the bar. It is different from the histogram where both the width as well as the height of the bar are important. Further, the bars of the bar diagram are separated from one another so that the gap between the successive bars is same, whereas in histogram they are placed adjacent to one another with out gap. Finally, in histogram the bars are always vertically placed whereas in bar diagram they can be placed both vertically as well as horizontally. Let us take a simple example to demonstrate the construction of a bar diagram.

Table 2.19
Number of students in four zones of a country

Zone	No. of students (lakhs)
North	6
South	10
East	2
West	4

The bar diagram of the above data is drawn in Fig. 2.7. To make the bar diagram beautiful we can either colour the bars or shade them in different ways. This is left to the aesthetic taste of the investigator.

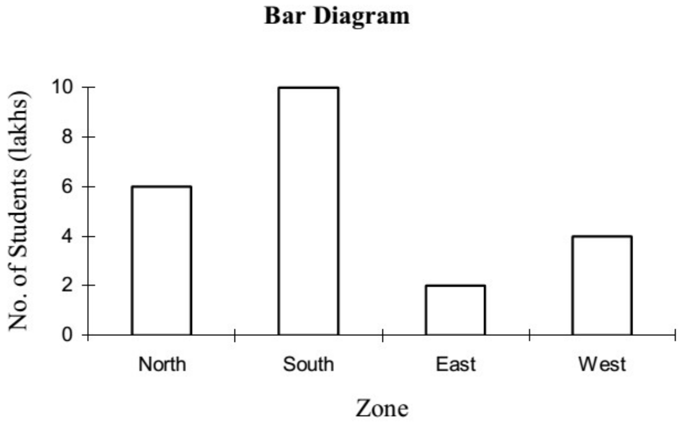


Fig. 2.7

i) Sub-divided or Component Bar Diagram

A sub-divided bar diagram is used when it is desired to represent the comparative values of different components of a phenomenon. In this diagram, the bars, corresponding to each phenomenon, is divided into various components. The portion of the bar occupied by each component denotes its share in the total. The sub-divisions of different bars must always be done in the same order and these should be distinguished from each other by using different colours or shades. A sub-divided bar diagram for the hypothetical data on sales of TV sets, given in Table 2.20 is drawn in Fig. 2.8.

Table 2.20

Zone-wise sale of TV sets (1995-1997)

Zone	Number of T.V. Sets sold (lakhs)		
	1995	1996	1997
North	12	20	28
South	8	9	15
East	5	7	10
West	6	8	11
Total	31	44	64

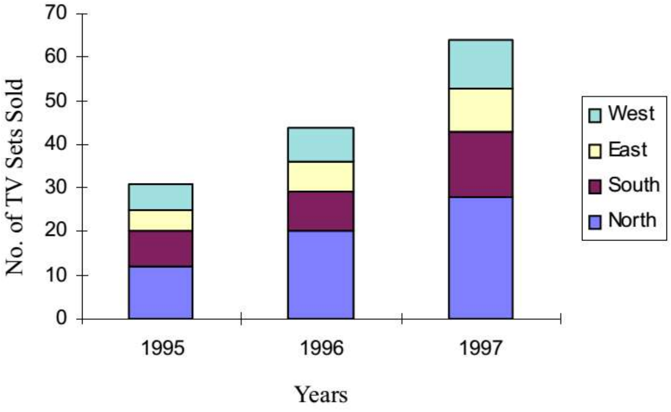


Fig. 2.8: Sub-divided or Component Bar Diagram

ii) Multiple Bar Diagram

This diagram is used when comparisons are to be shown between two or more sets of data. A set of bars for a period, place or a related phenomenon are drawn side by side without gap. Different bars are distinguished by different shades or colours. A multiple bar diagram for the hypothetical data given in Table 2.21 is drawn in Fig. 2.9.

Table 2.21
Total revenue, total cost and profit of M/S XYZ (1990-92)

(Rupees thousand)

Year	Total Revenue	Total cost	Profit
1990	30	25	5
1991	40	35	5
1992	50	40	10

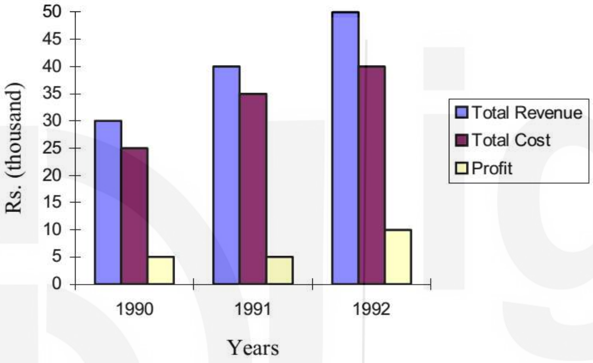


Fig. 2.9

2.5.2 Two Dimensional Diagrams or Area Diagrams

In the case of one dimensional diagrams only the height of the bar is important, and the width can be chosen according to convenience or aesthetic taste of the investigator. But in the case of two dimensional diagrams, area is more important. That is why they are also known as *Area diagrams*. There are three types of area diagrams.

a. *Rectangles*, where area equals width (or base) multiplied by the length (or height) of the rectangle.

b. *Squares* where area equals square of side (or base).

c. *Circles* where area equals πr^2, with π=22/7 and r = radius.

Let us consider data on, say, average salaries of three categories of University teachers, and prepare all the three types of area diagrams.

Table 2.22
Average Salaries of University Teachers as on 1/1/1998

Class of Teachers	Average Salaries (Rs.)
Professors	25,000
Readers	16,000
Lecturers	9,000

a) For drawing rectangles, a common base of, say, 100 is taken. Accordingly, the heights can be determined as:

1.	Salary of Rs.25,000	=	100 (base) × 250 (height)
2.	Salary of Rs.16,000	=	100 (base) × 160 (height)
3.	Salary of Rs. 9,000	=	100 (base) × 90 (height)

Now take a scale of 2 cm = 100, so that the first rectangle has dimensions of 2 cm. × 5 cm, the second one has the dimensions of 2 cm × 3.2 cm and the third one has the dimensions of 2 cm × 1.8 cm. After this, we are in a position to draw the rectangles as area diagrams (Fig. 2.10).

Average Salaries of University Teachers (Rs.)

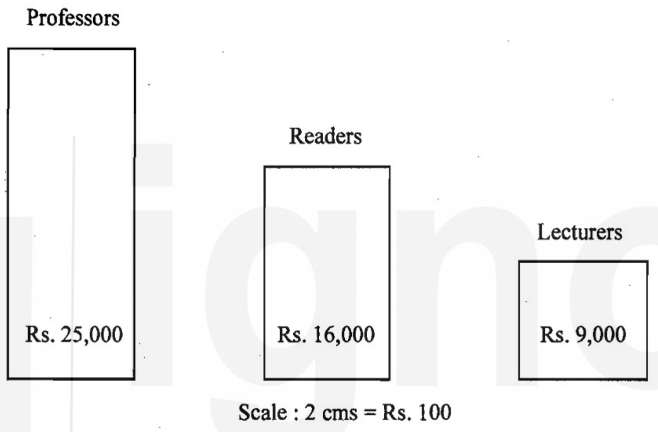


Fig. 2.10

b) For drawing squares, we find the square root of various incomes. We have,

1. $\sqrt{25{,}000} = 158.114$
2. $\sqrt{16{,}000} = 126.491$
3. $\sqrt{9000} = 94.868$

Choose a scale 1 cm = 50 so that the first square has each side approximately equal to 3.2 cm. (since 158.114/50 ≅ 3.2), second has the side of 2.53 cm. and the third has the side of 1.9 cm. The relevant squares are drawn in Fig. 2.11.

Average Salary of University Teachers (Rs.)

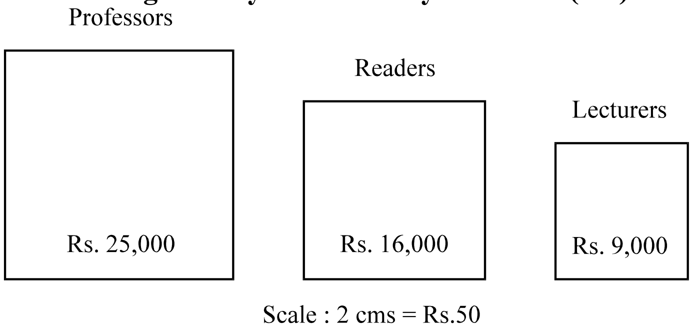
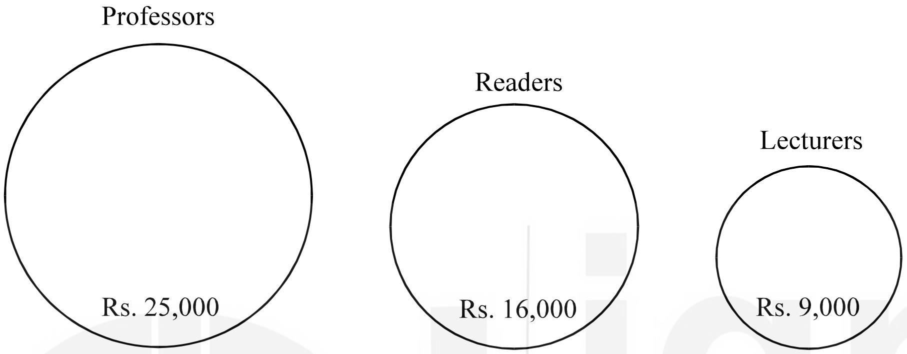


Fig. 2.11

c) For drawing Circles we take the squares of their radii in the ratio of areas, i.e., 25000: 16000: 9000 or 25: 16: 9. This is based on the property of the circles that area of a circle is proportional to the square of its radius. Let r_1, r_2 and r_3 denote the radii of the three circles, then we can write $r_1^2 : r_2^2 : r_3^2 = 25:16:9$ or $r_1 : r_2 : r_3 = 5:4:3$. Taking 1 cm = 2.5 units, the radii of the three circles will be 2.0, 1.6 and 1.2 centimetres respectively. Let us draw the required circles.

Average Salary of University Teachers (Rs.)



Scale : 1 cms = 2.5 unit.

Fig. 2.12

2.5.3 Pie Diagram or Pie Chart

It is also known as angular diagram. It is used to represent percentage break downs of the given data. For example the exports of a country to different countries and continents of the world can be expressed into ratios or percentages. These ratios or percentages can then be converted into angles by the formula

$$\frac{Share\ of\ \ the\ sub-division}{Total} \times 360°$$

Table 2.23

Exports of X to A, B, C and D in 1990

Country	Exports	Percentage Share	Degree
A	300	$(300 \times 100) \div 800 = 37.50$	$(37.5 \times 360^0) \div 100 = 135^0$
B	250	$(250 \times 100) \div 800 = 31.25$	$(31.25 \times 360^0) \div 100 = 12.5^0$
C	150	$(150 \times 100) \div 800 = 18.75$	$(18.75 \times 360^0) \div 100 = 67.5^0$
D	100	$(100 \times 100) \div 800 = 12.50$	$(12.5 \times 360^0) \div 100 = 45^0$
Total	800	100	360^0

Pie Diagram Representing Exports of X

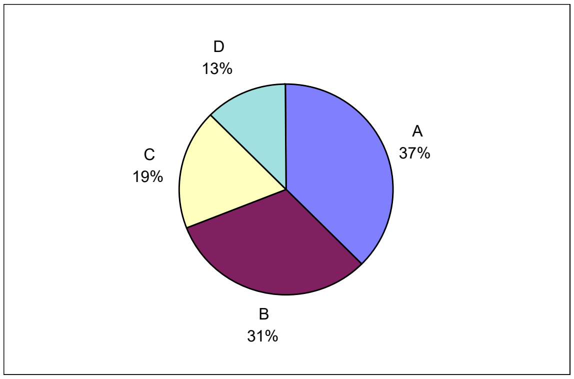


Fig. 2.13

Steps in the construction of Pie diagram

1. Find the total of all components.
2. Find ratio or percentage of the share of sub-division to the total and multiply by 360^0 to get the angle corresponding to the each sub-division.
3. Draw a circle of a suitable size.
4. Use protractor to draw different angles at the centre. Preferably start with the largest one.
5. Shade the different segments with different colours or shades.
6. Write the components with percentage values in the marked, shaded or coloured areas.

