STATISTICS FOR ECONOMICS

Textbook for Class XI

राष्ट्रीय शैक्षिक अनुसंधान और प्रशिक्षण परिषद्
NATIONAL COUNCIL OF EDUCATIONAL RESEARCH AND TRAINING

ISBN 81-7450-497-4

First Edition
February 2006 Phalguna 1927
Reprinted
December 2006 Pausa 1928
December 2007 Pausa 1929
January 2009 Magha 1930
January 2010 Magha 1931
January 2011 Magha 1932
January 2012 Magha 1933
December 2012 Agrahayana 1934
November 2013 Kartika 1935

PD 115T MJ

₹ 45.00

Printed on 80 GSM paper with NCERT watermark

Published at the Publication Division by the Secretary, National Council of Educational Research and Training, Sri Aurobindo Marg, New Delhi 110016 and printed at Gopsons Papers Limited, A-2 & 3, Sector-64, Noida - 201 301 (UP)

OFFICES OF THE PUBLICATION DIVISION, NCERT

NCERT Campus
Sri Aurobindo Marg
New Delhi 110 016 **Phone : 011-26562708**

108, 100 Feet Road
Hosdakere Halli Extension
Banashankari III Stage
Bengaluru 560 085 **Phone : 080-26725740**

Navjivan Trust Building
P.O.Navjivan
Ahmedabad 380 014 **Phone : 079-27541446**

CWC Campus
Opp. Dhankal Bus Stop
Panihati
Kolkata 700 114 **Phone : 033-25530454**

CWC Complex
Maligaon
Guwahati 781 021 **Phone : 0361-2674869**

Publication Team

Head, Publication Division : *Ashok Srivastava*
Chief Production Officer : *Kalyan Banerjee*
Chief Business Manager : *Gautam Ganguly*
Chief Editor (Contractual Service) : *Naresh Yadav*
Production Assistant : *Prakash Veer Singh*

Cover
Shweta Rao

Illustrations and Layout
Sarita Verma Mathur

Foreword

The National Curriculum Framework (NCF), 2005, recommends that children's life at school must be linked to their life outside the school. This principle marks a departure from the legacy of bookish learning which continues to shape our system and causes a gap between the school, home and community. The syllabi and textbooks developed on the basis of NCF signify an attempt to implement this basic idea. They also attempt to discourage rote learning and the maintenance of sharp boundaries between different subject areas. We hope these measures will take us significantly further in the direction of a child-centred system of education outlined in the National Policy on Education (1986).

The success of this effort depends on the steps that school principals and teachers will take to encourage children to reflect on their own learning and to pursue imaginative activities and questions. We must recognise that, given space, time and freedom, children generate new knowledge by engaging with the information passed on to them by adults. Treating the prescribed textbook as the sole basis of examination is one of the key reasons why other resources and sites of learning are ignored. Inculcating creativity and initiative is possible if we perceive and treat children as participants in learning, not as receivers of a fixed body of knowledge.

These aims imply considerable change in school routines and mode of functioning. Flexibility in the daily time-table is as necessary as rigour in implementing the annual calendar so that the required number of teaching days are actually devoted to teaching. The methods used for teaching and evaluation will also determine how effective this textbook

proves for making children's life at school a happy experience, rather than a source of stress or boredom. Syllabus designers have tried to address the problem of curricular burden by restructuring and reorienting knowledge at different stages with greater consideration for child psychology and the time available for teaching. The textbook attempts to enhance this endeavour by giving higher priority and space to opportunities for contemplation and wondering, discussion in small groups, and activities requiring hands-on experience.

The National Council of Educational Research and Training (NCERT) appreciates the hard work done by the textbook development team responsible for this book. We wish to thank the Chairperson of the advisory group for Social Sciences textbooks at Higher Secondary Level, Professor Hari Vasudevan and the Chief Advisor for this book, Professor Tapas Majumdar for guiding the work of this committee. Several teachers contributed to the development of this textbook; we are grateful to them and their principals for making this possible. We are indebted to the institutions and organisations which have generously permitted us to draw upon their resources, material and personnel. We are especially grateful to the members of the National Monitoring Committee, appointed by the Department of Secondary and Higher Education, Ministry of Human Resource Development under the Chairmanship of Professor Mrinal Miri and Professor G.P. Deshpande, for their valuable time and contribution. As an organisation committed to systemic reform and continuous improvement in the quality of its products, NCERT welcomes comments and suggestions which will enable us to undertake further revision and refinement.

Director
National Council of Educational
Research and Training

New Delhi
20 December 2005

TEXTBOOK DEVELOPMENT COMMITTEE

CHAIRPERSON, ADVISORY COMMITTEE FOR SOCIAL SCIENCE TEXTBOOKS AT HIGHER SECONDARY LEVEL

Hari Vasudevan, *Professor,* Department of History, University of Calcutta, Kolkata

CHIEF ADVISOR

Tapas Majumdar, *Emeritus Professor,* Jawaharlal Nehru University, New Delhi

MEMBERS

Bhawna Rajput, *Sr. Lecturer,* Aditi Mahavidyalaya, Delhi University, Delhi

E. Bijoykumar Singh, *Professor,* Department of Economics, Manipur University, Imphal

M. M. Goel, *Reader,* Department of Commerce, PGDAV College (M), Delhi University, Delhi

Meera Malhotra, *Head,* Economics, Modern School, Barakhamba Road, New Delhi

Sudhir Kumar, *Reader,* A. N. Sinha Institute of Social Studies, Patna

T. P. Sinha, *Reader,* Department of Economics, S.S.N. College, Delhi University, Delhi

MEMBER-COORDINATOR

Neeraja Rashmi, *Reader,* Economics, DESSH, NCERT, New Delhi

ACKNOWLEDGEMENTS

Acknowledgements are due to Savita Sinha, *Professor and Head,* Department of Education in Social Sciences and Humanities for her support in developing this textbook.

The Council is also thankful to J. Khuntia, *Sr. Lecturer,* School of Correspondence Courses, Delhi University; M. V. Srinivasan and Jaya Singh, *Lecturers,* DESSH, NCERT for helping in finalising the textbook.

Special thanks are due to Vandana R. Singh, *Consultant Editor* for going through the manuscript and suggesting relevant changes.

The Council also gratefully acknowledges the contributions of Girish Goyal, *DTP Operator,* Dillip Kumar Agasti, *Proof Reader,* Dinesh Kumar, *Incharge,* Computer Station, in shaping this book. The contribution of the Publication Department, NCERT in bringing out this book is also duly acknowledged.

Contents

THE CONSTITUTION OF INDIA

PREAMBLE

WE, THE PEOPLE OF INDIA, having solemnly resolved to constitute India into a [1]**[SOVEREIGN SOCIALIST SECULAR DEMOCRATIC REPUBLIC]** and to secure to all its citizens :

JUSTICE, social, economic and political;

LIBERTY of thought, expression, belief, faith and worship;

EQUALITY of status and of opportunity and to promote among them all;

FRATERNITY assuring the dignity of the individual and the [2][unity and integrity of the Nation];

IN OUR CONSTITUENT ASSEMBLY this twenty-sixth day of November, 1949 do **HEREBY ADOPT, ENACT AND GIVE TO OURSELVES THIS CONSTITUTION.**

---

1. Subs. by the Constitution (Forty-second Amendment) Act, 1976, Sec.2, for "Sovereign Democratic Republic" (w.e.f. 3.1.1977)
2. Subs. by the Constitution (Forty-second Amendment) Act, 1976, Sec.2, for "Unity of the Nation" (w.e.f. 3.1.1977)

CHAPTER

1

Introduction

Studying this chapter should enable you to:

- *know what the subject of economics is about;*
- *understand how economics is linked with the study of economic activities in consumption, production and distribution;*
- *understand why knowledge of statistics can help in describing consumption, production and distribution;*
- *learn about some uses of statistics in the understanding of economic activities.*

1. WHY ECONOMICS?

You have, perhaps, already had Economics as a subject for your earlier classes at school. You might have been told this subject is mainly around what *Alfred Marshall* (one of the founders of modern economics) called "*the study of man in the ordinary business of life*". Let us understand what that means.

When you buy goods (you may want to satisfy your own personal needs or those of your family or those of any other person to whom you want to make a gift) you are called a **consumer**.

When you sell goods to make a profit for yourself (you may be a shopkeeper), you are called a **seller**.

When you produce goods (you may be a farmer or a manufacturer), you are called a **producer**.

When you are in a job, working for some other person, and you get paid for it (you may be employed by somebody who pays you wages or a salary), you are called a **service-holder**.

When you provide some kind of service to others for a payment (you may be a lawyer or a doctor or a banker or a taxi driver or a transporter of goods), you are called a **service-provider**.

In all these cases you will be called **gainfully employed** in an **economic activity**. Economic activities are ones that are undertaken for a monetary gain. This is what economists mean by **ordinary business of life**.

Activities

- List different activities of the members of your family. Would you call them economic activities? Give reasons.
- Do you consider yourself a consumer? Why?

We cannot get something for nothing

If you ever heard the story of *Aladdin* and *his Magic Lamp*, you would agree that Aladdin was a lucky guy. Whenever and whatever he wanted, he just had to rub his magic lamp on when a genie appeared to fulfill his wish. When he wanted a palace to live in, the genie instantly made one for him. When he wanted expensive gifts to bring to the king when asking for his daughter's hand, he got them at the bat of an eyelid.

In real life we cannot be as lucky as Aladdin. Though, like him we have unlimited wants, we do not have a magic lamp. Take, for example, the pocket money that you get to spend. If you had more of it then you could have purchased almost all the things you wanted. But since your pocket money is limited, you have to choose only those things that you want the most. This is a basic teaching of Economics.

Activities

- Can you think for yourself of some other examples where a person with a given income has to choose which things and in what quantities he or she can buy at the prices that are being charged (called the current prices)?
- What will happen if the current prices go up?

Scarcity is the root of all economic problems. Had there been no scarcity, there would have been no economic problem. And you would not have studied Economics either. In our daily life, we face various forms of scarcity. The long queues at railway booking counters, crowded buses and trains, shortage of essential commodities, the rush to get a ticket to watch a new film, etc., are all manifestations of scarcity. We face scarcity because the things that satisfy our wants are limited in availability. Can you think of some more instances of scarcity?

The resources which the producers have are limited and also have

alternative uses. Take the case of food that you eat every day. It satisfies your want of nourishment. Farmers employed in agriculture raise crops that produce your food. At any point of time, the resources in agriculture like land, labour, water, fertiliser, etc., are given. All these resources have alternative uses. The same resources can be used in the production of non-food crops such as rubber, cotton, jute etc. Thus alternative uses of resources give rise to the problem of choice between different commodities that can be produced by those resources.

Activities

- Identify your wants. How many of them can you fulfill? How many of them are unfulfilled? Why you are unable to fulfill them?
- What are the different kinds of scarcity that you face in your daily life? Identify their causes.

Consumption, Production and Distribution

If you thought about it, you might have realised that Economics involves the study of man engaged in economic activities of various kinds. For this, you need to know reliable facts about all the diverse economic activities like production, consumption and distribution. Economics is often discussed in three parts: *consumption, production* and *distribution*.

We want to know how the consumer decides, given his income and many alternative goods to choose from, what to buy when he knows the prices. *This is the study of Consumption.*

We also want to know how the producer, similarly, chooses what to produce for the market when he knows the costs and prices. *This is the study of Production.*

Finally, we want to know how the national income or the total income arising from what has been produced in the country (called the Gross Domestic Product or GDP) is distributed through wages (and salaries), profits and interest (We will leave aside here income from international trade and investment). *This is study of Distribution.*

Besides these three conventional divisions of the study of Economics about which we want to know all the facts, modern economics has to include some of the basic problems facing the country for special studies.

For example, you might want to know why or to what extent some households in our society have the capacity to earn much more than others. You may want to know how many people in the country are really

poor, how many are middle-class, how many are relatively rich and so on. You may want to know how many are illiterate, who will not get jobs, requiring education, how many are highly educated and will have the best job opportunities and so on. In other words, you may want to know more facts in terms of numbers that would answer questions about poverty and disparity in society. If you do not like the continuance of poverty and gross disparity and want to do something about the ills of society you will need to know the facts about all these things before you can ask for appropriate actions by the government. If you know the facts it may also be possible to plan your own life better. Similarly, you hear of – some of you may even have experienced disasters like Tsunami, earthquakes, the bird flu – dangers threatening our country and so on that affect man's *'ordinary business of life'* enormously. Economists can look at these things provided they know how to collect and put together the facts about what these disasters cost systematically and correctly. You may perhaps think about it and ask yourselves whether it is right that modern economics now includes learning the basic skills involved in making useful studies for measuring poverty, how incomes are distributed, how earning opportunities are related to your education, how environmental disasters affect our lives and so on?

Obviously, if you think along these lines, you will also appreciate why we needed Statistics (which is the study of numbers relating to selected facts in a systematic form) to be added to all modern courses of modern economics.

Would you now agree with the following definition of economics that many economists use?

"Economics is the study of how people and society choose to employ scarce resources that could have alternative uses in order to produce various commodities that satisfy their wants and to distribute them for consumption among various persons and groups in society."

Activity

- Would you say, in the light of the discussion above, that this definition used to be given seems a little inadequate now? What does it miss out?

2. STATISTICS IN ECONOMICS

In the previous section you were told about certain special studies that concern the basic problems facing a country. These studies required that we know more about economic facts in terms of numbers. Such economic facts are also known as **data**.

The purpose of collecting data about these economic problems is to understand and explain these problems in terms of the various causes behind them. In other words, we try to analyse them. For example, when we **analyse** the hardships of poverty, we try to explain it in terms of the various factors such as

unemployment, low productivity of people, backward technology, etc.

But, what purpose does the analysis of poverty serve unless we are able to find ways to mitigate it. We may, therefore, also try to find those measures that help solve an economic problem. In Economics, such measures are known as **policies**.

So, do you realise, then, that no analysis of a problem would be possible without the availability of data on various factors underlying an economic problem? And, that, in such a situation, no policies can be formulated to solve it. If yes, then you have, to a large extent, understood the basic relationship between Economics and Statistics.

3. WHAT IS STATISTICS?

At this stage you are probably ready to know more about Statistics. You might very well want to know what the subject "Statistics" is all about. What are its specific uses in Economics? Does it have any other meaning? Let us see how we can answer these questions to get closer to the subject.

In our daily language the word ***'Statistics'*** is used in two distinct senses: **singular** and **plural**. In the *plural* sense, '*statistics*' means '*numerical facts systematically collected*' as described by Oxford Dictionary. Thus, the simple meaning of statistics in plural sense is data.

> Do you know that the term *statistics* in singular means the 'science of collecting, classifying and using statistics' or a 'statistical fact'.

By data or statistics, we mean both quantitative and qualitative facts that are used in Economics. For example, a statement in Economics like "the production of rice in India has increased from 39.58 million tonnes in 1974–75 to 58.64 million tonnes in 1984–85", is a quantitative fact. The numerical figures such as '39.58 million tonnes' and '58.64 million tonnes' are **statistics** of the production of rice in India for 1974–75 and 1984–85 respectively.

In addition to the quantitative data, Economics also uses qualitative data. The chief characteristic of such information is that they describe attributes of a single person or a group of persons that is important to record as accurately as possible even though they cannot be measured in quantitative terms. Take, for example, "gender" that distinguishes a person as man/woman or boy/girl. It is often possible (and useful) to state the information about an attribute of a person in terms of degrees (like better/ worse; sick/ healthy/ more healthy; unskilled/ skilled/ highly skilled etc.). Such qualitative information or statistics is often used in Economics and other social sciences and collected and stored systematically like quantitative information (on prices, incomes, taxes paid etc.), whether for a single person or a group of persons.

You will study in the subsequent chapters that *statistics* involves *collection* and *organisation* of data. The next step is to present the data in

tabular, diagrammatic and *graphic* forms. The data, then, is summarised by calculating various numerical indices such as *mean, variance, standard deviation* etc. that represent the broad characteristics of the collected set of information.

Activities

- Think of two examples of qualitative and quantitative data.
- Which of the following would give you qualitative data; beauty, intelligence, income earned, marks in a subject, ability to sing, learning skills?

4. WHAT STATISTICS DOES?

By now, you know that Statistics is an indispensable tool for an economist that helps him to understand an economic problem. Using its various methods, effort is made to find the causes behind it with the help of the qualitative and the quantitative facts of the economic problem. Once the causes of the problem are identified, it is easier to formulate certain policies to tackle it.

But there is more to Statistics. It enables an economist to present economic facts in a precise and definite form that helps in proper comprehension of what is stated. When economic facts are expressed in statistical terms, they become exact. Exact facts are more convincing than vague statements. For instance, saying that with precise figures, 310 people died in the recent earthquake in Kashmir, is more factual and, thus, a statistical data. Whereas, saying hundreds of people died, is not.

Statistics also helps in condensing the mass of data into a few numerical measures (such as mean, variance etc., about which you will learn later). These numerical measures help summarise data. For example, it would be impossible for you to remember the incomes of all the people in a data if the number of people is very large. Yet, one can remember easily a summary figure like the average income that is obtained statistically. In this way, Statistics summarises and presents a meaningful overall information about a mass of data.

Quite often, Statistics is used in finding relationships between different economic factors. An economist may be interested in finding out what happens to the demand for a commodity when its price increases or decreases? Or, would the supply of a commodity be affected by the changes in its own price? Or, would the consumption expenditure increase when the average income increases? Or, what happens to the general price level when the government expenditure increases? Such questions can only be answered if any relationship exists between the various economic factors that have been stated above. Whether such relationships exist or not can be easily verified by applying statistical methods to their data. In some cases the economist might assume certain relationships between them and like

to test whether the assumption she/he made about the relationship is valid or not. The economist can do this only by using statistical techniques.

In another instance, the economist might be interested in predicting the changes in one economic factor due to the changes in another factor. For example, she/he might be interested in knowing the impact of today's investment on the national income in future. Such an exercise cannot be undertaken without the knowledge of Statistics.

Sometimes, formulation of plans and policies requires the knowledge of future trends. For example, an economic planner has to decide in 2005 how much the economy should produce in 2010. In other words, one must know what could be the expected level of consumption in 2010 in order to decide the production plan of the economy for 2010. In this situation, one might make subjective judgement based on the guess about consumption in 2010. Alternatively, one might use statistical tools to predict consumption in 2010. That could be based on the data of consumption of past years or of recent years obtained by surveys. Thus, statistical methods help formulate appropriate economic policies that solve economic problems.

5. Conclusion

Today, we increasingly use Statistics to analyse serious economic problems such as rising prices, growing population, unemployment, poverty etc., to find measures that can solve such problems. Further it also helps evaluate the impact of such policies in solving the economic problems. For example, it can be ascertained easily using statistical techniques whether the policy of family planning is effective in checking the problem of ever-growing population.

In economic policies, Statistics plays a vital role in decision making. For example, in the present time of rising global oil prices, it might be necessary to decide how much oil India should import in 2010. The decision to import would depend on the expected domestic production of oil and the likely demand for oil in

Statistical methods are no substitute for common sense!

There is an interesting story which is told to make fun of statistics. It is said that a family of four persons (husband, wife and two children) once set out to cross a river. The father knew the average depth of the river. So he calculated the average height of his family members. Since the average height of his family members was greater than the average depth of the river, he thought they could cross safely. Consequently some members of the family (children) drowned while crossing the river.

Does the fault lie with the statistical method of calculating averages or with the misuse of the averages?

2010. Without the use of Statistics, it cannot be determined what the expected domestic production of oil and the likely demand for oil would be. Thus, the decision to import oil cannot be made unless we know the actual requirement of oil. This vital information that help make the decision to import oil can only be obtained statistically.

Recap

- Our wants are unlimited but the resources used in the production of goods that satisfy our wants are limited and scarce. Scarcity is the root of all economic problems.
- Resources have alternative uses.
- Purchase of goods by consumers to satisfy their various needs is Consumption.
- Manufacture of goods by producers for the market is Production.
- Division of the national income into wages, profits, rents and interests is Distribution.
- Statistics finds economic relationships using data and verifies them.
- Statistical tools are used in prediction of future trends.
- Statistical methods help analyse economic problems and formulate policies to solve them.

EXERCISES

1. Mark the following statements as true or false.
   (i) Statistics can only deal with quantitative data.
   (ii) Statistics solves economic problems.
   (iii) Statistics is of no use to Economics without data.
2. Make a list of activities that constitute the ordinary business of life. Are these economic activities?
3. 'The Government and policy makers use statistical data to formulate suitable policies of economic development'. Illustrate with two examples.
4. You have unlimited wants and limited resources to satisfy them. Explain by giving two examples.
5. How will you choose the wants to be satisfied?
6. What are your reasons for studying Economics?
7. Statistical methods are no substitute for common sense. Comment.

CHAPTER

2

Collection of Data

Studying this chapter should enable you to:

- *understand the meaning and purpose of data collection;*
- *distinguish between primary and secondary sources;*
- *know the mode of collection of data;*
- *distinguish between Census and Sample Surveys;*
- *be familiar with the techniques of sampling;*
- *know about some important sources of secondary data.*

1. INTRODUCTION

In the previous chapter, you have read about what is economics. You also studied about the role and importance of statistics in economics. In this chapter, you will study the sources of data and the mode of data collection. The purpose of collection of data is to collect evidence for reaching a sound and clear solution to a problem.

In economics, you often come across a statement like,

"After many fluctuations the output of food grains rose to 176 million tonnes in 1990–91 and 199 million tonnes in 1996–97, but fell to 194 million tonnes in 1997–98. Production of food grains then rose continuously and touched 212 million tonnes in 2001–02."

In this statement, you can observe that the food grains production in different years does not remain the same. It varies from year to year and from crop to crop. As these values

vary, they are called *variable*. The variables are generally represented by the letters X, Y or Z. The values of these *variables* are the *observation*. For example, suppose the food grain production in India varies between 100 million tonnes in 1970–71 to 220 million tonnes in 2001–02 as shown in the following table. The years are represented by variable X and the production of food grain in India (in million tonnes) is represented by variable Y:

TABLE 2.1

Production of Food Grain in India

(Million Tonnes)

X	*Y*
1970–71	108
1978–79	132
1979–80	108
1990–91	176
1996–97	199
1997–98	194
2001–02	212

Here, these values of the variables X and Y are the 'data', from which we can obtain information about the trend of the production of food grains in India. To know the fluctuations in the output of food grains, we need the 'data' on the production of food grains in India. 'Data' is a tool, which helps in understanding problems by providing information.

You must be wondering where do 'data' come from and how do we collect these? In the following sections we will discuss the types of data, method and instruments of data collection and sources of obtaining data.

2. What are the Sources of Data?

Statistical data can be obtained from two sources. The *enumerator* (person who collects the data) may collect the data by conducting an enquiry or an investigation. Such data are called *Primary Data,* as they are based on first hand information. Suppose, you want to know about the popularity of a film star among school students. For this, you will have to enquire from a large number of school students, by asking questions from them to collect the desired information. The data you get, is an example of primary data.

If the data have been collected and processed (scrutinised and tabulated) by some other agency, they are called *Secondary Data*. Generally, the published data are *secondary* data. They can be obtained either from published sources or from any other source, for example, a web site. Thus, the data are *primary* to the source that collects and processes them for the first time and *secondary* for all sources that later use such data. Use of secondary data saves time and cost. For example, after collecting the data on the popularity of the film star among students, you publish a report. If somebody uses the data collected by you for a similar study, it becomes secondary data.

3. How do we Collect the Data?

Do you know how a manufacturer decides about a product or how a political party decides about a candidate? They conduct a survey by

asking questions about a particular product or candidate from a large group of people. The purpose of surveys is to describe some characteristics like price, quality, usefulness (in case of the product) and popularity, honesty, loyalty (in case of the candidate). The purpose of the *survey* is to collect data. *Survey* is a method of gathering information from individuals.

Preparation of Instrument

The most common type of instrument used in surveys is questionnaire/ interview schedule. The questionnaire is either self administered by the respondent or administered by the researcher (enumerator) or trained investigator. While preparing the questionnaire/interview schedule, you should keep in mind the following points;

- *The questionnaire should not be too long*. The number of questions should be as minimum as possible. Long questionnaires discourage people from completing them.
- *The series of questions should move from general to specific*. The questionnaire should start from general questions and proceed to more specific ones. This helps the respondents feel comfortable. For example:

Poor Q

(i) Is increase in electricity charges justified?
(ii) Is the electricity supply in your locality regular?

Good Q

(i) Is the electricity supply in your locality regular?
(ii) Is increase in electricity charges justified?

- *The questions should be precise and clear.* For example,

Poor Q

What percentage of your income do you spend on clothing in order to look presentable?

Good Q

What percentage of your income do you spend on clothing?

- *The questions should not be ambiguous, to enable the respondents to answer quickly, correctly and clearly*. For example:

Poor Q

Do you spend a lot of money on books in a month?

Good Q

How much do you spend on books in a month?

(i) Less than Rs 200
(ii) Between Rs 200–300
(iii) Between Rs 300–400
(iv) More than Rs 400

- *The question should not use double negatives*. The questions starting with "Wouldn't you" or "Don't you" should be avoided, as they may lead to biased responses. For example:

Poor Q

Don't you think smoking should be prohibited?

Good Q

Do you think smoking should be prohibited?

- *The question should not be a leading question, which gives a clue about how the respondent should answer*. For example:

Poor Q

How do you like the flavour of this high-quality tea?

Good Q

How do you like the flavour of this tea?

- *The question should not indicate alternatives to the answer*. For example:

Poor Q

Would you like to do a job after college or be a housewife?

Good Q

Would you like to do a job, if possible?

The questionnaire may consist of *closed ended* (or structured) questions or *open ended* (or unstructured) questions.

C*losed ended or structured* questions can either be a *two-way question* or a *multiple choice* question. When there are only two possible answers, 'yes' or 'no', it is called a two-way question.

When there is a possibility of more than two options of answers, *multiple choice* questions are more appropriate. *Example,*

Q. Why did you sell your land?

(i) To pay off the debts.
(ii) To finance children's education.
(iii) To invest in another property.
(iv) Any other (please specify).

Closed -ended questions are easy to use, score and code for analysis, because all the respondents respond from the given options. But they are difficult to write as the alternatives should be clearly written to represent both sides of the issue. There is also a possibility that the individual's true response is not present among the options given. For this, the choice of '*Any Other*' is provided, where the respondent can write a response, which was not anticipated by the researcher. Moreover, another limitation of multiple-choice questions is that they tend to restrict the answers by providing alternatives, without which the respondents may have answered differently.