2.5.4 Three Dimensional Diagrams

These diagrams are not very popular and are used very rarely. Since these diagrams are three dimensional (involving length, breadth and width), they denote volumes. They can take the form of boxes, cubes, blocks, spheres and cylinders. They are very useful when the variations in magnitudes of the observations are very marked. Here we will explain only the presentation of data by cubes for which we take the following steps:

i) Find cube-root of each figure.
ii) Take a convenient scale, preferably in centimeters.
iii) Draw cubes, dimensions of which are calculated below for an example consisting of two classes of families: Poor and Very Rich.

Table 2.24

	Income class	Income (Rs.)	Cube-root	Side of cube
1.	Poor	216	$\sqrt[3]{216} = 6$	1.5 cms.
2.	Very Rich	3375	$\sqrt[3]{3375} = 15$	3.75 cms.

Scale: 1 cm. = 4 units.

iv) Now draw two cubes with sides equal to 1.5 cms. and 3.75 cms. respectively.

Income levels of Poor and Very Rich people (Rs.)

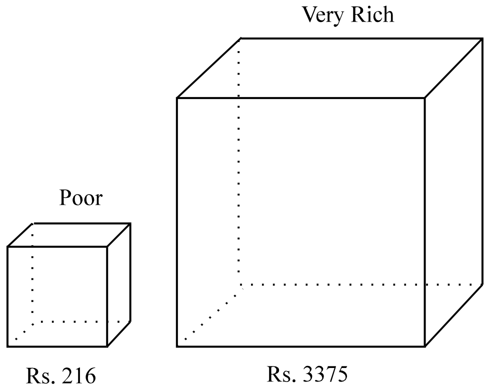


Fig. 2.14

2.5.5 Pictograms and Statistical Maps

These are also known as catrograms. Pictures are more attractive to laymen than other forms of graphic presentations.

But these are not suitable everywhere. It may suit cases involving population of people of a state or number of vehicles in a metropolitan city like Delhi or Bombay. For showing population of human beings, we draw human figures. Here also we have a scale. We may represent 1 lakh people by one human figure so that a population of three and half lakhs is shown by drawing 3 1/2 human figures, as given in Fig. 2.15.

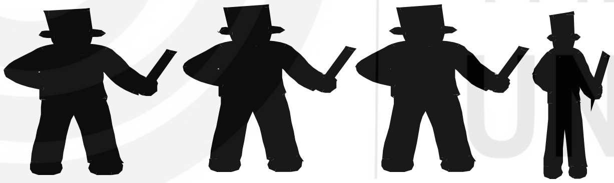


Fig. 2.15

Pictograms suffer from a defect that they present only approximate values. For more accurate presentations bar diagrams are preferable.

Check Your Progress 3

1) Distinguish between the following giving at least two points of distinction.

a) Histogram and historigram. b) Histogram and bar diagram.

c) Histogram and frequency polygon. d)“Less-than”and“More-than” ogives.

e) Pie diagram and circle.

....

....

....

2) Prepare a sub-divided bar chart and a pie diagram from the following data.

Academic Year	Expenditure on Books				
	Economics	Commerce	Maths	Languages	Total
1996 - 97	5200	10000	5000	4800	25000
1997 - 98	8000	14000	7000	6000	35000

....

....

....

....

....

3) Explain the following terms:

a. Line graph

b. Bar diagram

c. Sub-divided or component bar diagram

d. Multiple bar diagram

e. Area diagram

f. Volume diagram

....

....

....

....

....

4) Fill in the blanks with a suitable word out of those given in brackets:

a) A pie diagram is also called ________ diagram. (bar, angular, multiple bar).

b) In the case of vertical bars, the variable is measured on the ________. (X-axis, plane, Y- axis).

c) Bar diagrams, rectangles, squares, circles and pie charts are ________ forms of presenting data. (geometric, arithmetic, horizontal).

d) By joining the mid-points of the top of each rectangle of a histogram, we get ________. (an Ogive, a frequency curve, a frequency polygon)

e) Graph of "more-than" cumulative frequency distribution is also called "more - than" ________. (Ogive, frequency polygon, frequency curve)

f) The caption of a table labels data presented in the ________ of a table. (rows, columns, foot-note)

5) Are the following statements true or false? If false, what should be the correct statement?

a) A picture is worth a thousand words.

b) Squares and circles are examples of area diagrams.

c) We can have only vertical bar to present some data having one variable.

d) The graph of an ordinary frequency distribution is called ogive.

e) A time series graph is known as historigram.

f) Histogram is same as bar diagram.

....

....

....

....

....

2.6 LET US SUM UP

Collected data are unorganised and complex mass of figures. To draw some meaningful conclusions, they must be arranged in an orderly manner. This can be done in many ways, such as by forming simple and frequency array, discrete and continuous frequency distributions, etc.

Sometimes, it serves a useful purpose to form what is called "*less-than*" *or "more-than*" cumulative frequency distributions. Former is formed by successive totaling of frequencies from above and the latter by successive totaling from below.

After collection and condensation of data, good presentation of data is important. A good presentation helps to highlight important points of the data and makes possible useful comparisons and their intelligent use. This can be done through formal tables; line graphs; histograms, frequency polygon and frequency curves; "less-than" and "more-than" ogives; geometric forms – one, two and three dimensional diagrams such as bar diagrams, rectangles, squares, circles, cubes and pie diagrams; statistical maps. While using diagrams, their limitations must always be kept in mind. Diagrams give only a vague idea of the problem and can portray only a limited number of characteristics. Unlike a graphic presentation, the main limitation of a diagrammatic presentation is that it cannot be used as a tool of analysis. The level of accuracy of a graphic method is often lower than that of mathematical method.

2.7 ANSWER OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1. (a) See Sub-Section 2.2.2 and 2.2.3.

(b) See Sub-Section 2.2.4.

(c) See Sub-Section 2.2.3.

(d) See Sub-Section 2.2.1 and 2.2.2.

2. You may give examples from your surrounding. For exact meaning of the terms refer to Section 2.2.
3. In the text we have converted the monthly income data in Table 2.2 to a frequency distribution in Table 2.6. From this you can take a clue.
4. Refer to Sub-Section 2.2.3.
5. Refer to Sub-Section 2.2.4(c).
6. Refer to Sub-Section 2.2.4(d).

Check Your Progress 2

1. Refer to Table 2.16 and Sub-Section 2.3.2 for different parts of a table.
2. Refer to Sub-Section 2.3.2(2).
3. Refer to Table 2.16.
4. It can be presented in more than one ways. We have given one below. Try another.

Division of Students of XY College

Year	Hostelers		Non-Hostelers	
	Male	Female	Male	Female
First Year				
Second Year				
Third Year				

Check Your Progress 3

1. (a) See Sub-Sections 2.4.1 and 2.4.2
   (b) See Sub-Sections 2.4.2 and 2.5.1
   (c) See Sub-Section 2.4.2
   (d) See Sub-Section 2.4.3
   (e) See Sub-Section 2.5.2 and 2.5.3

2. Refer to Sub-Sections 2.5.1 and 2.5.3

3. (a) See Sub-Section 2.4.1
   (b) See Sub-Section 2.5.1
   (c) See Sub-Section 2.5.1
   (d) See Sub-Section 2.5.1
   (e) See Sub-Section 2.5.2
   (f) See Sub-Section 2.5.4

4. (a) angular
   (b) y-axis
   (c) geometric
   (d) a frequency polygon
   (e) ogive
   (f) columns

5. True: 1, 2, 5.
   False: 3, 4, 6.

UNIT 3 SUMMARISATION OF UNIVARIATE DATA[*]

Structure

3.0 OBJECTIVES

After going through this unit, you will be able to:

- compute numerical quantities that measure the central tendency of a set of data such as, mean, median, mode, geometric mean and harmonic mean;
- explain the concept of dispersion;
- compute numerical quantities that measure the dispersion of a set of data;
- explain chebychev's inequality;
- compute the coefficient of variation; and
- find a measure for concentration of certain distribution of data.

[*] Adapted from IGNOU study material of EEC 13: Elementary Statistical Methods and Survey Techniques, Units 4, 5 and 6 written by R S Bharadwaj with modifications by K. Barik

3.1 INTRODUCTION

In the previous Unit we had discussed about condensation of raw data by grouping them into a few class intervals and presenting in the form of a table or diagram. Such tables or diagrams provide a rough idea of the distribution of observations. Often we need to compare between distributions. In such situations it is difficult to compare tables or diagrams simply by looking at them. It is much more convenient and useful for comparison if we could find out a single numerical value for describing the data.

Measures of Central Tendency (or Location) constitute one of the major statistics designed for this purpose. There are five main measures of central tendency. These are Arithmetic Mean, Geometric Mean, Median and Mode. You will learn about each one of these measures below.

3.2 MEASURES OF CENTRAL TENDENCY

In frequency distributions of observations discussed in Unit 2 we notice that the observations tend to cluster around a central value. This phenomenon of clustering around a central value in a frequency distribution is called '*Central Tendency'*. Thus, it is of interest to locate such a value around which clustering of observations takes place. There are several measures of central tendency (or location) of a frequency distribution. These measures produce numbers that summarise a frequency distribution in terms of one its properties, namely, central tendency.

3.2.1 Arithmetic Mean

The *average* or the *arithmetic mean*, or simply the *mean* when there is no ambiguity, is the most common measure of central tendency. It is defined as the sum total of all values in the sample divided by the number of observations. It is denoted by a bar above the symbol of the variable being averaged. Thus $\overline{X}$ stands for the mean of *X*-values in the sample. If in a sample a particular *X*-value, say X_i occurs with frequency $f_i (i = 1,2,...,n)$, its contribution to the total of X-values is f_iX_i. Thus, we can compute the mean of X-values by

$$\overline{X} = \frac{1}{N}(f_1X_1 + f_2X_2 + ... + f_nX_n) = \frac{\sum_{i=1}^{n} f_iX_i}{N}, \qquad \text{where } N = \sum_{i=1}^{n} f_i.$$

When observations are classified into class intervals, as for continuous variables, individual observations falling into a class intervals are not separately identifiable and be contribution of the individual observations from a class intervals to the total cannot be calculated. To avoid this difficulty, it is assumed that every observation falling into a class interval has a value equal to the *mid-point* into which these observations fall. Such a procedure will not give the exact mean had we computed it from raw data and may require what is called corrections for grouping.

Example 3.1: Compute the mean for discrete frequency distribution of Table 3.1.

Table 3.1

Frequency distribution of 100 households by size

Household Size (X_i)	**Frequency (f_i)**
1	3
2	16
3	25
4	33
5	12
6	7
7	2
8	2
Total	100

Let us compute the arithmetic mean of the data given in the above table.

$$\overline{X} = \frac{\sum_{i=1}^{n} f_i X_i}{N} = \frac{1\times3+2\times16+3\times25+4\times33+5\times12+6\times7+7\times2+8\times2}{100} = \frac{.374}{100} = 3.74$$

Thus, mean household size based on 100 households is 3.74.

Example 3.2: Compute the mean for grouped frequency distribution of Table 3.2.

Table 3.2

Frequency distribution of 100 households by average monthly household expenditure on food

Expenditure class (Rs.)	**Frequency**
262.5 – 286.5	1
286.5 – 310.5	14
310.5 – 334.5	16
334.5 – 358.5	28
358.5 – 382.5	26
382.5 – 406.5	15
Total	**100**

For computation of the mean we have to construct table as given below.

Class interval (Rs.) (1)	**Mid-point (X_i) (2)**	**Frequency (f_i) (2)**	**f_i X_i (4)**
262.5 -286.5	274.5	1	274.5
286.5 – 310.5	298.5	14	4179.0
310.5 – 334.5	322.5	16	5160.0
334.5 – 358.5	346.5	28	9702.0
358.5 – 382.5	370.5	26	9633.0
382.5 – 406.5	394.5	15	5917.5
Total		**100**	**34866.0**

Thus, mean of monthly average household expenditure on food is

$$\overline{X} = \frac{34866}{100} = \text{Rs.}348.66.$$

We should note from the above example that to find column (3) we need to multiply the corresponding values of column (1) and (2), and often hand computations are long for each multiplication. These computations can be simplified, particularly when successive column (1) values are equidistant (but applicable otherwise also), by making the following simple transformation.