Open-ended questions allow for more individualised responses, but they are difficult to interpret and hard to score, since there are a lot of variations in the responses. *Example,*

Q. What is your view about globalisation?

Mode of Data Collection

Have you ever come across a television show in which reporters ask questions from children, housewives or general public regarding their examination performance or a brand of soap or a political party? The purpose of asking questions is to do a survey for collection of data. There are three basic ways of collecting data: (i) Personal Interviews, (ii) Mailing (questionnaire) Surveys, and (iii) Telephone Interviews.

Personal Interviews

This method is used when the researcher has access to all the members. The researcher (or investigator) conducts face to face interviews with the respondents.

Personal interviews are preferred due to various reasons. Personal contact is made between the respondent and the interviewer. The interviewer has the opportunity of explaining the study and answering any query of the respondents. The interviewer can request the respondent to expand on answers that are particularly important. Misinterpretation and misunderstanding can be avoided. Watching the reactions of the respondents can provide supplementary information.

Personal interview has some demerits too. It is expensive, as it requires trained interviewers. It takes longer time to complete the survey. Presence of the researcher may inhibit respondents from saying what they really think.

Mailing Questionnaire

When the data in a survey are collected by mail, the questionnaire is sent to each individual by *mail* with a request to complete and return it by a given date. The advantages of this method are that, it is less expensive. It allows the researcher to have access to people in remote areas too, who might be difficult to reach in person or by telephone. It does not allow influencing of the respondents by the interviewer. It also permits the respondents to take sufficient time to give thoughtful answers to the questions. These days online surveys or surveys through short messaging service i.e. SMS have become popular. Do you know how an online survey is conducted?

The disadvantages of mail survey are that, there is less opportunity to provide assistance in clarifying instructions, so there is a possibility of misinterpretation of questions. Mailing is also likely to produce low response rates due to certain factors such as returning the questionnaire without completing it, not returning the questionnaire at all, loss of questionnaire in the mail itself, etc.

Telephone Interviews

In a telephone interview, the investigator asks questions over the telephone. The advantages of telephone interviews are that they are cheaper than personal interviews and can be conducted in a shorter time. They allow the researcher to assist the respondent by clarifying the questions. Telephone interview is better in the cases where the respondents are reluctant to answer certain questions in personal interviews.

Activities

- You have to collect information from a person, who lives in a remote village of India. Which mode of data collection will be the most appropriate for collecting information from him?
- You have to interview the parents about the quality of teaching in a school. If the principal of the school is present there, what types of problems can arise?

The disadvantage of this method is access to people, as many people may not own telephones. Telephone Interviews also obstruct visual reactions of the respondents, which becomes helpful in obtaining information on sensitive issues.

Pilot Survey

Once the questionnaire is ready, it is advisable to conduct a try-out with a small group which is known as *Pilot Survey* or *Pre-Testing* of the questionnaire. The pilot survey helps in providing a preliminary idea about the survey. It helps in *pre-testing* of the questionnaire, so as to know the shortcomings and drawbacks of the questions. Pilot survey also helps in assessing the suitability of questions, clarity of instructions, performance of enumerators and the cost and time involved in the actual survey.

4. Census and Sample Surveys

Census or Complete Enumeration

A survey, which includes every element of the population, is known as *Census* or the *Method of Complete Enumeration*. If certain agencies are interested in studying the total population in India, they have to obtain information from all the households in rural and urban India.

Advantages	*Disadvantages*
• Highest Response Rate • Allows use of all types of questions • Better for using open-ended questions • Allows clarification of ambiguous questions.	 • Most expensive • Possibility of influencing respondents • More time taking.
• Least expensive • Only method to reach remote areas • No influence on respondents • Maintains anonymity of respondents • Best for sensitive questions.	 • Cannot be used by illiterates • Long response time • Does not allow explanation of unambiguous questions • Reactions cannot be watched.
• Relatively low cost • Relatively less influence on respondents • Relatively high response rate.	 • Limited use • Reactions cannot be watched • Possibility of influencing respondents.

The essential feature of this method is that this covers every individual unit in the entire population. You cannot select some and leave out others. You may be familiar with the *Census of India*, which is carried out every ten years. A house-to-house enquiry is carried out, covering all households in India. Demographic data on birth and death rates, literacy, workforce, life expectancy, size and composition of population, etc. are collected and published by the Registrar General of India. The last *Census of India* was held in February 2001.

According to the Census 2001, population of India is 102.70 crore. It was 23.83 crore according to Census 1901. In a period of hundred years, the population of our country increased by 78.87 crore. Census

Source: *Census of India, 2001.*

1981 indicated that the rate of population growth during 1960s and 1970s remained almost same. 1991 Census indicated that the annual growth rate of population during 1980s was 2.14 per cent, which came down to 1.93 per cent during 1990s according to Census 2001.

> "At 00.00 hours of first March, 2001 the population of India stood at 1027,015,247 comprising of 531,277,078 males and 495,738,169 females. Thus, India becomes the second country in the world after China to cross the one billion mark."

Source: *Census of India, 2001.*

Sample Survey

Population or the *Universe* in statistics means totality of the items under study. Thus, the *Population* or the *Universe* is a group to which the results of the study are intended to apply. *A population is always all the individuals/items who possess certain characteristics (or a set of characteris-*

tics), according to the purpose of the survey. The first task in selecting a sample is to identify the *population*. Once the population is identified, the researcher selects a *Representative Sample*, as it is difficult to study the entire population. *A sample refers to a group or section of the population from which information is to be obtained. A good sample (representative sample)* is generally smaller than the *population* and is capable of providing reasonably accurate information about the population at a much lower cost and shorter time.

Suppose you want to study the average income of people in a certain region. According to the Census method, you would be required to find out the income of every individual in the region, add them up and divide by number of individuals to get the average income of people in the region. This method would require huge expenditure, as a large number of enumerators have to be employed. Alternatively, you select a *representative sample,* of a few individuals, from the region and find out their income. The average income of the selected group of individuals is used as an estimate of average income of the individuals of the entire region.

Example

- *Research problem*: To study the economic condition of agricultural labourers in Churachandpur district of Manipur.
- *Population*: All agricultural labourers in Churachandpur district.
- *Sample*: Ten per cent of the agricultural labourers in Churachandpur district.

Most of the surveys are sample surveys. These are preferred in statistics because of a number of reasons. A sample can provide reasonably reliable and accurate information at a lower cost and shorter time. As samples are smaller than population, more detailed information can be collected by conducting intensive enquiries. As we need a smaller team of enumerators, it is easier to train them and supervise their work more effectively.

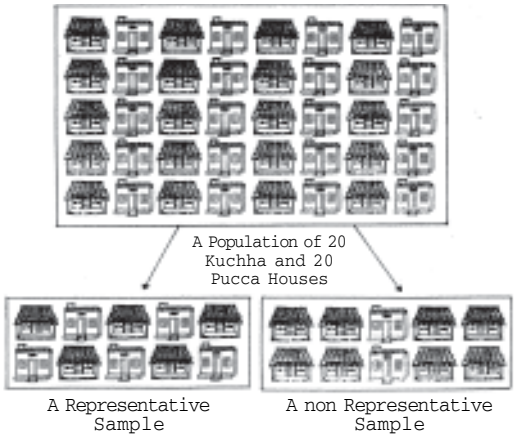
Now the question is how do you do the sampling? There are two main types of sampling, *random* and *non-random*. The following description will make their distinction clear.

Activities

- In which years will the next Census be held in India and China?
- If you have to study the opinion of students about the new economics textbook of class XI, what will be your *population* and *sample?*
- If a researcher wants to estimate the average yield of wheat in Punjab, what will be her/his *population* and *sample*?

Random Sampling

As the name suggests, random sampling is one where the individual units from the population (samples) are selected *at random*. The government wants to determine the



impact of the rise in petrol price on the household budget of a particular locality. For this, a representative (random) sample of 30 households has to be taken and studied. The names of all the 300 households of that area are written on pieces of paper and mixed well, then 30 names to be interviewed are selected one by one.

In the random sampling, every individual has an equal chance of being selected and the individuals who are selected are just like the ones who are not selected. In the above example, all the 300 sampling units (also called *sampling frame)* of the population got an equal chance of being included in the sample of 30 units and hence the sample, such drawn, is a random sample. This is also called *lottery method*. The same could be done using a Random Number Table also.

How to use the Random Number Tables?

Do you know what are the Random Number Tables? Random number tables have been generated to guarantee equal probability of selection of every individual unit (by their listed serial number in the sampling frame) in the population. They are available either in a published form or can be generated by using appropriate software packages (See Appendix B).You can start using the table from anywhere, i.e., from any page, column, row or point. In the above example, you need to select a sample of 30 households out of 300 total households. Here, the largest serial number is 300, a three digit number and therefore we consult three digit random numbers in sequence. We will skip the random numbers greater than 300 since there is no household number greater than 300. Thus, the 30 selected households are with serial numbers: 149, 219, 111, 165, 230, 007, 089, 212, 051, 244, 300, 051, 244, 155, 300, 051, 152, 156, 205, 070, 015, 157, 040, 243, 479, 116, 122, 081, 160, 162.

Exit Polls

You must have seen that when an election takes place, the television networks provide election coverage. They also try to predict the results. This is done through *exit polls,* wherein a random sample of voters who exit the polling booths are asked whom they voted for. From the data of the sample of voters, the prediction is made.

Activity

- You have to analyse the trend of foodgrains production in India for the last fifty years. As it is difficult to include all the years, you have to select a sample of production of ten years. Using the Random Number Tables, how will you select your sample?

Non-Random Sampling

There may be a situation that you have to select 10 out of 100 households in a locality. You have to decide which household to select and which to reject. You may select the households conveniently situated or the households known to you or your friend. In this case, you are using your judgement (bias) in selecting 10 households. This way of selecting 10 out of 100 households is not a random selection. *In a non-random sampling method all the units of the population do not have an equal chance of being selected and convenience or judgement of the investigator plays an important role in selection of the sample.* They are mainly selected on the basis of judgment, purpose, convenience or quota and are non-random samples.

5. SAMPLING AND NON-SAMPLING ERRORS

Sampling Errors

The purpose of the sample is to take an estimate of the population. Sampling error refers to the differences between the sample estimate and the actual value of a characteristic of the population (that may be the average income, etc.). It is the error that occurs when you make an observation from the sample taken from the population. *Thus, the difference between the actual value of a parameter of the population (which is not known) and its estimate (from the sample) is the sampling error.* It is possible to reduce the magnitude of sampling error by taking a larger sample.

Example

Consider a case of incomes of 5 farmers of Manipur. The variable x (income of farmers) has measurements 500, 550, 600, 650, 700. We note that the *population average* of (500 + 550 + 600 + 650 + 700) ÷ 5 = 3000 ÷ 5 = 600.

Now, suppose we select a sample of two individuals where x has measurements of 500 and 600. The *sample average* is (500 + 600) ÷ 2 = 1100 ÷ 2 = 550.

Here, the *sampling error* of the *estimate* = 600 (true value) – 550 (estimate) = 50.

Non-Sampling Errors

Non-sampling errors are more serious than sampling errors because a sampling error can be minimised by taking a larger sample. It is difficult to minimise non-sampling error, even by taking a large sample. Even a *Census* can contain non-sampling errors. Some of the non-sampling errors are:

Errors in Data Acquisition

This type of error arises from recording of incorrect responses. Suppose, the teacher asks the students to measure the length of the teacher's table in the classroom. The measurement by the students may differ. The differences may occur due to differences in measuring tape, carelessness of the students etc. Similarly, suppose we want to collect data on prices of oranges. We know that prices vary from shop to shop and from market to market. Prices also vary according to the quality. Therefore, we can only consider the average prices. Recording mistakes can also take place as the enumerators or the respondents may commit errors in recording or transcripting the data, for example, he/she may record 13 instead of 31.

Non-Response Errors

Non-response occurs if an interviewer is unable to contact a person listed in the sample or a person from the sample refuses to respond. In this case, the sample observation may not be representative.

Sampling Bias

Sampling bias occurs when the sampling plan is such that some members of the target population could not possibly be included in the sample.

6. Census of India and NSSO

There are some agencies both at the national and state level, which collect, process and tabulate the statistical data. Some of the major agencies at the national level are Census of India, National Sample Survey Organisation (NSSO), Central Statistical Organisation (CSO), Registrar General of India (RGI), Directorate General of Commercial Intelligence and Statistics (DGCIS), Labour Bureau etc.

The Census of India provides the most complete and continuous demographic record of population. The Census is being regularly conducted every ten years since 1881. The first Census after Independence was held in 1951. The Census collects information on various aspects of population such as the size, density, sex ratio, literacy, migration, rural-urban distribution etc. Census in India is not merely a statistical operation, the data is interpreted and analysed in an interesting manner.

The NSSO was established by the government of India to conduct nation-wide surveys on socio-economic issues. The NSSO does continuous surveys in successive rounds. The data collected by NSSO surveys, on different socio economic subjects, are released through reports and its quarterly journal *Sarvekshana*. NSSO provides periodic estimates of literacy, school enrolment, utilisation of educational services, employment, unemployment, manufacturing and service sector enterprises, morbidity, maternity, child care, utilisation of the public distribution system etc. The NSS 59th round survey (January-December

2003) was on land and livestock holdings, debt and investment. The NSS 60th round survey (January–June 2004) was on morbidity and health care. The NSSO also undertakes the fieldwork of Annual survey of industries, conducts crop estimation surveys, collects rural and urban retail prices for compilation of consumer price index numbers.

7. Conclusion

Economic facts, expressed in terms of numbers, are called data. The purpose of data collection is to understand, explain and analyse a problem and causes behind it. Primary data is obtained by conducting a survey. Survey includes various steps, which need to be planned carefully. There are various agencies which collect, process, tabulate and publish statistical data. These can be used as secondary data. However, the choice of source of data and mode of data collection depends on the objective of the study.

Recap

- Data is a tool which helps in reaching a sound conclusion on any problem by providing information.
- Primary data is based on first hand information.
- Survey can be done by personal interviews, mailing questionnaires and telephone interviews.
- Census covers every individual/unit belonging to the population.
- Sample is a smaller group selected from the population from which the relevant information would be sought.
- In a random sampling, every individual is given an equal chance of being selected for providing information.
- Sampling error arises due to the difference between the actual population and the estimate.
- Non-sampling errors can arise in data acquisition, by non-response or by bias in selection.
- Census of India and National Sample Survey Organisation are two important agencies at the national level, which collect, process and tabulate data.

EXERCISES

1. Frame at least four appropriate multiple-choice options for following questions:
   (i) Which of the following is the most important when you buy a new dress?

(ii) How often do you use computers?
(iii) Which of the newspapers do you read regularly?
(iv) Rise in the price of petrol is justified.
(v) What is the monthly income of your family?

2. Frame five two-way questions (with 'Yes' or 'No').

3. (i) There are many sources of data (true/false).
(ii) Telephone survey is the most suitable method of collecting data, when the population is literate and spread over a large area (true/false).
(iii) Data collected by investigator is called the secondary data (true/false).
(iv) There is a certain bias involved in the non-random selection of samples (true/false).
(v) Non-sampling errors can be minimised by taking large samples (true/false).

4. What do you think about the following questions. Do you find any problem with these questions? If yes, how?
(i) How far do you live from the closest market?
(ii) If plastic bags are only 5 percent of our garbage, should it be banned?
(iii) Wouldn't you be opposed to increase in price of petrol?
(iv) (a) Do you agree with the use of chemical fertilizers?
(b) Do you use fertilizers in your fields?
(c) What is the yield per hectare in your field?

5. You want to research on the popularity of Vegetable Atta Noodles among children. Design a suitable questionnaire for collecting this information.

6. In a village of 200 farms, a study was conducted to find the cropping pattern. Out of the 50 farms surveyed, 50% grew only wheat. Identify the population and the sample here.

7. Give two examples each of sample, population and variable.

8. Which of the following methods give better results and why?
(a) Census (b) Sample

9. Which of the following errors is more serious and why?

(a) Sampling error (b) Non-Sampling error

10. Suppose there are 10 students in your class. You want to select three out of them. How many samples are possible?

11. Discuss how you would use the lottery method to select 3 students out of 10 in your class?

12. Does the lottery method always give you a random sample? Explain.

13. Explain the procedure of selecting a random sample of 3 students out of 10 in your class, by using random number tables.

14. Do samples provide better results than surveys? Give reasons for your answer.

CHAPTER 3

Organisation of Data

Studying this chapter should enable you to:

- *classify the data for further statistical analysis;*
- *distinguish between quantitative and qualitative classification;*
- *prepare a frequency distribution table;*
- *know the technique of forming classes;*
- *be familiar with the method of tally marking;*
- *differentiate between univariate and bivariate frequency distributions.*

1. INTRODUCTION

In the previous chapter you have learnt about how data is collected. You also came to know the difference between census and sampling. In this chapter, you will know how the data, that you collected, are to be classified. The purpose of classifying raw data is to bring order in them so that they can be subjected to further statistical analysis easily.

Have you ever observed your local junk dealer or *kabadiwallah* to whom you sell old newspapers, broken household items, empty glass bottles, plastics etc. He purchases these things from you and sells them to those who recycle them. But with so much junk in his shop it would be very difficult for him to manage his trade, if he had not organised them properly. To ease his situation he suitably groups or "classifies" various junk. He puts old newspapers together and

ties them with a rope. Then collects all empty glass bottles in a sack. He heaps the articles of metals in one corner of his shop and sorts them into groups like "iron", "copper", "aluminium", "brass" etc., and so on. In this way he groups his junk into different classes — "newspapers, "plastics", "glass", "metals" etc. — and brings order in them. Once his junk is arranged and classified, it becomes easier for him to find a particular item that a buyer may demand.

Likewise when you arrange your schoolbooks in a certain order, it becomes easier for you to handle them. You may classify them

according to subjects where each subject becomes a group or a class. So, when you need a particular book on history, for instance, all you need to do is to search that book in the group "History". Otherwise, you would have to search through your entire collection to find the particular book you are looking for.

While classification of objects or things saves our valuable time and effort, it is not done in an arbitrary manner. The *kabadiwallah* groups his junk in such a way that each group consists of similar items. For example, under the group "Glass" he would put empty bottles, broken mirrors and windowpanes etc. Similarly when you classify your history books under the group "History" you would not put a book of a different subject in that group. Otherwise the entire purpose of grouping would be lost. *Classification,* therefore, *is arranging or organising similar things into groups or classes.*

Activity

- Visit your local post-office to find out how letters are sorted. Do you know what the pin-code in a letter indicates? Ask your postman.

2. RAW DATA

Like the *kabadiwallah's* junk, the unclassified data or *raw data* are highly disorganised. They are often very large and cumbersome to handle. To draw meaningful conclusions from them is a tedious task because they do not yield to statistical methods easily. Therefore proper organisation and presentation of such data is needed before any systematic statistical analysis is undertaken. Hence after collecting data the next step is to organise and present them in a classified form.

Suppose you want to know the performance of students in mathematics and you have collected data on marks in mathematics of 100

students of your school. If you present them as a table, they may appear something like Table 3.1.

TABLE 3.1
Marks in Mathematics Obtained by 100 Students in an Examination

47	45	10	60	51	56	66	100	49	40
60	59	56	55	62	48	59	55	51	41
42	69	64	66	50	59	57	65	62	50
64	30	37	75	17	56	20	14	55	90
62	51	55	14	25	34	90	49	56	54
70	47	49	82	40	82	60	85	65	66
49	44	64	69	70	48	12	28	55	65
49	40	25	41	71	80	0	56	14	22
66	53	46	70	43	61	59	12	30	35
45	44	57	76	82	39	32	14	90	25

Or you could have collected data on the monthly expenditure on food of 50 households in your neighbourhood to know their average expenditure on food. The data collected, in that case, had you

presented as a table, would have resembled Table 3.2. Both Tables 3.1 and 3.2 are *raw* or *unclassified data*. In both the tables you find that numbers are not arranged in any order. Now if you are asked what are the highest marks in mathematics

TABLE 3.2
Monthly Household Expenditure (in Rupees) on Food of 50 Households

1904	1559	3473	1735	2760
2041	1612	1753	1855	4439
5090	1085	1823	2346	1523
1211	1360	1110	2152	1183
1218	1315	1105	2628	2712
4248	1812	1264	1183	1171
1007	1180	1953	1137	2048
2025	1583	1324	2621	3676
1397	1832	1962	2177	2575
1293	1365	1146	3222	1396

from Table 3.1 then you have to first arrange the marks of 100 students either in ascending or in descending order. That is a tedious task. It becomes more tedious, if instead of 100 you have the marks of a 1,000 students to handle. Similarly in Table 3.2, you would note that it is difficult for you to ascertain the average monthly expenditure of 50 households. And this difficulty will go up manifold if the number was larger – say, 5,000 households. Like our *kabadiwallah*, who would be distressed to find a particular item when his junk becomes large and disarranged, you would face a similar situation when you try to get any information from raw data that are large. In one word, therefore, it is a tedious task to pull information from large unclassified data.

The raw data are summarised, and made comprehensible by classification. When facts of similar characteristics are placed in the same class, it enables one to locate them easily, make comparison, and draw inferences without any difficulty. You

have studied in Chapter 2 that the Government of India conducts Census of population every ten years. The raw data of census are so large and fragmented that it appears an almost impossible task to draw any meaningful conclusion from them. But when the data of Census are classified according to gender, education, marital status, occupation, etc., the structure and nature of population of India is, then, easily understood.

The raw data consist of observations on variables. Each unit of raw data is an observation. In Table 3.1 an observation shows a particular value of the variable "marks of a student in mathematics". The raw data contain 100 observations on "marks of a student" since there are 100 students. In Table 3.2 it shows a particular value of the variable "monthly expenditure of a household on food". The raw data in it contain 50 observations on "monthly expenditure on food of a household" because there are 50 households.

Activity

- Collect data of total weekly expenditure of your family for a year and arrange it in a table. See how many observations you have. Arrange the data monthly and find the number of observations.

3. Classification of Data

The groups or classes of a classification can be done in various ways. Instead of classifying your books according to subjects – "History", "Geography", "Mathematics", "Science" etc. – you could have classified them author-wise in an alphabetical order. Or, you could have also classified them according to the year of publication. The way you want to classify them would depend on your requirement.

Likewise the raw data could be classified in various ways depending on the purpose in hand. They can be grouped according to time. Such a classification is known as a **Chronological Classification.** In such a classification, data are classified either in ascending or in descending order with reference to time such as years, quarters, months, weeks, etc. The following example shows the population of India classified in terms of years. The variable 'population' is a **Time Series** as it depicts a series of values for different years.

Example 1

Population of India (in crores)

Year	*Population (Crores)*
1951	35.7
1961	43.8
1971	54.6
1981	68.4
1991	81.8
2001	102.7

In **Spatial Classification** the data are classified with reference to geographical locations such as countries, states, cities, districts, etc**.** *Example 2* shows the yield of wheat in different countries.

Example 2

Yield of Wheat for Different Countries

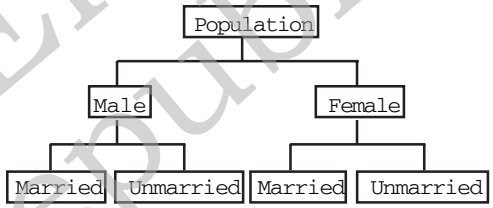
Country	*Yield of wheat (kg/acre)*
America	1925
Brazil	127
China	893
Denmark	225
France	439
India	862

Activities

- In the time-series of Example 1, in which year do you find the population of India to be the minimum. Find the year when it is the maximum.
- In Example 2, find the country whose yield of wheat is *slightly more* than that of India's. How much would that be in terms of percentage?
- Arrange the countries of Example 2 in the ascending order of yield. Do the same exercise for the descending order of yield.

Sometimes you come across characteristics that cannot be expressed quantitatively. Such characteristics are called *Qualities or Attributes.* For example, nationality, literacy, religion, gender, marital status, etc. They cannot be measured. Yet these *attributes* can be classified on the basis of either the presence or the absence of a qualitative characteristic. Such a classification of data on attributes is called a *Qualitative Classification.* In the following example, we find population of a country is grouped on the basis of the qualitative variable "gender". An observation could either be a male or a female. These two characteristics could be further classified on the basis of marital status (a qualitative variable) as given below:

Example 3



The classification at the first stage is based on the presence and absence of an attribute i.e. male or not male (female). At the second stage, each class – male and female, is further sub divided on the basis of the presence or absence of another attribute i.e. whether married or unmarried. On the other hand, characteristics like height, weight, age, income, marks of students, etc. are quantitative in nature. When the collected data of such characteristics are grouped into

Activity

- The objects around can be grouped as either living or non-living. Is it a quantitative classification?

classes, the classification is a *Quantitative Classification.*

Example 4

Frequency Distribution of Marks in Mathematics of 100 Students

Marks	*Frequency*
0–10	1
10–20	8
20–30	6
30–40	7
40–50	21
50–60	23
60–70	19
70–80	6
80–90	5
90–100	4
Total	*100*

Example 4 shows quantitative classification of the data of marks in mathematics of 100 students given in Table 3.1 as a Frequency Distribution.

Activity

- Express the values of frequency of Example 4 as proportion or percentage of total frequency. Note that frequency expressed in this way is known as *relative frequency*.
- In Example 4, which class has the maximum concentration of data? Express it as percentage of total observations. Which class has the minimum concentration of data?

4. VARIABLES: CONTINUOUS AND DISCRETE

A simple definition of variable, which you have read in the last chapter, does not tell you how it varies. Different variables vary differently and depending on the way they vary, they are broadly classified into *two* types:

(i) *Continuous* and
(ii) *Discrete.*

A *continuous variable* can take *any* numerical value. It may take integral values (1, 2, 3, 4, ...), fractional values (1/2, 2/3, 3/4, ...), and values that are not exact fractions ($\sqrt{2}$ =1.414, $\sqrt{3}$ =1.732, … , $\sqrt{7}$ =2.645). For example, the height of a student, as he/she grows say from 90 cm to 150 cm, would take all the values in between them. It can take values that are whole numbers like 90cm, 100cm, 108cm, 150cm. It can also take fractional values like 90.85 cm, 102.34 cm, 149.99cm etc. that are not whole numbers. Thus the variable "height" is capable of manifesting in every conceivable value and its values can also be broken down into infinite gradations. Other examples of a continuous variable are weight, time, distance, etc.