For *i* = 1, 2, …, n

$$u_i = \frac{X_i}{h} \qquad \text{i.e., } X_i = A + hu_i \text{ and so } \overline{X} = A + h\overline{u}.$$

Often *A* is called the 'assumed mean' and $h\overline{u}$ as its correction to get $\overline{X}$. Choice of *A* and *h* are made so that computation of $\overline{u}$ becomes simple. Usually *A* is take as that *X* value for which the frequency is largest. For equidistant successive *X*-values is column (1), h may be taken as the difference between two successive *X*-values. For equal length class intervals, the difference between successive mid-points is the same as the length of each class interval.

We will explain this method by re-computing the mean of the monthly average household food expenditure data given in Table 3.2. We construct Table 3.3 by using *A* and *h* as explained below.

We define A = Mid-point of the class with largest frequency = 346.5 and

h = Common length of each class interval = 24.

Thus, $$u_i = \frac{X_i - 346.5}{24}$$

Table 3.3

Computation of Mean of Frequency Distribution of Table 3.2

Class interval (Rs.)	**Mid-point (X_i)**	$u_i = \frac{X_i - 346.5}{24}$	**frequency (f_i)**	**f_i u_i**
262.5 – 286.5	274.5	– 3	1	–3
286.5 – 310.5	298.5	–2	14	–28
310.5 – 334.5	322.5	–1	16	–16
334.5 – 358.5	346.5	0	28	0
358.5 – 382.5	370.5	1	26	26
382.5 – 406.5	394.5	2	15	30
Total			**100**	**9**

We find out that

$$\bar{u} = \frac{1}{N}\sum_{i=1}^{n} f_i u_i = \frac{1}{100} \times 9 = \frac{9}{100}$$

Thus, $X = A + h \times \bar{u} = 346.5 + 24 \times \frac{9}{100} =$ Rs.348.66 as was computed earlier.

Properties of Arithmetic Mean

1) *The algebraic sum of deviations of a given set of observations is zero when taken from the arithmetic mean.*

   Let $X_1, X_2, ..., X_n$ be n observations with respective frequencies as $f_1, f_2, ..., f_n$. Mathematically, this property implies that $\sum_{i=1}^{n} f_i (X_i - \bar{X}) = 0,$ where $X_I - \bar{X}$ is the deviation of ith observation from mean. To prove the above property, we write

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

   Hence, the result,

2) The sum of squares of deviations of a given set of observations is minimum when taken from the arithmetic mean.

   Mathematically, this property implies that for any arbitrarily chosen origin, A,

$$S = \sum_{i=1}^{n} f_i (X_i - A)^2 \text{ is minimum when } A = \bar{X}.$$

To prove this property, we note that the magnitude of *S* will depend upon the selected value of *A*. thus, we can say that *S* is a function of *A*.

We want to find that value of *A* for which *S* is minimum. Using calculus, this value is given by the equation $\frac{dS}{dA} = 0$ such that $\frac{d^2S}{dA^2} > 0.$

(Remember that the value of a function is minimum when first derivative is zero and second derivative is positive.)

Differentiating *S* with respect to *A* and equating to zero, we get

$$\frac{dS}{dA} = -2\sum_{i=1}^{n} f_i (X_i - A) = 0$$

This implies that

$$\sum_{i=1}^{n} f_i X_i - A\sum_{i=1}^{n} f_i = 0 \quad \text{or} \quad A = \frac{\sum_{i=1}^{n} f_i X_i}{\sum_{i=1}^{n} f_i} = \overline{X}.$$

Further, it can be shown that $\frac{d^2S}{dA^2} > 0$ when $A = \overline{X}.$

3.2.2 Median

Median of a distribution locates a central point which divides a distribution into two equal halves, i.e., it is the middle most value among a set of observations. Let us start with examples in a discrete case. Consider a data set having 5 distinct observations: 2, 4, 9, 12, 19 (arranged is ascending order). Here 9 is the middle most value since an equal number of observations are to its left and to its right. Thus, 9 is the median of the above observations. Consider another data set having 6 distinct observations: 3, 8, 15, 25, 35, 43. Here any point between 15 and 25 has the property that equal number of observations are to its left and to its right. Any point in the interval 15 to 25 may be used as a median. Conventionally we take the middle point of such an interval to define median uniquely. Thus 20 is the median of 3, 8, 15, 25, 35, 43.

When a data set has non-distinct observations – a situation more common in practice – difficulties' may arise. In such situations, it may not be always possible to locate the middle most value or the central point that divides the distribution into two equal halves, For example, in the case of the data set having 5 observations 2, 9, 9, 12, 19 the value 9 is repeated twice. Thus, a form definition of median is needed to overcome such difficulties.

A median of a distribution is a point or a central value such that ***at least*** *50% of the observations are less than or equal to it and* ***at least*** *50% of the observations are greater than or equal to it*. With this definition of median and the convention of taking the middle point of a class in which each point i a median, median of a distribution can always be specified uniquely. Thus, median of observations 2, 9, 9, 12, 19 is 9 because 3 of the 5 observations (60%) are less than or equal to 9 and 4 of the 5 observations (80%) are greater than or equal to 9.

Let us find out the median household size from the frequency distribution in Table 3.1. We notice that 77 (out of 100) households have family size of less than or equal to 4 and 56 households have family size of more than or equal to 4. Thus median family size in this case is 4.

Median for a grouped frequency distribution of a continuous variable is easier to understand if we look at the associated histogram with height of a rectangle equal to the frequency density, $\frac{f}{h}$, of the class. In such a histogram, the area of a rectangle gives the frequency of the corresponding class. The median, in this case, is a point in one of the classes such that the areas to its left and to its right are 50% each. First step is to locate the class, up to the right boundary of which the total areas is at least 50% (*called the median class*). Then the median is computed by adding, to the lower boundary value of this class, the length of a part of this class interval in proportion to the frequency needed to achieve 50%. A convenient method of finding out the median class is to compute the cumulative frequency (discussed in Unit 2, Section 2.3.3) and identifying the class interval in which the $\frac{N}{2}$th observation lies.

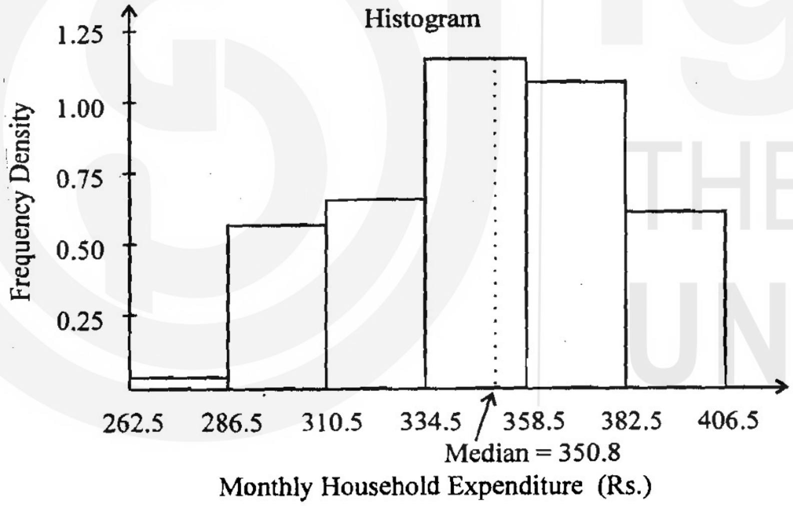


Fig. 3.1

Area up to the class boundary 334.5 is 131 and up to 358.5 is 59. Hence the median lies in the class 334.5 – 358.5. We now want to find a point in this classes so that the area from 334.5 to the point is (50 – 31) = 19, where are up to 334.5 is 31. Since the rectangle over the interval 334.5 – 358.5 has an area of 28, and is of length 24, to get an area of 19 we need $\frac{19}{28}$th part of 24. This works out to be $\frac{19}{28} \times 24 = 16.3$. Thus the median is 334.5 + 16.3 = 350.8. Note also that the area in the class 350.8 to 358.5 is 28 – 19 = 9 and to the right of 350.8 is 9 + 41 = 50, as it should be.

Based on the above procedure, we can write a formula for the computation of median.

$$M_d = l_m + \frac{\frac{N}{2} - C}{f_m} \times h, \text{ where}$$

l_m is the lower limit of the median class, i.e., the class in which median lies,

N is the total frequency,

C is the cumulative frequency of classes preceding the median class (not that C = 31 in the above example).

f_m is the frequency of median class, and

h is the width of median class.

3.2.3 Mode

As has been pointed out earlier, often observations tend to cluster around a central value. A simple measure of this phenomenon is called mode.

Mode or modal value of a discrete variable is defined as that value of the variable for which frequency is the maximum. Mode, however, is not the majority, i.e., it does not imply that most (50% or more) of the observations have the modal value.

From Table 3.1 we find that the mode or modal value of household size is 4 as this value occurs with largest frequency of 33 among 100 households.

There are, however, data sets when mode cannot be defined uniquely, i.e., the distribution has multiple mode. Raw data with 7 hypothetical observations with values 4, 3, 4, 1, 2, 5, 3 have two modes, 3 and 4. Distributions having two modes are called *bimodal distributions*, though the frequently encountered distributions have only one mode or are *unimodal*.

For observations on the continuous variable, like monthly household expenditure on food, no two observations are likely to have same value and so mode is not a meaningful measure of such raw data. However, central tendency comes out clearly when these raw data are grouped into various class intervals. For grouped data *modal class* is defined as the class having largest frequency. Since large class intervals are likely to include large number of observations and smaller class intervals are likely to have few observations, definition of modal class is meaningful only when class intervals have equal length.

For discrete data it is easier to find out the mode. But in the case of continuous data computation of the mode is done by the following formula:

$$M_0 = l_m + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times h, \text{ where}$$

l_m is the lower limit of the modal class, i.e., the class in which mode lies,

$\Delta_1 (= f_m - f_{m-1})$ is the difference of the frequencies of the modal class and its preceding class.

$\Delta_2 (= f_m - f_{m+1})$ is the difference of the frequencies of the modal class and its following class, and

h is the width of the modal class.

Let us look back to Table 3.2. Here modal class is 334.5 – 358.5 as it has the highest frequency, 28.

Thus, $l_m = 334.5,\ \Delta_1 = 28 - 16 = 12, \Delta_2 = 28 - 26 = 2$ and $h = 24$.

Hence $M_0 = 334.5 + \frac{12}{12+2} \times 24 = 355.07$

Mode is a useful measure of central tendency when a frequency distribution has a strong peak and it is particularly useless when a frequency distribution is almost flat.

Check Your Progress 1

1) The frequency distribution of a family size for 250 families in a ward of an industrial town is given below:

   Find the mean, median and mode.

Family Size	Frequency
1	4
2	22
3	25
4	45
5	52
6	41
7	36
8	15
9	7
10	3
Total	**250**

……………………………………………………………………………………......

……………………………………………………………………………….........

……………………………………………………………………………....…..

……………………………………………………………………………….........

2) Compute the mean, median and mode for the following frequency distribution.

I.Q.	Frequency
160 – 169	2
150 – 159	3
140 – 149	7
130 – 139	19
120 – 129	37
110 – 119	79
100 – 109	69
90 – 99	65
80 – 89	17
70 – 79	5
60 – 69	3
50 – 59	2
40 – 49	1
Total	**309**

………………………………………………………………………………………......

………………………………………………………………………………………........

………………………………………………………………………………………......…

……………………………………………………………………………………….........

……………………………………………………………………………………….........

………………………………………………………………………………………....…..

3.3 OTHER MEASURES OF CENTRAL TENDENCY

Besides the arithmetic mean, median and mode there are other averages which are relatively unimportant but may be appropriate in particular situations. These are Geometric Mean and Harmonic Mean.

Often we see that all the observations do not have equal importance. In such cases we need to give differential importance to different items. Here we use weighted means – arithmetic, geometric or harmonic – instead of simple means. This we will discuss in Section 3.3.2.

3.3.1 Geometric Mean and Harmonic Mean

Often we have to deal with that are time dependent, i.e., time series data which are unlike one-time data of Tables 3.1 and 3.2. For time dependent data, it is often of interest to find the pattern of change over time. Consider the following two data sets.

Set I: 1000 1100 1200 1300 1400 1500 1600

Set II: 1100 1210 1331 1464 1611 1772 1949

The first set looks like the basic salary (in Rs.) of an employee for 7 years with annual increment of Rs. 100 per year.

The second set looks more like his gross salary (in Rs.). Annual increase in the two sets is given below.

Set I: 100 100 100 100 100 100

Set II: 110 121 133 147 161 177

Arithmetic mean of the annual increase is 100 for Set I and 141.5 for Set II. On the basis of these average annual increases, if we find out the figures for the two sets, starting from the initial values, we would get the following:

Set I: 1000 1100 1200 1300 1400 1500 1600

Set II: 1100 1241.5 1383 1524.5 1666 1807.5 1949

We find that arithmetic mean has worked well for Set I. However, it has not worked well for Set II. It is because the progression of original numbers in the two sets is different. In set I, increment has been a fixed quantum whereas in Set II, figures have increased at a fixed rate. Fixed quantum of increase is called *arithmetic progression* and arithmetic mean is appropriate to describe the increase. Fixed rate of increase is called *geometric progression* and geometric mean is most appropriate to describe the increase.

For *n* numbers $X_1, X_2, ... X_n$ the geometric mean (GM) is defined as the nth root of the product of these n numbers, i.e.,

$$GM = (X_1, X_2, ..., X_n)^{\frac{1}{n}} = \left[\prod_{i=1}^{n} X_i\right]^{\frac{1}{n}}$$

Clearly, GM is not defined unless all the n numbers are positive. If any number is negative or zero, we cannot calculate GM. By taking logarithm of GM, we have

$$\log GM = \left(\frac{1}{n}\right)\left(\log X_1 + \log X_2 + ... + \log X_n\right) = \frac{1}{n}\sum_{i=1}^{n} \log X_i$$

which shows that now GM can be computed by using a log-table. Anti-logarithm of the arithmetic mean of log X values is GM. For the second data set, gross salary increased at the rate of 11% every year. In practice, however, increase/decrease will not be at a fixed rate over the years; and it is meaningful to talk about average rate because fixed rate situation is rare.

In general, GM is more appropriate average for percentage (or proportionate) rates of change than arithmetic mean as in the case of rise in various price indices, cost of living indices, etc.

Finally, we discuss about another measure of location called the 'harmonic mean' (HM). This measure of central tendency comes naturally in many situations as in the following illustration. A stockist stocks Rs. 5000 worth of an item at the beginning of every month. Unit rate (in Rs.) of the item for five successive months had been 10.75, 11.80, 14.00, 11.45. and 12.00. The stockist wants to find average rate per unit of the item he has stocked for five months. Computation is presented below:

Month	**Amount Spent (Rs.)**	**Unit Rate (Rs.)**
1	5000	10.75
2	5000	11.80
3	5000	14.00
4	5000	11.45
5	5000	12.00
Total	**25000**	

Average price (in Rs. of his entire stock $= \dfrac{\text{Total Money Spent}}{\text{Total Quantity Purchased}}$

$$= \frac{5 \times 5000}{\frac{5000}{10.75} + \frac{5000}{11.80} + \frac{5000}{14.00} + \frac{5000}{11.45} + \frac{5000}{12.00}}$$

$$= \frac{5}{\frac{1}{10.75} + \frac{1}{11.80} + \frac{1}{14.00} + \frac{1}{11.45} + \frac{1}{12.00}}$$

$$= \frac{1}{\frac{1}{5}\left(\frac{1}{10.75} + \frac{1}{11.80} + \frac{1}{14.00} + \frac{1}{11.45} + \frac{1}{12.00}\right)} = 11.91$$

The last expression is 'the reciprocal of the arithmetic mean of the reciprocals' and is called harmonic mean (HM). For a set of *n* values $X_1, X_2, ..., X_n$, the HM is defined as

$$\text{HM} = \frac{n}{\frac{1}{X_1} + \frac{1}{X_2} + ... + \frac{1}{X_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}}$$

You should note that HM is not defined when any observation is zero.

If the stockiest, instead of stocking Rs. 5000 worth of items, stocks 3000 items at the beginning of every month at the given prices, the appropriate average would be arithmetic mean. To verify this, we can write

$$\text{Average Price } = \frac{\text{Total Money Spent}}{\text{Total Quantity Purchased}}$$

$$= \frac{3000\times 10.75 + 3000\times 11.80 + 3000\times 14.00 + 3000\times 11.45 + 3000\times 12.00}{3000\times 5}$$

$$= \frac{10.75 + 11.870 + 14.00 + 11.45 + 12.00}{5} = \text{AM of the given prices.}$$

3.3.2 Weighted Means

For many practical applications weighted means (arithmetic, geometric or harmonic) reflect phenomenon more clearly than unweighted or simple means that have been computed so far. For computation of, say, consumer price index, not all commodities are equally important. Increase in fuel cost may affect consumer price index more than an increase in agricultural prices. For stock market, stock of some key companies may be a trend setter. Weighted means are more appropriate in such situations. To find weighted mean, a weight w_i is attached to each X_i and the means are computed as if wi's are, symbolically, frequencies of the corresponding X_i's. The computational formulae are as given below:

$$\text{Weighted AM} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

$$\text{Weighted GM} = \left(\prod_{i=1}^{n} X_i^{w_i}\right)^{\frac{1}{\sum w_i}} \quad \text{and}$$

$$\text{Weighted HM} = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{X_i}}$$

Weighted mean is equal to unweighted mean when each w_i is the same or equal to unity.

3.3.3 Pooled Mean

Often we come across situations when the means have been computed for different sources or samples. In such situations we become interested to find an overall mean if it is meaningful. This is done by computing what is called a *pooled mean*. The procedure of computing a pooled mean is given below.

Let $m_1, m_2, ..., m_r$ be *r* arithmetic (or geometric or harmonic) means, computed on the basis of $n_1, n_2, ..., n_r$ observations respectively. Then

$$\text{Pooled arithmetic mean } = \frac{1}{n}\sum_{i=1}^{r} m_i n_i, \text{ where } n = \sum_{i=1}^{r} n_i$$

$$\text{Pooled geometric mean } = \left(\prod_{i=1}^{r} m_i^{n_i}\right)^{\frac{1}{n}} \text{ and}$$

$$\text{Pooled harmonic mean } = \frac{n}{\sum_{i=1}^{r}\frac{n_i}{m_i}}$$

where $n = n_1 + n_2 + ... + n_r$

Note that the above expressions are similar to the expressions for weighted means.

3.3.4 Choosing a Measure of Central Tendency

It has already been discussed when a particular mean, AM or GM or HM, is more appropriate than the other two. However, when we have grouped data in which either of the end classes are open ended, i.e., of the type 'up to c_1' and / or c_{k-1} and above', mid-points of such classes cannot be computed. Consequently, no mean can be computed. There is, however, no problem in computing median or mode in such cases. On the other hand, a pooled median or mode cannot be computed, like the case for mean, unless all the sets of data are made available in their entirety. These problems are related to computational difficulties and not to appropriateness of a measure.

Since graphical representation of data is more appealing, median or mode are more useful in such a situation because their crude values can be obtained easily without having to go through any computations. Also, median and mode are simple concepts for communication and comparison between graphs. It has, however, been observed that median is less stable than arithmetic mean in repeated sampling and we need to be careful when comparing graphs.

For data that has a distribution close in shape to what is called the normal distribution, with one peak and going down symmetrically on either side, we may use of mean, median or mode. It is because, for a normal distribution, these measures have the same value.

You should note that choosing an appropriate measure of central tendency is not an end to data analysis, and much still remains. For example, by saying that household average monthly expenditure on food is Rs. 348.66, it does not say whether a large number of households have very low monthly average expenditure on food or a few households have a very good menu. Next set of analysis aims at answering such questions.

3.4 PERCENTILES

Concept of percentiles will be explained by using mainly Table 3.2 data on average monthly household expenditure. Percentiles are used in two directions, depending on the question to be answered. Direction of a question may be, what per cent of households have monthly average food expenditure upto Rs. 350.80? Or it may be, what is the maximum monthly average food expenditure of the lower 50% of the households? Note, from our earlier computation of median of Table 3.2 distribution, that the answer to one question is the figure in other, i.e., 50% of the households have Rs. 350.80 as maximum average monthly food expenditure. Depending on interest, percentage below a cut-off point may be called for: when a poverty line is decided, it is of interest to know the percentage below the poverty line. In the other direction, it may also be of interest to find the status of lower 10% or upper 5% of the population. These are answered by using what are called percentiles.

3.4.1 Percentile: Definition and Computation

For any given percentage v, the v^{th} percentile is P_v, a value of the variable being studied, so that at leasat v percent of the observations are less than or equal to P_v and at least (100 – v) percent of the observations are greater than or equal to P.

For example, for Table 3.1, distribution of household size, $P_v = 5$ for any from 78 to 79.

For grouped data, percentiles are more clearly understood when we look at the cumulative distribution function. Let F(*X*) be the proportion of observations less than or equal to *X*. Any given value X_0 is then the 100 $F(X_0)$th percentile. For Table 3.2, class boundaries, we have F(286.5) = 0.01, F(310.5) = 0.15, F(334.5) = 0.31, F(358.5) = 0.59 and F(382.5) = 0.85, and consequently Rs. 286.5 = P_{10}, Rs. 310.5 = P_{15}, Rs 334.5 = P_{31}, Rs. 358.5 = P_{59}, and Rs. 382.5 = P_{85}.

You should note that any amount less than Rs. 262.5 (lower boundary of first class interval) is zero-th percentile and any amount more than Rs. 406.5 (upper boundary of last class interval) is 100th percentile.

3.4.2 Quartiles and Deciles

Depending on its use, some specific percentiles go by different names. Every 25th percentile is called a quartile, and every 10th percentile is called a decile. For example,

$$25^{th} \text{ percentile } = P_{25} = Q_1 = \text{ first quartile}$$

$$50^{th} \text{ percentile } = P_{50} = Q_2 = \text{ second quartile}$$

$$75^{th} \text{ percentile } = P_{75} = Q_3 = \text{third quartile}$$

$$10^{th} \text{ percentile } = P_{10} = d_1 = \text{first decile}$$

$$20^{th} \text{ percentile } = P_{20} = d_{21} = \text{ second decile, etc., and}$$

$$P_{50} = Q_2 = d_5 = \text{median}$$

The formulae for Q_1 and Q_2 are similar to the formula for the median. These can be directly written as given below.

$$Q_1 = l_{Q_1} + \frac{\frac{N}{4} - C}{f_{Q_1}} \times h, \text{ and}$$

$$Q_{31} = l_{Q_3} + \frac{\frac{3N}{4} - C}{f_{Q_3}} \times h,$$

where *C* denotes the cumulative frequency of classes preceding the first (or third quartile class and *h* is the corresponding class width.

Using similar notations, it is possible to write the formula for any partition value. For example, the formula for 40th percentile can be written as

$$P_{40} = l_{P_{40}} + \frac{\frac{40N}{100} - C}{f_{P_{40}}}$$

Percentiles also go by the name of fractiles when proportions, instead of percentage, are used. For example, P_{30} is 0.3 fractile.

Just as we do not get a complete picture of a distribution by looking at a measure of location, too many percentiles may be needed to describe the spread or dispersion of a distribution. It is felt that there should be some simple measures of dispersion. This is the topic of discussion of the next Section.

Check Your Progress 2

1) Given below are the prices is ratios for five commodities with the corresponding weights. Calculate the Weighted Arithmetic Mean and Geometric Mean.

Commodity	Price Ratio	weight
1	2.20	30
2	1.85	25
3	1.80	22
4	2.05	13
5	1.75	10

…………………………………………………………………………………......…

…………………………………………………………………………………........

…………………………………………………………………………………........

…………………………………………………………………………………........

…………………………………………………………………………………...........

…………………………………………………………………………………......…

2) The earnings of five nationalised banks, in crores of rupees, is given below.

217.40	330.50	682.55	1263.59	2249.63

Find the Geometric Mean of the earnings.