Unlike a continuous variable, a *discrete variable* can take only certain values. Its value changes only by finite "jumps". It "jumps" from one value to another but does not take any intermediate value between them. For example, a variable like the "number of students in a class", for different classes, would assume values that are only whole numbers. It cannot take

any fractional value like 0.5 because "half of a student" is absurd. Therefore it cannot take a value like 25.5 between 25 and 26. Instead its value could have been either 25 or 26. What we observe is that as its value changes from 25 to 26, the values in between them – the fractions are not taken by it. But do not have the impression that a discrete variable cannot take any fractional value. Suppose X is a variable that takes values like 1/8, 1/16, 1/32, 1/64, ... Is it a discrete variable? Yes, because though X takes fractional values it cannot take any value between two adjacent fractional values. It changes or "jumps" from 1/8 to 1/16 and from 1/16 to 1/32. But cannot take a value in between 1/8 and 1/16 or between 1/16 and 1/32

Activity

- Distinguish the following variables as continuous and discrete:
  Area, volume, temperature, number appearing on a dice, crop yield, population, rainfall, number of cars on road, age.

Earlier we have mentioned that example 4 is the frequency distribution of marks in mathematics of 100 students as shown in Table 3.1. It shows how the marks of 100 students are grouped into classes. You will be wondering as to how we got it from the raw data of Table 3.1. But, before we address this question, you must know what a frequency distribution is.

5. WHAT IS A FREQUENCY DISTRIBUTION?

A frequency distribution is a comprehensive way to classify raw data of a quantitative variable. It shows how the different values of a variable (here, the marks in mathematics scored by a student) are distributed in different classes along with their corresponding class frequencies. In this case we have ten classes of marks: 0–10, 10–20, … , 90–100. The term *Class Frequency* means the number of values in a particular class. For example, in the class 30–40 we find 7 values of marks from raw data in Table 3.1. They are 30, 37, 34, 30, 35, 39, 32. The frequency of the class: 30–40 is thus 7. But you might be wondering why 40–which is occurring *twice* in the raw data – is not included in the class 30–40. Had it been included the class frequency of 30–40 would have been 9 instead of 7. The puzzle would be clear to you if you are patient enough to read this chapter carefully. So carry on. You will find the answer yourself.

Each class in a frequency distribution table is bounded by *Class Limits.* Class limits are the two ends of a class. The lowest value is called the *Lower Class Limit* and the highest value the *Upper Class Limit.* For example, the class limits for the class: 60–70 are 60 and 70. Its lower class limit is 60 and its upper class limit is 70. *Class Interval or Class Width* is

the difference between the upper class limit and the lower class limit. For the class 60–70, the class interval is 10 (upper class limit *minus* lower class limit).

The *Class Mid-Point* or *Class Mark* is the middle value of a class. It lies halfway between the lower class limit and the upper class limit of a class and can be ascertained in the following manner:

Class Mid-Point or Class Mark = (Upper Class Limit + Lower Class Limit) /2 .....(1)

The class mark or mid-value of each class is used to represent the class. Once raw data are grouped into classes, individual observations are not used in further calculations. Instead, the class mark is used.

TABLE 3.3
The Lower Class Limits, the Upper Class Limits and the Class Mark

Class	*Frequency*	*Lower Class Limit*	*Upper Class Limit*	*Class Marks*
0–10	1	0	10	5
10–20	8	10	20	15
20–30	6	20	30	25
30–40	7	30	40	35
40–50	21	40	50	45
50–60	23	50	60	55
60–70	19	60	70	65
70–80	6	70	80	75
80–90	5	80	90	85
90–100	4	90	100	95

Frequency Curve is a graphic representation of a frequency distribution. Fig. 3.1 shows the diagrammatic presentation of the frequency distribution of the data in our example above. To obtain the frequency curve we plot the class marks on the X-axis and frequency on the Y-axis.

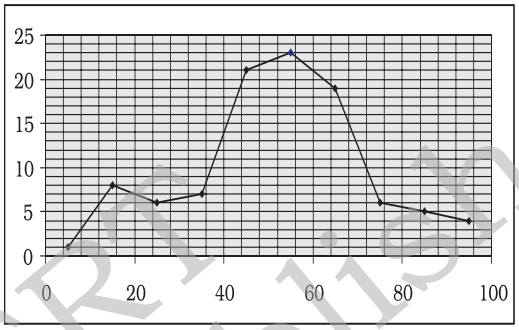


Fig.3.1: *Diagrammatic Presentation of Frequency Distribution of Data.*

How to prepare a Frequency Distribution?

While preparing a frequency distribution from the raw data of Table 3.1, the following four questions need to be addressed:

1. How many classes should we have?
2. What should be the size of each class?
3. How should we determine the class limits?
4. How should we get the frequency for each class?

How many classes should we have?

Before we determine the number of classes, we first find out as to what extent the variable in hand changes in value. Such variations in variable's value are captured by its *range*. The *Range is the difference between the largest and the smallest values of the*

variable. A large range indicates that the values of the variable are widely spread. On the other hand, a small range indicates that the values of the variable are spread narrowly. In our example the range of the variable "marks of a student" are 100 because the minimum marks are 0 and the maximum marks 100. It indicates that the variable has a large variation.

After obtaining the value of range, it becomes easier to determine the number of classes once we decide the class interval. Note that *range is the sum of all class intervals.* If the class intervals are equal then range is the product of the number of classes and class interval of a single class.

Range = Number of Classes × Class Interval .....(2)

Activities

Find the range of the following:

- population of India in Example 1,
- yield of wheat in Example 2.

Given the value of range, the number of classes would be large if we choose small class intervals. A frequency distribution with too many classes would look too large. Such a distribution is not easy to handle. So we want to have a reasonably compact set of data. On the other hand, given the value of range if we choose a class interval that is too large then the number of classes becomes too small. The data set then may be too compact and we may not like the loss of information about its diversity. For example, suppose the range is 100 and the class interval is 50. Then the number of classes would be just 2 (i.e.100/50 = 2). Though there is no hard-and-fast rule to determine the number of classes, the rule of thumb often used is that the number of classes should be between 5 and 15. In our example we have chosen to have 10 classes. Since the range is 100 and the class interval is 10, the number of classes is 100/10 =10.

What should be the size of each class?

The answer to this question depends on the answer to the previous question. The equality (2) shows that given the range of the variable, we can determine the number of classes once we decide the class interval. Similarly, we can determine the class interval once we decide the number of classes. Thus we find that these two decisions are inter-linked with one another. We cannot decide on one without deciding on the other.

In Example 4, we have the number of classes as 10. Given the value of range as 100, the class intervals are automatically 10 by the equality (2). Note that in the present context we have chosen class intervals that are equal in magnitude. However we could have chosen class intervals that are not of equal magnitude. In that case, the classes would have been of unequal width.

How should we determine the class limits?

When we classify raw data of a continuous variable as a frequency distribution, we in effect, group the individual observations into classes. *The value of the upper class limit of a class is obtained by adding the class interval with the value of the lower class limit of that class.* For example, the upper class limit of the class 20–30 is 20 + 10 = 30 where 20 is the lower class limit and 10 is the class interval. This method is repeated for other classes as well.

But how do we decide the lower class limit of the first class? That is to say, why 0 is the lower class limit of the first class: 0–10? It is because we chose the minimum value of the variable as the lower limit of the first class. In fact, we could have chosen a value less than the minimum value of the variable as the lower limit of the first class. Similarly, for the upper class limit for the last class we could have chosen a value greater than the maximum value of the variable. It is important to note that, when a frequency distribution is being constructed, the class limits should be so chosen that the mid-point or class mark of each class coincide, as far as possible, with any value around which the data tend to be concentrated.

In our example on marks of 100 students, we chose 0 as the lower limit of the first class: 0–10 because the minimum marks were 0. And that is why, we could not have chosen 1 as the lower class limit of that class. Had we done that we would have excluded the observation 0. The upper class limit of the first class: 0–10 is then obtained by adding class interval with lower class limit of the class. Thus the upper class limit of the first class becomes 0 + 10 = 10. And this procedure is followed for the other classes as well.

Have you noticed that the upper class limit of the first class is equal to the lower class limit of the second class? And both are equal to 10. This is observed for other classes as well. Why? The reason is that we have used the *Exclusive Method* of classification of raw data. Under the method we form classes in such a way that the lower limit of a class coincides with the upper class limit of the previous class.

The problem, we would face next, is how do we classify an observation that is not only equal to the upper class limit of a particular class but is also equal to the lower class limit of the next class. For example, we find observation 30 to be equal to the upper class limit of the class 20–30 and it is equal to the lower class limit of class 30–40. Then, in which of the two classes: 20–30 or 30–40 should we put the observation 30? We can put it either in class 20–30 or in class 30–40. It is a dilemma that one commonly faces while classifying data in overlapping classes. This problem is solved by the rule of classification in the *Exclusive Method.*

Exclusive Method

The classes, by this method, are formed in such a way that the upper class limit of one class equals the lower class limit of the next class. In this way the continuity of the data is maintained. That is why this method of classification is most suitable in case of data of a continuous variable. *Under the method, the upper class limit is excluded but the lower class limit of a class is included in the interval.* Thus an observation that is exactly equal to the upper class limit, according to the method, would not be included in that class but would be included in the next class. On the other hand, if it were equal to the lower class limit then it would be included in that class. In our example on marks of students, the observation 40, that occurs twice, in the raw data of Table 3.1 is not included in the class: 30–40. It is included in the next class: 40–50. That is why we find the frequency corresponding to the class 30–40 to be 7 instead of 9.

There is another method of forming classes and it is known as the Inclusive Method of classification.

Inclusive Method

In comparison to the *exclusive method*, the *Inclusive Method* does not exclude the upper class limit in a class interval. It includes the upper class in a class. Thus both class limits are parts of the class interval.

For example, in the frequency distribution of Table 3.4 we include

TABLE 3.4
Frequency Distribution of Incomes of 550 Employees of a Company

Income (Rs)	*Number of Employees*
800–899	50
900–999	100
1000–1099	200
1100–1199	150
1200–1299	40
1300–1399	10
Total	*550*

in the class: 800–899 those employees whose income is either Rs 800, or between Rs 800 and Rs 899, or Rs 899. If the income of an employee is exactly Rs 900 then he is put in the next class: 900–999.

Adjustment in Class Interval

A close observation of the *Inclusive Method* in Table 3.4 would show that though the variable "income" is a continuous variable, no such continuity is maintained when the classes are made. We find *"gap"* or discontinuity between the upper limit of a class and the lower limit of the next class. For example, between the upper limit of the first class: 899 and the lower limit of the second class: 900, we find a "gap" of 1. Then how do we ensure the continuity of the variable while classifying data? This is achieved by making an adjustment in the class interval. The adjustment is done in the following way:

1. Find the difference between the lower limit of the second class and the upper limit of the first class. For example, in Table 3.4 the lower limit of the second class is 900 and

the upper limit of the first class is 899. The difference between them is 1, i.e. (900 – 899 = 1)

2. Divide the difference obtained in (1) by two i.e. (1/2 = 0.5)
3. Subtract the value obtained in (2) from lower limits of all classes (lower class limit – 0.5)
4. Add the value obtained in (2) to upper limits of all classes (upper class limit + 0.5).

After the adjustment that restores continuity of data in the frequency distribution, the Table 3.4 is modified into Table 3.5

After the adjustments in class limits, the equality (1) that determines the value of class-mark would be modified as the following:

Adjusted Class Mark = (Adjusted Upper Class Limit + Adjusted Lower Class Limit)/2.

TABLE 3.5
Frequency Distribution of Incomes of 550 Employees of a Company

Income (Rs)	*Number of Employees*
799.5–899.5	50
899.5–999.5	100
999.5–1099.5	200
1099.5–1199.5	150
1199.5–1299.5	40
1299.5–1399.5	10
Total	*550*

How should we get the frequency for each class?

In simple terms, *frequency of an observation means how many times that observation occurs in the raw data.* In our Table 3.1, we observe that the value 40 occurs thrice; 0 and 10 occur only once; 49 occurs five times and so on. Thus the frequency of 40 is 3, 0 is 1, 10 is 1, 49 is 5 and so on. But when the data are grouped into

TABLE 3.6
Tally Marking of Marks of 100 Students in Mathematics

Class	*Observations*	*Tally Mark*	*Frequency*	*Class Mark*
0–10	0	/	1	5
10–20	10, 14, 17, 12, 14, 12, 14, 14	𝍸 ///	8	15
20–30	25, 25, 20, 22, 25, 28	𝍸 /	6	25
30–40	30, 37, 34, 39, 32, 30, 35,	𝍸 //	7	35
40–50	47, 42, 49, 49, 45, 45, 47, 44, 40, 44, 49, 46, 41, 40, 43, 48, 48, 49, 49, 40, 41	𝍸 𝍸 𝍸 𝍸 /	21	45
50–60	59, 51, 53, 56, 55, 57, 55, 51, 50, 56, 59, 56, 59, 57, 59, 55, 56, 51, 55, 56, 55, 50, 54	𝍸 𝍸 𝍸 𝍸 ///	23	55
60–70	60, 64, 62, 66, 69, 64, 64, 60, 66, 69, 62, 61, 66, 60, 65, 62, 65, 66, 65	𝍸 𝍸 𝍸 ////	19	65
70–80	70, 75, 70, 76, 70, 71	𝍸 //	6	75
80–90	82, 82, 82, 80, 85	𝍸	5	85
90–100	90, 100, 90, 90	////	4	95
	Total		*100*	

classes as in example 3, the Class Frequency refers to the number of values in a particular class. The counting of class frequency is done by tally marks against the particular class.