…………………………………………………………………………………......…

…………………………………………………………………………………........

…………………………………………………………………………………........

…………………………………………………………………………………........

…………………………………………………………………………………...........

…………………………………………………………………………………......…

3) The distribution of age of males at the time of marriage was as follows:

Age (years)	**No. of Males**
18 – 20	5
20 – 22	18
22 – 24	28
24 – 26	37
26 – 28	24
28 – 30	22

Find at the time of marriage (i) the average age, (ii) modal age, (iii) the median age, (iv) third quartile, (v) sixth decile, (vi) nineteenth percentile.

…………………………………………………………………………………....…....

…………………………………………………………………………………..........

…………………………………………………………………………………..............

…………………………………………………………………………………..........

4) In a factory, a mechanic takes 15 days to fabricate a machine, the second mechanic takes 18 days, the third mechanic takes 30 days and the fourth mechanic takes 90 days. Find the average number of days taken the workers to fabricate the machine. Which average would you use, and why?

…………………………………………………………………………….........

……………………………………………………………………....……......

……………………………………………………………………….......…..

…………………………………………………………………………........

……………………………………………………………………....….......

5) The amount of interest paid on each of the three different sums of money yielding 10%, 12% and 15% simple interest per annum are equal. What is the average yield percent on the total sum invested?

....

…………………………………………………………………………….........

…………………………………………………………………………….........

…………………………………………………………………………........

…………………………………………………………………………….........

3.5 MEASURES OF DISPERSION

So far we have discussed various measures of central tendency, viz., arithmetic mean, median, mode geometric mean and harmonic mean. However, in many situations these measures do not represent the distribution of data. For example, look into the following three sets of data:

Set A: 2, 5, 17, 17, 44.

Set B: 17, 17, 17, 17, 17.

Set C: 13, 14, 17, 17, 24.

In all the sets the numerical value of the mean, median and mode are the same, that is, 17. Still all three sets are so different! While in Set B all the observations are equal, in Set A they are so dispersed. Definitely we need another measure which will account for such dispersion of data.

The word dispersion is used to denote the degree of heterogeneity in the data. It is an important characteristic indicating the extent to which observations vary amongst themselves. The dispersion of a given set of observations will be zero when all of them are equal (as in Set B given above). The wider the discrepancy from one observation to another, the larger would be the dispersion. (Thus dispersion in Set A should be larger than that is Set C). A measure of dispersion is designed to state numerically the extent to which individual observations vary on the average. There are quite a few measures of dispersion. We discuss them below.

3.5.1 Range

Of all measures dispersions, range is the smallest. It is defined as *the difference between the largest and the smallest observations*. Thus for the data given at Set A the range is 44 – 2 = 42. Similarly, for Set B the range is 17 – 17 = 0 and for Set C it is 11. Now let us look into some grouped data. For Table 3.2 data (look back to the previous Unit), the range is Rs. 406.5 – Rs. 262.5 = Rs. 144. Notice that, for grouped data, largest and the smallest observations are not identifiable. Hence we take *the difference between two extreme boundaries of the classes*.

It is intuitive that, because of central tendency, if one selects a small sample, observations are more likely to be around its mode than away from it. Less likely or extreme values will be included in the sample when its size is large. This, in other words, implies that range will increase with increase in sample size. Also, it is known that in repeated sampling with same sample size, range varies considerably making it a less suitable measure for comparisons. However, range is a measure which is easy to understand and can be computed quickly.

3.5.2 Inter-quartile Range

Range as a measure of dispersion does not reflect a frequency distribution well, as it depends on the two extreme values. Even one very large or small observation, away from general pattern of other observations in the data set, makes the range very large. For example, in Set A, the range is found to be excessively large (44 – 2 = 42) because of the present of very large one observation, that is 44. To avoid such extreme observations, particularly when there is a strong central tendency, inter-quartile range is useful as a measure of dispersion. It is defined as

Inter-quartile Range $= Q_3 - Q_1 = P_{75} - P_{25}$.

Inter-quartile range is the range of the middle most 50% of the observations. If the observations are compact around median, i.e., a strong mode close to the median exists inter-quartile range will be smaller than half of the range. If the data are flat, having no central tendency, this measure will be large, and its value will be close to half of the range.

Let us look into the discrete data given in Table 3.1 of the previous Unit. Here, $P_{75} = 4$ and $P_{25} = 3$. Hence, the inter-quartile range of household size is 4 – 3 = 1. This shows that a strong central tendency exists in the distribution of household size range was observed to be 7 (since 8 – 1 = 7).

For Table 3.2 data, P_{25} of the average monthly expenditure on food was seen to be Rs. 325.50; P_{75} computed similarly works out to be Rs. 377.88 and inter-quartile range is Rs 377.88 – Rs. 325.50 = Rs. 52.38. Compared to Rs. 52.38, the range was observed to be Rs. 146.00 or 2.79 times larger. This shows not so strong central tendency for average monthly household expenditure on food.

3.5.3 Mean Deviation

While the range depends on the two extreme observations, inter-quartile range depends on the two extreme observations among the middle most 50 percent of the observations. Thus, one talks only about the percentage of observations between minimum, P_{25} and maximum, P_{75}. Thus both range and inter-quartile range do not depend upon all the observations in the sample. Hence, while computing range or inter-quartile range we do not say anything about the distribution of observations within the group.

Among many possibilities to quantify spread or dispersion of observations, one possibility is to used the deviation of observations from some central value.

Since mean is the most commonly used measure of central tendency, it is often taken as the central value with reference to which the deviations are computed. These deviations are then suitably combined to a get a measure of dispersion.

Mean deviation treats every single observation with equal weight, in the form of arithmetic mean of deviations based on each observation.

For observations $X_1, X_2, ..., X_n$, if we take deviations as simple difference, then for the i^{th} observations the deviation is $(X_i - \overline{X})$ where $\overline{X}$ is the mean. Mean of these deviations is

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

Since simple difference do not lead to any measure, absolute differences are used to define mean deviation.

$$\text{Mean Deviation } = \frac{1}{n}\sum_{i=1}^{n}\left|X_i - \overline{X}\right|, \text{ where}$$

the two vertical bars indicate that the sign of the difference with in two bars is to be taken as positive. For example, $|2 - 4| = 4$ (and not –2).

For frequency data, discrete or continuous type, the formula becomes

$$\text{Mean Deviation } = \frac{1}{N}\sum_{i=1}^{n} f_i\left|X_i - \overline{X}\right|,$$

where $N = \sum_{i=1}^{n} f_i$ and X'_i s are distinct observations and f_i is the frequency of X_i in the discrete case and X_i is the mid-point of i^{th} class and f_i is its frequency for the continuous case. The need for such a measure is illustrated below.

Following summary values have been computed for two data sets.

	Data Set I	**Data Set II**
Number of observations	7	7
P_{25}	7	7
Median $= P_{75}$	12	12
P_{75}	17	17
Range	20	20
inter-quartile range	10	10
Mean	12	12

Thus, based on the above measures only, and not looking at the data sets I and II, it would appear that two persons separately may have worked out on the same data set. However, the two data sets may have been as given below.

Data Set I:	3	7	8	12	14	17	23
Data Set II:	2	7	11	12	13	17	22

One may construct much more different looking data sets having identical values for the above type of measures. This comparison indicates that more measures are needed and mean deviation is one such. This is not to imply that the above measures and mean deviation together completely describe a data set.

For data set I

Mean deviation =

$$\frac{1}{7}\left(|3-12|+|7-12|+|8-12|+|12-12|+|14-12|+|17-12|+|23-12|\right)$$

$$=\frac{9+5+4+0+2+5+11}{7}=\frac{36}{7}=51.4$$

For data set II

Mean deviation =

$$\frac{1}{7}\left(|2-12|+|7-12|+|11-12|+|12-12|+|13-12|+|17-12|+|22-12|\right)$$

$$=\frac{10+5+1+0+1+5+10}{7}=\frac{32}{7}4.57.$$

Thus, observations in data set I are more dispersed from mean than that of data set II.

Let us now compute mean deviation of household size and household average monthly food expenditure.

For household size data of Table 3.1, mean $= \overline{X} = 3.74.$ Mean deviation is now computed as

$$\text{Mean deviation } = \frac{1}{N}\sum_{I=1}^{N} f_i \left|X_i - \overline{X}\right|$$

$$= \frac{1}{100}(3\,|\,1-3.74\,|+16\,|\,2-3.74\,|+...+2\,|\,8-3.74) = \frac{109.12}{100} = 1.0912.$$ For Table 3.2 distribution on average household expenditure on food, mean $\overline{X} = \text{Rs.}348.66.$

The mean deviation=

$$= \frac{1}{100}\left(2\,|\,274.5-348.66\,|+...+15\,|\,394.5-348.66\,|\right) = \frac{2510.88}{100} = 25.11$$

So far we have considered mean deviation from mean. The mean deviation from median or from mode can also be defined in a similar way.

3.4.4 Variance and Standard Deviation

The most frequently used measures of dispersion are variance and standard deviation. Variance is so commonly used that it is also called dispersion.

Variance is a measure which suitably combines individual deviations from the mean, treating each observation with equal weight as in mean deviation. For variance, however, measure of individual deviation is taken as the *squared difference from the mean*. Since it is more manageable to use the squared difference rather than absolute difference, particularly while doing format mathematics, use of variance has become more popular. Conventionally variance for a population is denoted by σ^2 (pronounced *sigma-squared*) and variance for a sample is denoted by σ^2. Variance is defined as the mean of the squared deviations of observations form their mean. Variance from raw data is computed by

$$\text{Variance } = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

For frequency data, discrete or continuous type, the formula becomes

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n} f_i\left(X_i - \overline{X}\right)^2, \text{ where } N = \sum_{I=1}^{N} f_i$$

In the same scale of measurement, for example, observations with a variance of 2 are less dispersed then observations with variance more than 2. To talk about a distribution in terms of a measure of central tendency and a measure dispersion, it is a practical need to use both measures in the same unit. Mean and mean deviation are in the same unit. Since each deviation has been squared for Based on variance, an equally or more popular measure of dispersion in the same unit as that of observations is *standard deviation* , abbreviated as s.d. Standard deviation as the positive *square root of variance*, i.e., s.d. = σ. As it is the positive square root of variance, it cannot be negative.

Let us compute the s.d. for household size data of Table 3.1

$$\sigma^2 = \frac{1}{100}\left[3(1-3.74)^2 + 16(2-3.74)^2 + ... + 2(8-3.74)^2\right] = \frac{199.24}{100} = 1.9924 \text{ and}$$

$\sigma = 1.4115$.

Similarly for Table 3.2 distribution of average monthly household expenditure on food, variance in Rs. Square is given by

$$\sigma^2 = \frac{1}{100}\left[2(274.50-348.66)^2 + ... + 15(394.5-348.66)^2\right] = \frac{95725.437}{100} = 957.25,$$

and s.d. is

$\sigma = \text{Rs.}3094.$

For computational convenience, the formula for variance is written in alternative form as

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} f_i(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{N} X_i^2 - \overline{X}^2$$

or

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} f_i(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{N} f_i X_i^2 - \overline{X}^2$$

as the case may be. Thus, variance is viewed as

variance = Mean of Squares – Square of the Mean

Using the above formulae, you may compute the variance for the data given in Tables 3.1 and 3.2 and verify the earlier results.

The computation of variance may be greatly simplified by changing X_i to $u_i = \frac{X_i - A}{h}$, as was done in the computation of mean.

Note that, since

$u_i - \overline{u} = \frac{X_i - A}{h} - \frac{\overline{X} - A}{h} = \frac{X_i - \overline{X}}{h}$, we can write

$$X_i - \overline{X} = h(u_i - \overline{u})$$

Hence,

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left\{h\left(u_i - \overline{u}\right)\right\}^2 = h^2\sigma_u^2$$

where σ_x^2 is the variance of X_i and σ_u^2 is the variance of the *u* values.

Since the magnitude of *u* values is smaller, it is easier to compute variance of *u* values. Then the variance of X values can be easily computed by using the above formula.

Let us compute the variance by applying the above method for the data given in Table 3.2.