Finding class frequency by tally marking

A tally (/) is put against a class for each student whose marks are included in that class. For example, if the marks obtained by a student are 57, we put a tally (/) against class 50–60. If the marks are 71, a tally is put against the class 70–80. If someone obtains 40 marks, a tally is put against the class 40–50. Table 3.6 shows the tally marking of marks of 100 students in mathematics from Table 3.1.

The counting of tally is made easier when four of them are put as //// and the fifth tally is placed across them as ////. Tallies are then counted as groups of five. So if there are 16 tallies in a class, we put them as //// //// //// / for the sake of convenience. Thus frequency in a class is equal to the number of tallies against that class.

Loss of Information

The classification of data as a frequency distribution has an inherent shortcoming. While it summarises the raw data making it concise and comprehensible, it does not show the details that are found in raw data. There is a loss of information in classifying raw data though much is gained by summarising it as a classified data. Once the data are grouped into classes, an individual observation has no significance in further statistical calculations. In Example 4, the class 20–30 contains 6 observations: 25, 25, 20, 22, 25 and 28. So when these data are grouped as a class 20–30 in the frequency distribution, the latter provides only the number of records in that class (i.e. frequency = 6) but not their actual values. *All values in this class are assumed to be equal to the middle value of the class interval or class mark (i.e. 25). Further statistical calculations are based only on the values of class mark and not on the values of the observations in that class.* This is true for other classes as well. Thus the use of class mark instead of the actual values of the observations in statistical methods involves considerable loss of information.

Frequency distribution with unequal classes

By now you are familiar with frequency distributions of equal class intervals. You know how they are constructed out of raw data. But in some cases frequency distributions with unequal class intervals are more appropriate. If you observe the frequency distribution of Example 4, as in Table 3.6, you will notice that most of the observations are concentrated in classes 40–50, 50–60 and 60–70. Their respective frequen-

cies are 21, 23 and 19. It means that out of 100 observations, 63 (21+23+19) observations are concentrated in these classes. These classes are densely populated with observations. Thus, 63 percent of data lie between 40 and 70. The remaining 37 percent of data are in classes 0–10, 10–20, 20–30, 30–40, 70–80, 80–90 and 90–100. These classes are sparsely populated with observations. Further you will also notice that observations in these classes deviate more from their respective class marks than in comparison to those in other classes. But if classes are to be formed in such a way that class marks coincide, as far as possible, to a value around which the observations in a class tend to concentrate, then in that case unequal class interval is more appropriate.

Table 3.7 shows the same frequency distribution of Table 3.6 in terms of unequal classes. Each of the classes 40–50, 50–60 and 60–70 are split into *two* classes. The class 40–50 is divided into 40–45 and 45–50. The class 50–60 is divided into 50– 55 and 55–60. And class 60–70 is divided into 60–65 and 65–70. The new classes 40–45, 45–50, 50–55, 55–60, 60–65 and 65–70 have class interval of 5. The other classes: 0–10, 10–20, 20–30, 30–40, 70–80, 80–90 and 90–100 retain their old class interval of 10. The last column of this table shows the new values of class marks for these classes. Compare them with the old values of class marks in Table 3.6. Notice that the observations in these classes deviated more from their old class mark values than their new class mark values. Thus the new class mark values are more representative of the data in these classes than the old values.

TABLE 3.7

Frequency Distribution of Unequal Classes

Class	*Observations*	*Frequency*	*Class Mark*
0–10	0	1	5
10–20	10, 14, 17, 12, 14, 12, 14, 14	8	15
20–30	25, 25, 20, 22, 25, 28	6	25
30–40	30, 37, 34, 39, 32, 30, 35,	7	35
40–45	42, 44, 40, 44, 41, 40, 43, 40, 41	9	42.5
45–50	47, 49, 49, 45, 45, 47, 49, 46, 48, 48, 49, 49	12	47.5
50–55	51, 53, 51, 50, 51, 50, 54	7	52.5
55–60	59, 56, 55, 57, 55, 56, 59, 56, 59, 57, 59, 55, 56, 55, 56, 55	16	57.5
60–65	60, 64, 62, 64, 64, 60, 62, 61, 60, 62,	10	62.5
65–70	66, 69, 66, 69, 66, 65, 65, 66, 65	9	67.5
70–80	70, 75, 70, 76, 70, 71	6	75
80–90	82, 82, 82, 80, 85	5	85
90–100	90, 100, 90, 90	4	95
	Total	*100*	

Figure 3.2 shows the frequency curve of the distribution in Table 3.7. The class marks of the table are plotted on X-axis and the frequencies are plotted on Y-axis.

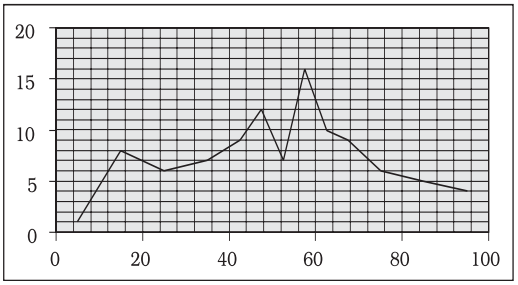


Fig. 3.2: Frequency Curve

Activity

- If you compare Figure 3.2 with Figure 3.1, what do you observe? Do you find any difference between them? Can you explain the difference?

Frequency array

So far we have discussed the classification of data for a continuous variable using the example of percentage marks of 100 students in mathematics. For a discrete variable, the classification of its data is known as a *Frequency Array*. Since a discrete variable takes values and not intermediate fractional values between two integral values, we have frequencies that correspond to each of its integral values.

The example in Table 3.8 illustrates a *Frequency Array*.

TABLE 3.8
Frequency Array of the Size of Households

Size of the Household	*Number of Households*
1	5
2	15
3	25
4	35
5	10
6	5
7	3
8	2
Total	*100*

The variable "size of the household" is a discrete variable that only takes integral values as shown in the table. Since it does not take any fractional value between two adjacent integral values, there are no classes in this frequency array. Since there are no classes in a frequency array there would be no class intervals. As the classes are absent in a discrete frequency distribution, there is no class mark as well.

6. Bivariate Frequency Distribution

The frequency distribution of a single variable is called a Univariate Distribution. The example 3.3 shows the univariate distribution of the single variable "marks of a student". *A Bivariate Frequency Distribution is the frequency distribution of two variables.*

Table 3.9 shows the frequency distribution of two variable sales and advertisement expenditure (in Rs. lakhs) of 20 companies. The values of sales are classed in different columns

TABLE 3.9
Bivariate Frequency Distribution of Sales (in Lakh Rs) and Advertisement Expenditure (in Thousand Rs) of 20 Firms

	115–125	*125–135*	*135–145*	*145–155*	*155–165*	*165–175*	*Total*
62–64	2	1					3
64–66	1		3				4
66–68	1	1	2	1			5
68–70		2		2			4
70–72		1	1		1	1	4
Total	*4*	*5*	*6*	*3*	*1*	*1*	*20*

and the values of advertisement expenditure are classed in different rows. Each cell shows the frequency of the corresponding row and column values. For example, there are 3 firms whose sales are between Rs 135–145 lakhs and their advertisement expenditures are between Rs 64–66 thousands. The use of a bivariate distribution would be taken up in Chapter 8 on correlation.

7. CONCLUSION

The data collected from primary and secondary sources are raw or unclassified. Once the data is collected, the next step is to classify them for further statistical analysis. Classification brings order in the data.

The chapter enables you to know how data can be classified through a frequency distribution in a comprehensive manner. Once you know the techniques of classification, it will be easy for you to construct a frequency distribution, both for continuous and discrete variables.

Recap

- Classification brings order to raw data.
- A Frequency Distribution shows how the different values of a variable are distributed in different classes along with their corresponding class frequencies.
- The upper class limit is *excluded* but lower class limit is included in the Exclusive Method.
- Both the upper and the lower class limits are *included* in the Inclusive Method.
- In a Frequency Distribution, further statistical calculations are based only on the class mark values, instead of values of the observations.
- The classes should be formed in such a way that the class mark of each class comes as close as possible, to a value around which the observations in a class tend to concentrate.

EXERCISES

1. Which of the following alternatives is true?

(i) The class midpoint is equal to:
   (a) The average of the upper class limit and the lower class limit.
   (b) The product of upper class limit and the lower class limit.
   (c) The ratio of the upper class limit and the lower class limit.
   (d) None of the above.

(ii) The frequency distribution of two variables is known as
   (a) Univariate Distribution
   (b) Bivariate Distribution
   (c) Multivariate Distribution
   (d) None of the above

(iii) Statistical calculations in classified data are based on
   (a) the actual values of observations
   (b) the upper class limits
   (c) the lower class limits
   (d) the class midpoints

(iv) Under Exclusive method,
   (a) the upper class limit of a class is excluded in the class interval
   (b) the upper class limit of a class is included in the class interval
   (c) the lower class limit of a class is excluded in the class interval
   (d) the lower class limit of a class is included in the class interval

(v) Range is the
   (a) difference between the largest and the smallest observations
   (b) difference between the smallest and the largest observations
   (c) average of the largest and the smallest observations
   (d) ratio of the largest to the smallest observation

2. Can there be any advantage in classifying things? Explain with an example from your daily life.

3. What is a variable? Distinguish between a discrete and a continuous variable.

4. Explain the 'exclusive' and 'inclusive' methods used in classification of data.

5. Use the data in Table 3.2 that relate to monthly household expenditure (in Rs) on food of 50 households and
   (i) Obtain the range of monthly household expenditure on food.
   (ii) Divide the range into appropriate number of class intervals and obtain the frequency distribution of expenditure.
   (iii) Find the number of households whose monthly expenditure on food is
      (a) less than Rs 2000
      (b) more than Rs 3000

(c) between Rs 1500 and Rs 2500

6. In a city 45 families were surveyed for the number of domestic appliances they used. Prepare a frequency array based on their replies as recorded below.

1	3	2	2	2	2	1	2	1	2	2	3	3	3	3
3	3	2	3	2	2	6	1	6	2	1	5	1	5	3
2	4	2	7	4	2	4	3	4	2	0	3	1	4	3

7. What is 'loss of information' in classified data?

8. Do you agree that classified data is better than raw data?

9. Distinguish between univariate and bivariate frequency distribution.

10. Prepare a frequency distribution by inclusive method taking class interval of 7 from the following data:

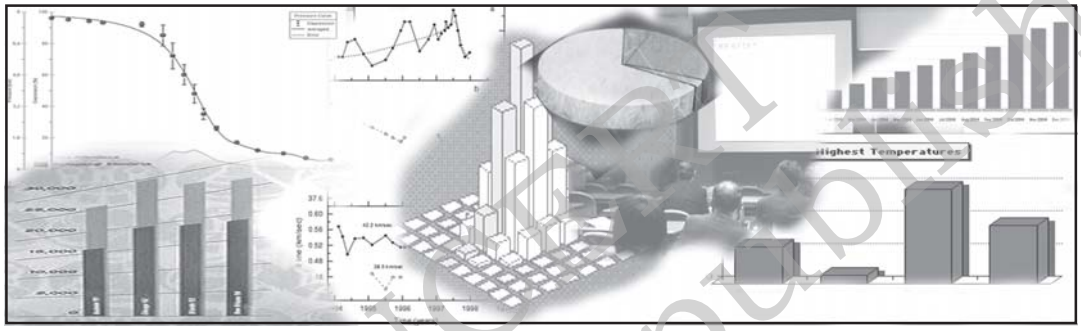
28	17	15	22	29	21	23	27	18	12	7	2	9	4	6
1	8	3	10	5	20	16	12	8	4	33	27	21	15	9
3	36	27	18	9	2	4	6	32	31	29	18	14	13	
15	11	9	7	1	5	37	32	28	26	24	20	19	25	
19	20													

Suggested Activity

- From your old mark-sheets find the marks that you obtained in mathematics in the previous classes. Arrange them year-wise. Check whether the marks you have secured in the subject is a variable or not. Also see, if over the years, you have improved in mathematics.

CHAPTER 4

Presentation of Data



Studying this chapter should enable you to:

- *present data using tables;*
- *represent data using appropriate diagrams.*

1. INTRODUCTION

You have already learnt in previous chapters how data are collected and organised. As data are generally voluminous, they need to be put in a compact and presentable form. This chapter deals with presentation of data precisely so that the voluminous data collected could be made usable readily and are easily comprehended. There are generally three forms of presentation of data:

- Textual or Descriptive presentation
- Tabular presentation
- Diagrammatic presentation.

2. TEXTUAL PRESENTATION OF DATA

In textual presentation, data are described within the text. When the quantity of data is not too large this form of presentation is more suitable. Look at the following cases:

Case 1

In a *bandh* call given on 08 September 2005 protesting the hike in prices of petrol and diesel, 5 petrol pumps were found open and 17 were closed whereas 2 schools were closed and remaining 9 schools were found open in a town of Bihar.

Case 2

Census of India 2001 reported that Indian population had risen to 102 crore of which only 49 crore were females against 53 crore males. 74 crore people resided in rural India and only 28 crore lived in towns or cities. While there were 62 crore non-worker population against 40 crore workers in the entire country, urban population had an even higher share of non-workers (19 crores) against the workers (9 crores) as compared to the rural population where there were 31 crore workers out of a 74 crore population....

In both the cases data have been presented only in the text. A serious drawback of this method of presentation is that one has to go through the complete text of presentation for comprehension but at the same time, it enables one to emphasise certain points of the presentation.

3. TABULAR PRESENTATION OF DATA

In a tabular presentation, data are presented in rows (read horizontally) and columns (read vertically). For example see Table 4.1 below tabulating information about literacy rates. It has 3 rows (for male, female and total) and 3 columns (for urban, rural and total). It is called a 3 × 3 Table giving 9 items of information in 9 boxes called the "cells" of the Table. Each cell gives information that relates an attribute of gender ("male", "female" or total) with a number (literacy percentages of rural people, urban people and total). The most important advantage of tabulation is that it organises data for further statistical treatment and decision-making. Classification used in tabulation is of four kinds:

- Qualitative
- Quantitative
- Temporal and
- Spatial

Qualitative classification

When classification is done according to qualitative characteristics like social status, physical status, nationality, etc., it is called qualitative classification. For example, in Table 4.1 the characteristics for classification are sex and location which are qualitative in nature.

TABLE 4.1
Literacy in Bihar by sex and location (per cent)

	Location		*Total*
Sex	*Rural*	*Urban*	
Male	57.70	80.80	60.32
Female	30.03	63.30	33.57
Total	*44.42*	*72.71*	*47.53*

Source: *Census of India 2001, Provisional Population Totals.*

Quantitative classification

In quantitative classification, the data are classified on the basis of

characteristics which are quantitative in nature. In other words these characteristics can be measured quantitatively. For example, age, height, production, income, etc are quantitative characteristics. Classes are formed by assigning limits called class limits for the values of the characteristic under consideration. An example of quantitative classification is Table 4.2.

TABLE 4.2
Distribution of 542 respondents by their age in an election study in Bihar

Age group (yrs)	*No. of respondents*	*Per cent*
20–30	3	0.55
30–40	61	11.25
40–50	132	24.35
50–60	153	28.24
60–70	140	25.83
70–80	51	9.41
80–90	2	0.37
All	*542*	*100.00*

Source: *Assembly election Patna central constituency 2005, A.N. Sinha Institute of Social Studies, Patna.*

Here classifying characteristic is age in years and is quantifiable.

Activities

- Construct a table presenting data on preferential liking of the students of your class for Star News, Zee News, BBC World, CNN, Aaj Tak and DD News.
- Prepare a table of
  (i) heights (in cm) and
  (ii) weights (in kg) of students of your class.

Temporal classification

In this classification time becomes the classifying variable and data are categorised according to time. Time may be in hours, days, weeks, months, years, etc. For example, see Table 4.3.

TABLE 4.3
Yearly sales of a tea shop from 1995 to 2000

Years	*Sale (Rs in lakhs)*
1995	79.2
1996	81.3
1997	82.4
1998	80.5
1999	100.2
2000	91.2

Data Source: *Unpublished data.*

In this table the classifying characteristic is year and takes values in the scale of time.

Activity

- Go to your library and collect data on the number of books in economics, the library had at the end of the year for the last ten years and present the data in a table.

Spatial classification

When classification is done in such a way that place becomes the classifying variable, it is called spatial classification. The place may be a village/town, block, district, state, country, etc.

Here the classifying characteristic is country of the world. Table 4.4 is an example of spatial classification.

TABLE 4.4
Export from India to rest of the world in one year as share of total export (per cent)

Destination	*Export share*
USA	21.8
Germany	5.6
Other EU	14.7
UK	5.7
Japan	4.9
Russia	2.1
Other East Europe	0.6
OPEC	10.5
Asia	19.0
Other LDCs	5.6
Others	9.5
All	*100.0*

(**Total Exports**: *US $ 33658.5 million*)

Activity

- Construct a table presenting data collected from students of your class according to their native states/residential locality.

4. TABULATION OF DATA AND PARTS OF A TABLE

To construct a table it is important to learn first what are the parts of a good statistical table. When put together in a systematically ordered manner these parts form a table. The most simple way of conceptualising a table may be data presented in rows and columns alongwith some explanatory notes. Tabulation can be done using one-way, two-way or three-way classification depending upon the number of characteristics involved. A good table should essentially have the following:

(i) Table Number

Table number is assigned to a table for identification purpose. If more than one table is presented, it is the table number that distinguishes one table from another. It is given at the top or at the beginning of the title of the table. Generally, table numbers are whole numbers in ascending order if there are many tables in a book. Subscripted numbers like 1.2, 3.1, etc. are also in use for identifying the table according to its location. For example, Table number 4.5 may read as fifth table of the fourth chapter and so on. (See Table 4.5)

(ii) Title

The title of a table narrates about the contents of the table. It has to be very clear, brief and carefully worded so that the interpretations made from the table are clear and free from any ambiguity. It finds place at the head of the table succeeding the table number or just below it. (See Table 4.5).

(iii) Captions or Column Headings

At the top of each column in a table a column designation is given to explain figures of the column. This is called caption or column heading. (See Table 4.5)

(iv) Stubs or Row Headings

Like a caption or column heading each row of the table has to be given a heading. The designations of the rows are also called stubs or stub items, and the complete left column is known as

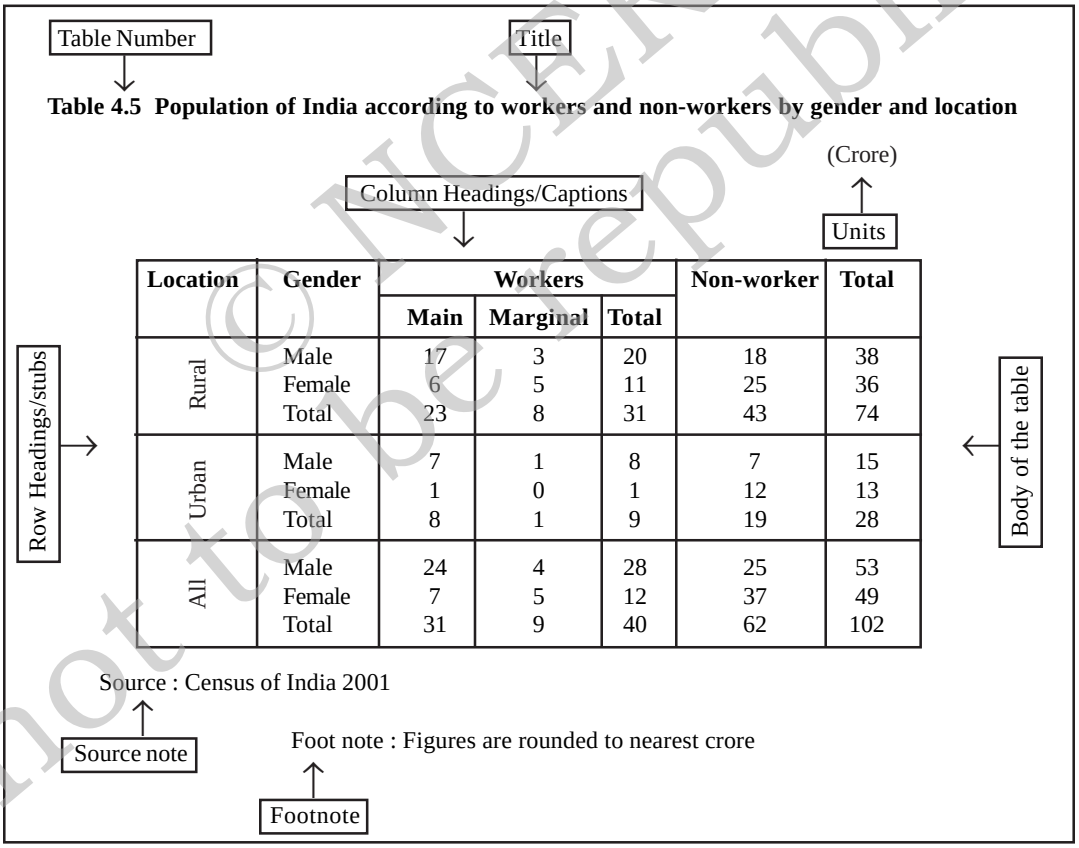
stub column. A brief description of the row headings may also be given at the left hand top in the table. (See Table 4.5).

(v) Body of the Table

Body of a table is the main part and it contains the actual data. Location of any one figure/data in the table is fixed and determined by the row and column of the table. For example, data in the second row and fourth column indicate that 25 crore females in rural India were non-workers in 2001. (See Table 4.5).

(vi) Unit of Measurement

The unit of measurement of the figures in the table (actual data) should always be stated alongwith the title if the unit does not change throughout the table. If different units are there for rows or columns of the table, these units must be stated alongwith 'stubs' or 'captions'. If figures are large, they should be rounded up and the method



Location	Gender	Workers			Non-worker	Total
		Main	Marginal	Total		
Rural	Male	17	3	20	18	38
	Female	6	5	11	25	36
	Total	23	8	31	43	74
Urban	Male	7	1	8	7	15
	Female	1	0	1	12	13
	Total	8	1	9	19	28
All	Male	24	4	28	25	53
	Female	7	5	12	37	49
	Total	31	9	40	62	102

(Note : Table 4.5 presents the same data in tabular form already presented through *case 2* in textual presentation of data)

of rounding should be indicated (See Table 4.5).

(vii) **Source Note**

It is a brief statement or phrase indicating the source of data presented in the table. If more than one source is there, all the sources are to be written in the source note. Source note is generally written at the bottom of the table. (See Table 4.5).

(viii) **Footnote**

Footnote is the last part of the table. Footnote explains the specific feature of the data content of the table which is not self explanatory and has not been explained earlier.

Activities

- How many rows and columns are essentially required to form a table?
- Can the column/row headings of a table be quantitative?

5. Diagrammatic Presentation of Data

This is the third method of presenting data. This method provides the quickest understanding of the actual situation to be explained by data in comparison to tabular or textual presentations. Diagrammatic presentation of data translates quite effectively the highly abstract ideas contained in numbers into more concrete and easily comprehensible form.

Diagrams may be less accurate but are much more effective than tables in presenting the data.

There are various kinds of diagrams in common use. Amongst them the important ones are the following:

(i) Geometric diagram
(ii) Frequency diagram
(iii) Arithmetic line graph

Geometric Diagram

Bar diagram and pie diagram come in the category of geometric diagram for presentation of data. The bar diagrams are of three types – simple, multiple and component bar diagrams.

Bar Diagram

Simple Bar Diagram

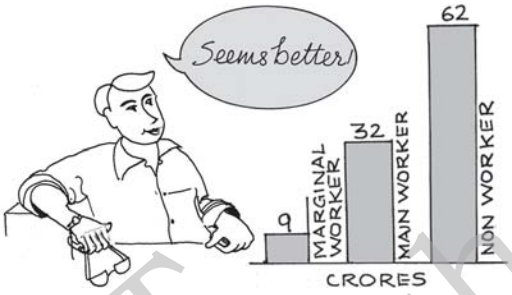
Bar diagram comprises a group of equispaced and equiwidth rectangular bars for each class or category of data. Height or length of the bar reads the magnitude of data. The lower end of the bar touches the base line such that the height of a bar starts from the zero unit. Bars of a bar diagram can be visually compared by their relative height and accordingly data are comprehended quickly. Data for this can be of frequency or non-frequency type. In non-frequency type data a particular characteristic, say production, yield, population, etc. at various points of time or of different states are noted and corresponding bars are made of the respective heights according to the values of the characteristic to construct the diagram. The values of the characteristics (measured or counted)

retain the identity of each value. Figure 4.1 is an example of a bar diagram.

Activity

- You had constructed a table presenting the data about the students of your class. Draw a bar diagram for the same table.

Different types of data may require different modes of diagrammatical representation. Bar diagrams are suitable both for frequency type and non-frequency type variables and attributes. Discrete variables like family size, spots on a dice, grades in an examination, etc. and attributes such as gender, religion, caste, country, etc. can be represented by bar diagrams. Bar diagrams are more convenient for non-frequency data such as income-expenditure profile, export/imports over the years, etc.