If we write $u_i = \dfrac{X_i - 346.5}{24}$, the u values are

–3, –2, –1, 0, 1, 2 and the respective frequencies are 1, 14, 16, 28, 26, 15.

The mean of u values = $\dfrac{-3\times1-2\times14-1\times16+0\times28+1\times26+2\times15}{100} = 0.09$

The mean of squares of u values =

$$\frac{9\times1+4\times14+1\times16+0\times28+1\times26+4\times15}{100} = 1.67$$

Thus $\sigma_u^2 = 1.67 - (0.09)^2 = 1.6619$ and

$$\sigma_X^2 = (24)^2.(1.6619) = 957.25.$$

Even though change from *X* to *u* is for computational ease, it brings up an important issue. Notice that $\sigma_2^2 = 1.6619$ but $\sigma_X^2 = 957.25,$ where *u* was obtained from X by a simple linear transformation, i.e., by change of origin and scale of X values. Typical such natural case are pounds and kilograms for weight, gallons and litres for liquid volume, etc. Since 1 kg. = 2.2046 lbs., s.d. of 5 kg. when measured in kilograms is same as 11.023 lbs. when measured in pounds; or since 1 litre = 0.22 gallon, s.d. of 5 litres when measured in litres is same as s.d of 1.1 gallons when measured in gallons. Thus, whereas variance and standard deviation are supposed to measure spread of observations, not much can be made out of these measures due to their dependence on the unit of measurement.

In this context, the single most useful result about the spread of observations based on mean and standard deviation, irrespective of unit of measurement, is due to Chebychev.

Check Your Progress 3

1) What is dispersion? What are the common measures of dispersion?

………………………………………………………………………………...……....

……………………………………………………………………………….............

………………………………………………………………………………...…......

……………………………………………………………………………….............

……………………………………………………………………………….............

………………………………………………………………………………...…......

2) In a batch of 10 children the marks obtained by a dull boy are 25 marks below the average marks of other children. Show that the standard deviation of marks for all the children is at least 7.5. If this standard deviation is actually 12.0, find the standard deviation when the dull boy is left out.

…………………………………………………………………………...…....

…………………………………………………………………………............

…………………………………………………………………………............

…………………………………………………………………………........

…………………………………………………………………………............

…………………………………………………………………………......…..

3) The following data shows the daily profits (in Rs.) made by a shopkeeper on 15 successive days.

116, 87, 91, 81, 98, 102, 97, 100, 105, 101, 115, 98, 102, 98, 93

Determine the range, the mean deviation about mean and the standard deviation for the data.

…………………………………………………………………………….........

……………………………………………………………………………............

…………………………………………………………………………….........

…………………………………………………………………………….........

……………………………………………………………………………...........

…………………………………………………………………………….........

4) Compute the arithmetic mean, standard deviation and the mean devaiton of the following data.

Scores	4 – 5	6 – 7	8 – 9	10 – 11	12 – 13	14 – 15	Total
f	4	10	20	15	8	3	60

……………………………………………………………………………......…..

……………………………………………………………………………............

……………………………………………………………………………......…..

……………………………………………………………………………............

…………………………………………………………………………........

5) The mean and the s.d. of a sample of 100 observations were calculated as 40 and 5.1 respectively by a student who by mistake took one observation as 50 instead of 40. Calculated the correct s.d.

……………………………………………………………………………...........

……………………………………………………………………………...........

……………………………………………………………………………...........

……………………………………………………………………………...........

3.6 RELATIONSHIP BETWEEN DISPERSION AND STANDARD DEVIATION

You have earlier learnt that when all the values in a set of data are located near their mean, then exhibit a small amount of dispersion or variation and those set of data in which some values are located far form their mean have a large amount of dispersion. A useful rule that illustrates the relationship between dispersion and standard deviation is given by Chebychev's theorem.

3.6.1 Chebychev's Theorem

For any set of observation and positive constant k (>1), the proportion of observations lying within k standard deviations of the mean is certain to be at least $1-\frac{1}{k^2}$.

Note that the theorem is not useful for any positive k less than or equal to 1, since $1-\frac{1}{k^2}$ is at the most equal to zero. For other values of k, the minimum proportion can be computed easily. For example, proportion of observations within 1.5 s.d. of the mean is certain to be at least $1-\frac{1}{1.5^2}=0.556$ or 55.6%. The following figure indicates spread of data based on Chebychev's theorem. For the household size data of Table 3.1, $\overline{X}=3.74$ and s = 1.4115. If we take k=2, we can say that at least $\left[\left(1-\frac{1}{2^2}\right)\times 100\right]=75\%$ of the households are certain to have their size between $3.74\pm 2\times 1.4115$, i.e., between 0.917 and 6.563.

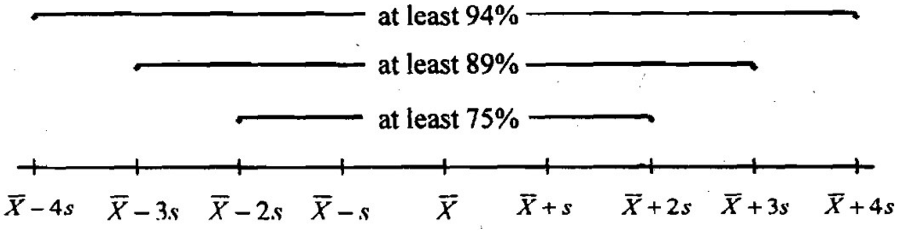


Fig. 3.2

For the Table 3.2 distribution of average monthly household expenditure on food $\overline{X} = \text{Rs.}348.66$ and s = Rs. 30.94, at least 55.6% (for *k* = 1.5) of households are certain to have monthly average food expenditure between Rs. 302.25 and Rs. 395.07. You can find the relevance of this theorem when we study normal distribution later in Unit 11.

3.6.2 Shape of Distribution

For methodological studies in many situations, a distribution is adequately described by measures of central tendency and dispersion. Yet other measures are also in use to describe distribution in practical situations, particularly for economic variables such as income, consumption, economic assets, etc., which are non-negative. Two such measures are *coefficient of variation and concentration ratio*. These measures will be viewed here essentially as measures of inequality in the distribution of economic variables.

3.6.3 Coefficient of Variation

Let us propose to economic status of households in two villages. The summary figures of monthly calories intake of households are given below for the two villages.

	Villages	
	A	**B**
Number of Households (n)	817	561
Mean calorie intake $(\overline{X})$	2417	2235
s.d. of calorie intake (σ)	418	232

The problem is to identify the village that has more inequality as far as calorie intake is concerned. Village A has higher mean calorie intake but has larger s.d. and larger number of households compared to village B. Village A may actually have more number of poorer households in than in village B. Therefore, in village A, inequality between households may be more than that in village B. One index which measures the quantum of such disparity is called the coefficient of variation, abbreviated as c.v. It is defined as percentage standard deviation per unit of mean, i.e.,

$$\text{c.v} = \frac{\sigma}{\overline{X}} \times 100$$

Since σ and $\overline{X}$ have the same unit of measurement, c.v. is unit free and is not affected by the choice of unit of measurement.

For village A, c.v. $= \frac{418}{2417} \times 100 = 17.29$ and for village B,

$$\text{c.v.} = \frac{232}{2235} \times 100 = 10.38.$$

Since the coefficient of variation in village A is greater than the coefficient of variation in village B, the inequalities are given in village A compared to village B.

To compare the extent of inequalities, we compute

$\frac{17.29 - 10.38}{10.38} \times 100 = 66.57$ which implies that compared to village B, 66.57% more inequality exists in village A.

3.6.4 Concentration Ratio

Above was a comparison of inequality between two villages, without quantifying the level of inequality within each village. If a distribution has a long right tail, it shows that a few have a large share. In other words, a majority of population has a very small share. Let us consider the distribution of income of a hypothetical economy.

Suppose there are three classes of people in the economy – the upper class, the middle class and the lower class. Let 10%, 30% and 60% be the share of population in these three classes respectively. Suppose the lower class receives only 20% of the national income, the middle class 30% and the upper class the rest, i.e., the remaining 50%. We can now present the data in a percentage cumulative frequency distribution form. Thus, the lowest 60% of the population receives only 20% of the income, the lowest 90% receive 50% (= 20 + 30) of the income and obviously, 100% of the population receive 100% of the income. If we take a graph paper where the percent cumulative total income is plotted on the vertical axis and we plot the point (0, 0), (60, 20), (90, 50) and (100, 100), then the curve joining these points is what we call the ***curve of concentration*** or ***Lorenz curve***. The straight line joining the points (0,0) and (100, 100) give the line of *equal distribution* or the *equitable line*. The equitable line is that one which shows that the proportion of share is exactly the same as the proportion of population who are supposed to share. The area between the line of equal distribution and the curve of concentration, called the *area of concentration* is an indicator of the degree of concentration; the larger the area the greater is the concentration.

Coefficient of Inequality

Let us take coordinates of the above points in per unit terms instead of percentage terms. Thus, the coordinates of the points, in the above example, can be written as (0,0), (0.60), (0.20), (0.90, 0.50) and (1.00, 1.00). The coefficient of inequality of income distribution is then defined as the ratio of the area of concentration to total area of the triangle. Since the area of the triangle is 0.5 (since $\frac{1}{2} \times 1 \times 1 = 0.5$), the coefficient of inequality is equal to twice the area of concentration when coordinates of various points are taken per unit rather than in percentage.

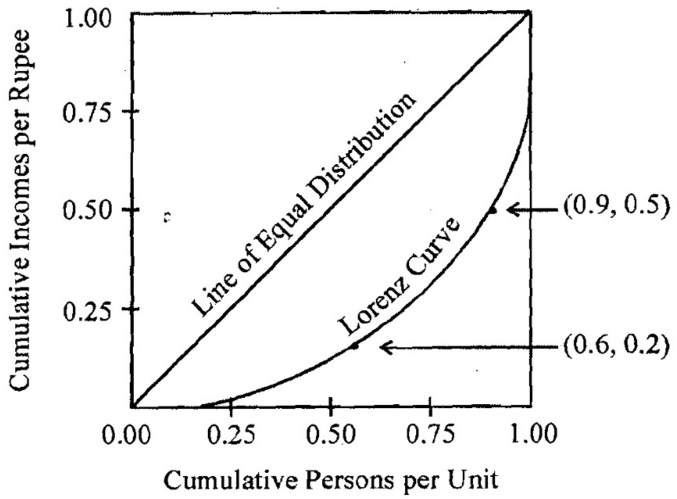


Fig. 3.3: Lorenz Curve

Check Your Progress 4

1) The following figures give the crude birth rate per 1000 people in Switzerland from 1968 to 1980.

   Crude birth rate (*X*): 17.1, 16.5, 15.8, 15.2, 14.3, 13.6, 12.9, 12.3, 11.7, 11.5, 11.3, 11.3, 11.6.

   Calculate the Variance, Standard Deviation and Coefficient of Variation.

   ……………………………………………………………………………………..……….

   ……………………………………………………………………………………..……...

   ……………………………………………………………………………………..……….

   ……………………………………………………………………………………...........

   ……………………………………………………………………………………...…….

2) The following table gives the distribution of age of lady teachers of a school as revealed by records.

Age Group (years)	No. of lady teachers
15 – 19	3
20 – 24	13
25 – 29	21
30 – 34	15
35 – 39	5
40 – 44	4
45 – 49	2

Calculate coefficient of variation, and (ii) number of teachers between the age 26 and 33 years.

……………………………………………………………………………......…..

……………………………………………………………………………............

……………………………………………………………………………........

……………………………………………………………………………............

……………………………………………………………………………........

3.7 LET US SUM UP

In this unit you have learned to compute various measures of central tendency. These measures of central tendency can be divided into two broad categories, namely mathematical averages and positional averages. Positional averages are mode, median, quartile, percentiles, etc., while arithmetic mean, geometric mean and harmonic mean are mathematical averages. Geometric Mean is most suitable for averaging ratio and proportional rates of growth while Arithmetic mean or Harmonic mean can be used to find average rates like price, speed, etc. depending upon the nature of the given condition.

You also learned about the measures of dispersion. The most important measures of dispersion you learned about in the unit are the variance, standard deviation and the concentration ratio. You have also learned to compute variance, standard deviation and coefficient of variation using both ungrouped and grouped data. The coefficient of variation is used to compare the dispersion of two distributions having either different means (even when their variables are measured in same units) or different units of measurement of their variables.

3.8 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) 5.1, 5, 5

2) 108.48 ; 108.41 ; 111.42

Check Your Progress 2

1) Rs. 1.96 ; Rs. 1.95

2) Rs. 674.31 crore

3) i) 25.83 years (ii) 24.82 years (iii) 24.86 years (iv) 27.30 years
(v) 25.59 years (vi) 28.79 years.

4) Arithmetic Mean, 38.25 days

5) Harmonic Mean, 12%.

Check Your Progress 3

1) Do it yourself.

2) 9.9

3) 35, 6.46, 8.85

4) 9.23, 2.49, 2.03

5) 5.0

Check Your Progress 4

1) 3.085, 2.021, 15.004%

2) 23.47%, 25 (rounded figure)

UNIT 4 MEASURES OF SKEWNESS AND KURTOSIS[*]

Structure

4.0 OBJECTIVES

After going through this Unit, you will be able to:

- distinguish between a symmetrical and a skewed distribution;
- compute various coefficients to measure the extent of skewness in a distribution;
- distinguish between platykurtic, mesokurtic and leptokurtic distributions and
- compute the coefficient of kurtosis.

4.1 INTRODUCTION

In this unit you will learn various techniques to distinguish between various shapes of a frequency distribution. This is the final unit with regard to the summarization of univariate data. This unit will make you familiar with the concept of skewness and kurtosis. The need to study these concepts arises from the fact that the measures of central tendency and dispersion fail to describe a distribution completely. It is possible to have frequency distributions which differ widely in their nature and composition and yet may have same central tendency and dispersion. Thus, there is need to supplement the measures of central tendency and dispersion. Consequently, in this unit, we shall discuss two such measures, viz., measures of skewness and kurtosis.

[*] Adapted from IGNOU study material of EEC 13: Elementary Statistical Methods and Survey Techniques, Units 4, 5 and 6 written by R S Bharadwaj with modifications by Kaustuva Barik

4.2 CONCEPT OF SKEWNESS

The skewness of a distribution is defined as the lack of *symmetry*. In a symmetrical distribution, the Mean, Median and Mode are equal to each other and the ordinate at mean divides the distribution into two equal parts such that one part is the mirror image of the other (Fig. 6.1). If some observations, of very high (low) magnitude, are added to such a distribution, its right (left) tail gets elongated.

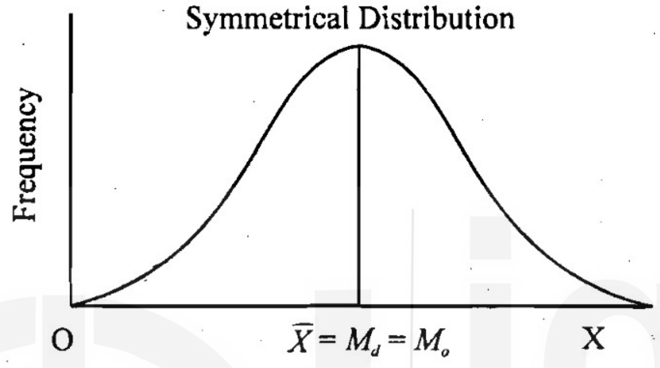


Fig. 4.1

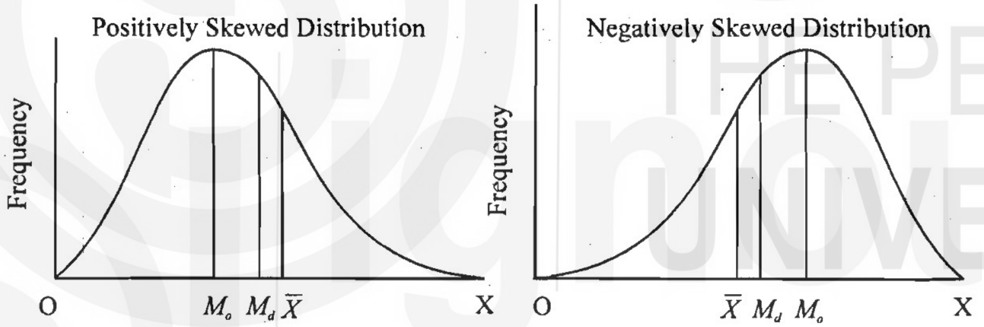


Fig. 4.2

These observations are also known as extreme observations. The presence of extreme observations on the right hand side of a distribution makes it positively skewed and the three averages, viz., mean, median and mode, will no longer be equal.

For positively skewed distribution we find that:

Mean > Median > Mode

On the other hand, the presence of extreme observations to the left hand side of a distribution make it negatively skewed and the relationship between mean, median and mode is:

Mean < Median < Mode

In Fig. 4.2 we depict the shapes of positively skewed and negatively skewed distributions. The direction and extent of skewness can be measured in various ways. We will discuss four measures of skewness in this unit.

4.2.1 Karl Pearson's Measure of Skewness

In Fig. 4.2 you noticed that the mean, median and mode are not equal in a skewed distribution. The Karl Pearson's measure of skewness is based upon the *divergence of mean from mode* in a skewed distribution.

Since Mean = Mode in a symmetrical distribution, (Mean – Mode) can be taken as an *absolute measure of skewnes.* The absolute measure of skewness for a distribution depends upon the unit of measurement. For example, if the mean = 2.45 metre and mode = 2.14 metre, then absolute measure of skewness will be 2.45 metre – 2.14 metre = 0.31 metre. For the same distribution, if we change the unit of measurement to centimeters, the absolute measure of skewness is 245 centimetre – 214 centimetre = 31 centimatre. In order to avoid such a problem Karl Pearson takes a relative measure of skewness.

A relative measure, independent of the units of measurement, is defined as the *Karl Pearson's Coefficient of skewness S_k*, given by

$$S_k = \frac{\text{Mean} - \text{Mode}}{\text{s.d.}}$$

The sign of S_k gives the direction and its magnitude gives the extent of skewness. If $S_k > 0$, the distribution is positively skewed, and if $S_k < 0$ it is negatively skewed.

Example 4.1: Compute the Karl Pearson's coefficient of skewness from the following data:

Height (in inches)	Number of Persons
58	10
59	18
60	30
61	42
62	35
63	28
64	16
65	8

So far we have seen that S_k is strategically dependent upon mode. If mode is not defined for a distribution we cannot find S_k. The empirical relation between mean, median and mode states that, for a moderately symmetrical distribution, we have median and mode states that, for a moderately symmetrical distribution, we have

$$\text{Mean} - \text{Mode} \approx 3\,(\text{Mean} - \text{Median})$$

Hence Karl Pearson's coefficient of skewness is defined in terms of median as

$$S_k = \frac{3(\text{Mean} - \text{Mode})}{\text{s.d.}}$$

Table for the computation of mean and standard deviation (s.d.)

Height (*X*)	*U = X – 61*	No. of persons (*f*)	*fu*	*fu*^{2}
58	-3	10	-30	90
59	-2	18	-36	72
60	-1	30	-30	30
61	0	42	0	0
62	1	35	35	35
63	2	28	56	112
64	3	16	48	144
65	4	8	32	128
Total		**187**	**75**	**611**

$$\text{Mean} = 61 + \frac{75}{187} = 61.4$$

$$\text{s.d.} = \sqrt{\frac{611}{187} - \left(\frac{75}{187}\right)^2} = 1.76$$

To find mode, we note that height is a continuous variable. It is assumed that the height has been measured under the approximation that a measurement on height that is, e.g., greater than 58 but less than 58.5 is taken as 58 inches while a measuring greater than or equal to 58.5 but less than 59 is taken as 59 inches. Thus the given data can be written as

Height (in inches)	Number of Persons
57.5 - 58.5	10
58.5 - 59.5	18
59.5 - 60.5	30
60.5 - 61.5	42
61.5 - 62.5	35
62.5 - 63.5	28
63.5 - 64.5	16
64.5 - 65.5	8

By inspection, the modal class in 60.5 – 61.5. Thus, we have

$$l_m = 60.5, \Delta_1 = 42 - 30 = 12, \Delta_2 = 42 - 35 = 7 \text{ and } h = 1$$

$$\therefore \ \text{Mode} = 60.5 + \frac{12}{12+7} \times 1 = 61.13$$

Hence, the Karl Pearson's coefficient of skewness $S_k = \frac{61.4 - 61.13}{1.76} = 0.153$.

Thus the distribution is positively skewed.

4.2.2 Bowley's Measure of Skewness

This measure is based on quartiles. For a symmetrical distribution, it is seen that Q_1 and Q_2 are equidistant from median. Thus $(Q_3 - M_d) - (M_d - Q_1)$ can be taken as an absolute measure of skewness.

A relative measure of skewness, known as Bowley's coefficient (S_Q), is given by

$$S_Q = \frac{(Q_3 - M_d) - (M_d - Q_1)}{(Q_3 - M_d) - (M_d - Q_1)}$$

$$= \frac{Q_3 - 2M_d + Q_1}{Q_3 - Q_1}$$

Let us calculate the Bowley's coefficient for the following data on height of 187 persons:

Height (in inches)	Number of Persons(*f*)	Cumulative Frequency
57.5 - 58.5	10	10
58.5 - 59.5	18	28
59.5 - 60.5	30	58
60.5 - 61.5	42	100
61.5 - 62.5	35	135
62.5 - 63.5	28	163
63.5 - 64.5	16	179
64.5 - 65.5	8	187

Computation of Q_1:

Since $\frac{N}{4} = 46.75$, the first quartile class is 59.5 – 60.5. Thus
$l_{Q_1} = 59.5, C = 28,\ f_{Q_1} = 30$ and $h = 1$.
$\therefore Q_1 = 59.59.5 + \frac{46.75-28}{30} \times 1 = 60.125.$

Computation of $M_d(Q_2)$:

Since $\frac{N}{2} = 93.5$, the median class is 60.5 – 61.5. Thus
$l_m = 60.5,\ C = 58,\ f_m = 42$ and $h = 1$.
$\therefore M_d = 60.5 + \frac{93-58}{42} \times 1 = 61.345.$

Computation of Q_3:

Since $\frac{3N}{4} = 140.25$, the third quartile class is 62.5 – 63.5. Thus
$l_{Q_3} = 62.5,\ C = 135,\ f_{Q_3} = 28$ and $h = 1$.
$\therefore Q_3 = 62.5 + \frac{140.25-135}{28} \times 1 = 62.688.$

Hence, Bowley's coefficient $S_Q = \frac{62.88-2\times61.345+60.125}{62.688-60.125} = 0.048.$

4.2.3 Kelly's Measure of Skewness

Bowley's measure of skewness is based on the middle 50% of the observation because it leaves 25% of the observations on each extreme of the distribution. As an improvement over Bowley's measure, Kelly has suggested a measure based on P_{10} and P_{90} so that only 10% of the observations on each extreme are ignored. Kelly's coefficient of skewness, denoted by S_p, is given by

$$S_p = \frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{(P_{90} - P_{50}) + (P_{50} - P_{10})}$$

$$= \frac{P_{90} - 2.P_{50} + P_{10}}{P_{90} - P_{10}}$$

Note that $P_{50} = M_d$ (median).

The value of S_p, for the data given in Table 6.1, can be computed as given below.

Computation of P_{10}:

Since $\frac{10N}{100} = \frac{10\times187}{100} = 18.7$, 10th percentile lies in the class 58.5 – 59.5. Thus $l_{P_{10}} = 58.5,\ C = 10,\ f_{P_{10}} = 18$ and $h = 1$.

$\therefore P_{10} = 58.5 + \frac{1807-10}{18} \times 1 = 58.983$

Computation of P_{90}:

Since $\frac{90N}{100} = \frac{90\times187}{100} = 168.3$, 90th percentile lies in the class 63.5 – 64.5. Thus

$l_{P_{90}} = 63.5,\ C = 163,\ f_{P_{90}} = 16$ and $h = 1$.

$\therefore P_{90} = 63.5 + \frac{168.5-163}{16} \times 1 = 63.831$

Hence, Kelly's coefficient $S_P = \frac{63.831-2\times61.345 \quad .983}{63.68 \quad .983} = 0.026.$

It may be noted here that although the coefficient S_k, S_Q and S_P are not comparable, however, in the absence of skewness, each of them will be equal to zero.

Check Your Progress 1

1) Compute the Karl Pearson's coefficient of skewness from the following data:

Daily Expenditure (Rs.):	0-20	20-40	40-60	60-80	80-100
No. of families:	13	25	27	19	16

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

2) The following figures relate to the size of capital of 285 companies:

Capital (in Rs. lacs.)	1-5	6-10	11-15	16-20	21-25	26-30	31-35	Total
No. of companies	20	27	29	38	48	53	70	285

Compute the Bowley's and Kelly's coefficients of skewness and interpret the results.

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

3) The following measures were computed for a frequency distribution: Mean = 50, coefficient of Variation = 35% and Karl Pearson's Coefficient of Skewness = − 0.25. Compute Standard Deviation, Mode and Median of the distribution.

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

4.3 MOMENTS

The r^{th} moment about mean of a distribution, denoted by μ_r, is given by

$$\mu_r = \frac{1}{N}\sum_{i-1}^{n} f_i(X_i - \bar{X})^r, \qquad \text{where } r = 0, 1, 2, 3, 4, \dots \dots .$$

Thus, r^{th} moment about mean is the mean of the r^{th} power of deviations of observations from their arithmetic mean. In particular,

$$\text{if } r = 0, \text{we have } \mu_0 = \frac{1}{N}\sum_{i-1}^{n} f_i(X_i - \bar{X})^0 = 1,$$

$$\text{if } r = 1, \text{we have } \mu_1 = \frac{1}{N}\sum_{i-1}^{n} f_i(X_i - \bar{X})^1 = 0,$$

$$\text{if } r = 2, \text{we have } \mu_2 = \frac{1}{N}\sum_{i-1}^{n} f_i(X_i - \bar{X})^2 = \sigma^2,$$

$$\text{if } r = 3, \text{we have } \mu_3 = \frac{1}{N}\sum_{i-1}^{n} f_i(X_i - \bar{X})^3 \text{ and so on.}$$

These moments are also known as *central moments.*

In addition to the above, we can define *raw moments* as moments about any arbitrary mean.

Let *A* denote an arbitrary mean, then r^{th} moment about *A* is defined as

$$\mu_r' = \frac{1}{N}\sum_{i-1}^{n} f_i(X_i - A)^r, \;\; r = 0, 1, 2, 3, \ldots..$$

When $A = 0$, we get various moments about origin.

Moment Measure of Skewness

The moment measure of skewness is based on the property that, for a symmetrical distribution, all odd ordered central moments are equal to zero. We note that $\mu_0 = 0$, for every distribution, therefore, the lowest order moment that can provide an absolute measure of skewness is $\mu_{3.}$

Further, a coefficient of skewness, independent of the units of measurement, is given by

$\alpha_3 = \frac{\mu_3}{\sigma^3} = \sqrt[\pm]{\beta_1 = \gamma_1}$, where β_1 and γ_1 are defined as the *first beta* and *first gamma* coefficients respectively. Note that β_2 is a measure of kurtosis as you will come to know in the next Section.

Very often, the skewness is measured in terms of $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$, where the sign of skewness is determined by the sign of μ_3.

Example 4.2: Compute the Moment coefficient of skewness (β_1) from the following data.

Marks Obtained:	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency:	6	12	22	24	16	12	8

Table for the computations of mean, s.d. and μ_3.

Class Intervals	Frequency (*f*)	Mid-values(*X*)	$u = \frac{X - 35}{10}$	fu	fu^2	fu^3
0-10	6	5	-3	-18	54	-162
10-20	12	15	-2	-24	48	-96
20-30	22	25	-1	-22	22	-22
30-40	24	35	0	0	0	0
40-50	16	45	1	16	16	16
50-60	12	55	2	24	48	96
60-70	8	65	3	24	72	216
Total	**100**			**0**	**260**	**48**

Since $\sum fu = 0$, the mean of the distribution is 35.

The second moment μ_2 is equal to the variance (σ^2) and its positive square root is equal to standard deviation (σ).

$\mu_2 = \frac{260}{100} \times 100 = 260$, and
$\text{s.d.}(\sigma) = \sqrt{260} = 16.12.$

Also, $\mu_3 = \frac{48}{100} \times 1000 = 480.$

Thus, $\beta_1 = \frac{(480)^2}{(260)^3} = 0.01.$

Since the sign of μ_3 is positive and β_1 is small, the distribution is slightly positively skewed.

If the mean of a distribution is not a convenient figure like 35, as in the above example, the computation of various central moments may become a cumbersome task. Alternatively, we can first compute raw moments and then convert them into central moments by using the equations obtained below.

Conversion of Raw Moments into Central Moments

We can write

$$\mu_r = \frac{1}{N}\sum_{i-1}^{n} f_i(X_i - \bar{X})^r = \frac{1}{N}\sum_{i-1}^{n} f_i[(X_i - A) - (\bar{X} - A)]^r$$

$$= \frac{1}{N}\sum_{i-1}^{n} f_i[(X_i - A) - \mu'_1]^r \quad (\text{Since}\mu'_1 == \frac{1}{N}\sum_{i-1}^{n} f_i(X_i - A) = \bar{X} - A)$$

Expanding the term within brackets by *binomial theorem,* we get

$$= \frac{1}{N}\sum_{i-1}^{n} f_i\left[r_{C_0}(X_i - A)^r \mu_1'^{\,0} - r_{C_1}(X_i - A)^{r-1}\mu'_1 + r_{C_2}(X_i - A)^{r-2}\mu_1'^2 - \cdots\right]$$

$$= \frac{1}{N}\sum_{i-1}^{n} f_i(X_i - A)^r - r_{C_1}\frac{1}{N}\sum_{i-1}^{n} f_i(X_i - A)^{r-1}\,\mu'_1 + r_{C_2}$$

$$+ \frac{1}{N}\sum_{i-1}^{n} f_i(X_i - A)^r\ \mu_1'^2 - \cdots$$

From the above, we can write

$$\mu_r = \mu'_2 - r_{C_1}\mu'_{r-1}\mu'_2 + r_{C_2}\mu'_{r-2}\mu_1'^2 - r_{C_3}r_{C_2}\mu'_{r-3}\mu_1'^3 + \cdots$$

In particular, taking r = 2, 3, 4, etc., we get

$$\mu_2 = \mu'_2 - 2_{C_1}\mu_r'^2 + 2_{C_2}\mu'_0\mu_1'^2 = \mu'_2 - \mu_1'^2(\mu'_0 = 1)$$

$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_0 + 3\mu'_3 - \mu_2'^3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu_1'^3$$

$$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu_1'^2 - 4\mu_1'^4 + \mu_1'^4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu_1'^2 - 3\mu_1'^4$$

Example 4.3: Compute the first four moments about mean from the following data.

Class Intervals:	0 – 10	10 – 20	20 – 30	30 – 40
Frequency (*f*) :	1	3	4	2

Table for computations of raw moments (Take *A* = 25).

Class Intervals	f	Mid-Value	$u = \frac{X-25}{10}$	fu	fu^2	fu^3	fu^4
0 – 10	1	5	– 2	– 2	4	-8	16
10 – 20	2	15	– 1	– 3	3	-3	3
20 – 30	3	25	0	0	0	0	0
30 – 40	4	35	1	2	2	2	2
Total	**10**			**– 3**	**9**	**– 9**	**21**

From the above table, we can write

$$\mu_1' = \frac{-3 \times 10}{10} = -3,$$

$$\mu_2' = \frac{9 \times 10^2}{10} = 90,$$

$$\mu_3' = \frac{-9 \times 10^3}{10} = -900 \text{ and}$$

$$\mu_4' = \frac{21 \times 10^4}{10} = 2100$$

Moments about Mean

By definition,

$\mu_1 = 0,$

$\mu_2 = 90 - 9 = 81,$

$\mu_3 = -900 - 3 \times 90 \times (-3) + 2 \times (-3)^3 = -900 + 810 - 54 = -144$ and

$\mu_4 = 21000 - 4 \times (-900) \times (-3) + 6 \times 90 \times (-3)^3 - 3 \times (-3)^4$

$= 21000 - 10800 + 4860 - 243 = 14817.$

Check Your Progress 2

1) Calculate the first four moments about mean for the following distribution. Also, calculate β_1 and comment upon the nature of skewness.

Class Intervals:	0 – 20	20 – 40	40 – 60	60 – 80	80 – 100
Frequency (*f*) :	8	28	35	17	12

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

2) The first three moment of distribution about the value 3 of a variable are 2, 10 and 30 respectively. Obtain $\overline{X}, \mu_2, \mu_3$ and hence $\beta_{1.}$ Comment upon the nature of skewness.

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

……………………………………………………………………………

4.4 CONCEPT AND MEASURE OF KURTOSIS

Kurtosis is another measure of the shape of a distribution. Whereas skewness measures the lack of symmetry of the frequency curve of a distribution, kurtosis is a measure of the relative peakedness of its frequency curve. Various frequency curves can be divided into three categories depending upon the shape of their peak. The three shapes are termed as Leptokurtic, Mesokurtic and Platykurtic as shown in Fig. 4.3.

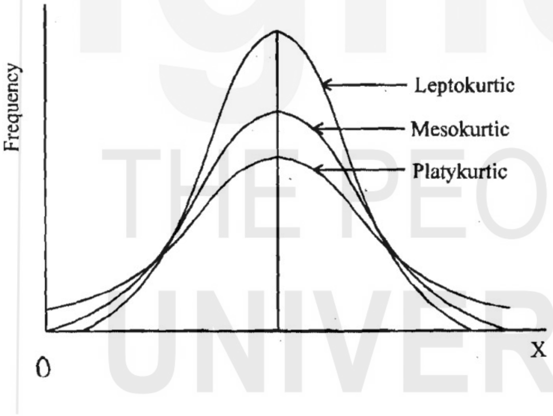


Fig. 4.3

A measure of Kurtosis is given by $\beta_2 = \frac{\mu_4}{\mu_2^2}$, a coefficient given by Karl Pearson. The value of $\beta_2 = 3$ for a mesokurtic curve. When $\beta_2 > 3$, the curve is more peaked than the mesokurtic curve and is termed as lepotokurtic. Similarly, when $\beta_2 < 3$, the curve is less peaked than the mesokurtic curve and is called as platykuritc curve.

Example 4.4: The first four central moments of a distribution are 0, 2,5,0,7 and 18.75. Examine the skewness and kurtosis of the distribution.

To examine skewness, we compute β_1.

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0.7)^2}{(2.5)^3} = 0.031$$

Since $\mu_3 > 0$ and β_1 is small, the distribution is moderately positively skewed.

Kurtosis is given by the coefficient

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{18.75}{(2.5)^2} = 3.0.$$

Hence the curve is mesokutic.

Check Your Progress 3

1) Compute the first four central moments from the following data. Also find the two beta coefficients.

Value	:	5	10	15	20	25	30	35
Frequency	:	8	15	20	32	23	17	5

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

2) The first four moments of a distribution are 1, 4, 10 and 46 respectively. Compute the moment coefficients of skewness and kurtosis and comment upon the nature of the distribution.

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

………………………………………………………………………………

4.5 LET US SUM UP

In this Unit you have learned about the measures of skeweness and kurtosis. These two concepts are used to get an idea about the shape of the frequency curve of a distribution. Skewness is a measure of the lack of symmetry whereas kurtosis is a measure of the relative peakedness of the top of a frequency curve.

4.6 ANSWERS OR HINTS TO CHECK YOUR PROGRSS EXERCISES

Check Your Progress 1

1) 0.237

2) – 0.12, – 0.243

3) 17.5, 54.38, 51.46

Check Your Progress 2

1) 0,499.64, 2579.57, 589111.61, 0.053, skewness is positive.

2) 5,6, – 14, 0,907 since μ_3 is negative the distribution is negatively skewed.

Check Your Progress 3

1) 0, 59.99, – 50.18, 8356.64, 0.012, (negatively skewed), 2.32 (platykurtic).

2) 0, 3. Thus the distribution is symmetrical and mesokurtic. Such a distribution is also known as a Normal Distribution.