A category that has a longer bar (literacy of Kerala) than another category (literacy of West Bengal), has more of the measured (or enumerated) characteristics than the other. Bars (also called columns) are usually used in time series data (food grain produced between 1980-2000, decadal variation in work participation

TABLE 4.6
Literacy Rates of Major States of India

	2001			*1991*		
Major Indian States	*Person*	*Male*	*Female*	*Person*	*Male*	*Female*
Andhra Pradesh (AP)	60.5	70.3	50.4	44.1	55.1	32.7
Assam (AS)	63.3	71.3	54.6	52.9	61.9	43.0
Bihar (BR)	47.0	59.7	33.1	37.5	51.4	22.0
Jharkhand (JH)	53.6	67.3	38.9	41.4	55.8	31.0
Gujarat (GJ)	69.1	79.7	57.8	61.3	73.1	48.6
Haryana (HR)	67.9	78.5	55.7	55.8	69.1	40.4
Karnataka (KA)	66.6	76.1	56.9	56.0	67.3	44.3
Kerala (KE)	90.9	94.2	87.7	89.8	93.6	86.2
Madhya Pradesh (MP)	63.7	76.1	50.3	44.7	58.5	29.4
Chhattisgarh (CH)	64.7	77.4	51.9	42.9	58.1	27.5
Maharashtra (MR)	76.9	86.0	67.0	64.9	76.6	52.3
Orissa (OR)	63.1	75.3	50.5	49.1	63.1	34.7
Punjab (PB)	69.7	75.2	63.4	58.5	65.7	50.4
Rajasthan (RJ)	60.4	75.7	43.9	38.6	55.0	20.4
Tamil Nadu (TN)	73.5	82.4	64.4	62.7	73.7	51.3
Uttar Pradesh (UP)	56.3	68.8	42.2	40.7	54.8	24.4
Uttaranchal (UT)	71.6	83.3	59.6	57.8	72.9	41.7
West Bengal (WB)	68.6	77.0	59.6	57.7	67.8	46.6
India	**64.8**	**75.3**	**53.7**	**52.2**	**64.1**	**39.3**

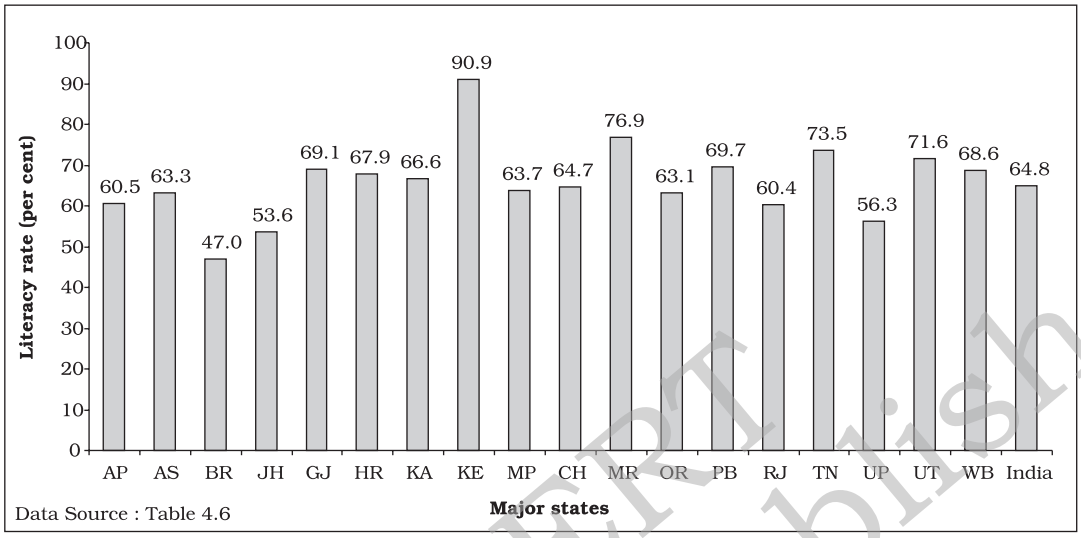


Fig. 4.1: *Bar diagram showing literacy rates (person) of major states of India, 2001.*

rate, registered unemployed over the years, literacy rates, etc.) (Fig 4.2).

Bar diagrams can have different forms such as multiple bar diagram and component bar diagram.

Activities

- How many states (among the major states of India) had higher female literacy rate than the national average in 2001?
- Has the gap between maximum and minimum female literacy rates over the states in two consecutive census years 2001 and 1991 declined?

Multiple Bar Diagram

Multiple bar diagrams (Fig.4.2) are used for comparing two or more sets of data, for example income and expenditure or import and export for different years, marks obtained in different subjects in different classes, etc.

Component Bar Diagram

Component bar diagrams or charts (Fig.4.3), also called sub-diagrams, are very useful in comparing the sizes of different component parts (the elements or parts which a thing is made up of) and also for throwing light on the relationship among these integral parts. For example, sales proceeds from different products, expenditure pattern in a typical Indian family (components being food, rent, medicine, education, power, etc.), budget outlay for receipts and expenditures, components of labour force, population etc. Component bar diagrams are usually shaded or coloured suitably.

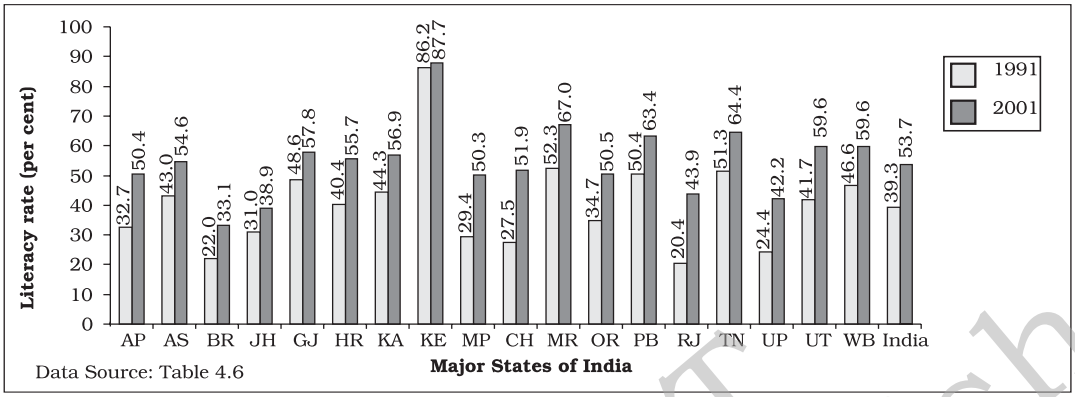


Fig. 4.2: *Multiple bar (column) diagram showing female literacy rates over two census years 1991 and 2001 by major states of India.*

Interpretation: It can be very easily derived from Figure 4.2 that female literacy rate over the years was on increase throughout the country. Similar other interpretations can be made from the figure like the state of Rajasthan experienced the sharpest rise in female literacy, etc.

TABLE 4.7

Enrolment by gender at schools (per cent) of children aged 6–14 years in a district of Bihar

Gender	*Enrolled (per cent)*	*Out of school (per cent)*
Boy	91.5	8.5
Girl	58.6	41.4
All	78.0	22.0

Data Source: *Unpublished data*

A component bar diagram shows the bar and its sub-divisions into two or more components. For example, the bar might show the total population of children in the age-group of 6–14 years. The components show the proportion of those who are enrolled and those who are not. A component bar diagram might also contain different component bars for boys, girls and the total of children in the given age group range, as shown in Figure 4.3. To construct a component bar diagram, first of all, a bar is constructed on the x-axis with its height equivalent to the total value of the bar [for per cent data the bar height is of 100 units (Figure 4.3)]. Otherwise the height is equated to total value of the bar and proportional heights of the components are worked out using unitary method. Smaller components are given priority in parting the bar.

Pie Diagram

A pie diagram is also a component

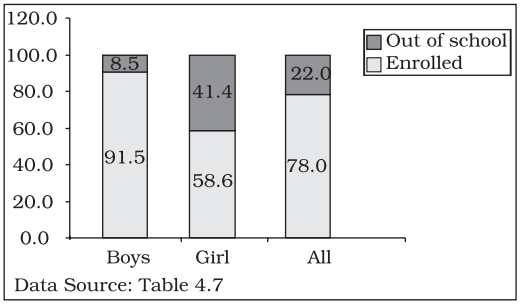


Fig. 4.3: *Enrolment at primary level in a district of Bihar (Component Bar Diagram)*

diagram, but unlike a component bar diagram, a circle whose area is proportionally divided among the components (Fig.4.4) it represents. It

is also called a pie chart. The circle is divided into as many parts as there are components by drawing straight lines from the centre to the circumference.

Pie charts usually are not drawn with absolute values of a category. The values of each category are first expressed as percentage of the total value of all the categories. A circle in a pie chart, irrespective of its value of radius, is thought of having 100 equal parts of 3.6° (360°/100) each. To find out the angle, the component shall subtend at the centre of the circle, each percentage figure of every component is multiplied by 3.6°. An example of this conversion of percentages of components into angular components of the circle is shown in Table 4.8.

It may be interesting to note that data represented by a component bar diagram can also be represented equally well by a pie chart, the only requirement being that absolute values of the components have to be converted into percentages before they can be used for a pie diagram.

TABLE 4.8
Distribution of Indian population by their working status (crore)

Status	*Population*	*Per cent*	*Angular Component*
Marginal Worker	9	8.8	32°
Main Worker	31	30.4	109°
Non-Worker	62	60.8	219°
All	*102*	*100.0*	*360°*

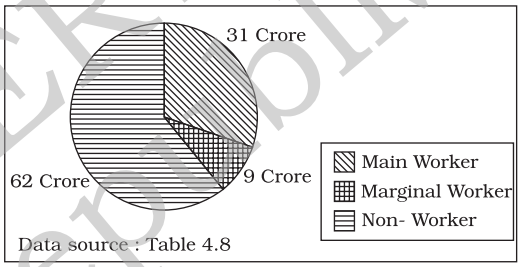


Fig. 4.4: *Pie diagram for different categories of Indian population according to working status 2001.*

Activities

- Represent data presented through Figure 4.4 by a component bar diagram.
- Does the area of a pie have any bearing on total value of the data to be represented by the pie diagram?

Frequency Diagram

Data in the form of grouped frequency distributions are generally represented by frequency diagrams like histogram, frequency polygon, frequency curve and ogive.

Histogram

A histogram is a two dimensional diagram. It is a set of rectangles with bases as the intervals between class boundaries (along X-axis) and with areas proportional to the class frequency (Fig.4.5). If the class intervals are of equal width, which they generally are, the area of the rectangles are proportional to their respective frequencies. However, in some type of data, it is convenient, at times necessary, to use varying width of class intervals. For example, when tabulating deaths by age at death, it would be very meaningful as well as useful too to have very short age intervals (0, 1, 2, ..., yrs/ 0, 7, 28, ..., days) at the beginning when death rates are very high compared to deaths at most other higher age segments of the population. For graphical representation of such data, height for area of a rectangle is the quotient of height (here frequency) and base (here width of the class interval). When intervals are equal, that is, when all rectangles have the same base, area can conveniently be represented by the frequency of any interval for purposes of comparison. When bases vary in their width, the heights of rectangles are to be adjusted to yield comparable measurements. The answer in such a situation is frequency density (class frequency divided by width of the class interval) instead of absolute frequency.

TABLE 4.9
Distribution of daily wage earners in a locality of a town

Daily earning (Rs)	No. of wage earners (f)	Cumulative Frequencey 'Less than'	'More than'
45-49	2	2	85
50-54	3	5	83
55-59	5	10	80
60-64	3	13	75
65-69	6	19	72
70-74	7	26	66
75-79	12	38	59
80-84	13	51	47
85-89	9	60	34
90-94	7	67	25
95-99	6	73	18
100-104	4	77	12
105-109	2	79	8
110-114	3	82	6
115-119	3	85	3

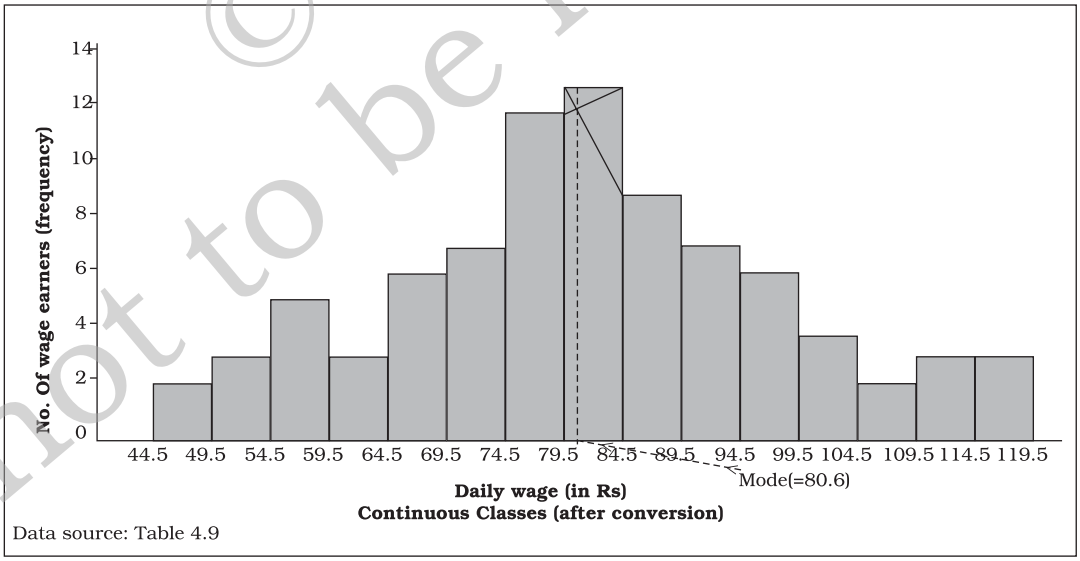
Source: *Unpublished data*

Since histograms are rectangles, a line parallel to the base line and of the same magnitude is to be drawn at a vertical distance equal to frequency (or frequency density) of the class interval. A histogram is never drawn for a discrete variable/data. Since in an interval or ratio scale the lower class boundary of a class interval fuses with the upper class boundary of the previous interval, equal or unequal, the rectangles are all adjacent and there is no open space between two consecutive rectangles. If the classes are not continuous they are first converted into continuous classes as discussed in Chapter 3. Sometimes the common portion between two adjacent rectangles (Fig.4.6) is omitted giving a better impression of continuity. The resulting figure gives the impression of a double staircase.

A histogram looks similar to a bar diagram. But there are more differences than similarities between the two than it may appear at the first impression. The spacing and the width or the area of bars are all arbitrary. It is the height and not the width or the area of the bar that really matters. A single vertical line could have served the same purpose as a bar of same width. Moreover, in histogram no space is left in between two rectangles, but in a bar diagram some space must be left between consecutive bars (except in multiple bar or component bar diagram). Although the bars have the same width, the width of a bar is unimportant for the purpose of comparison. The width in a histogram is as important as its height. *We can have a bar diagram both for discrete and continuous variables, but histogram is drawn only for a continuous variable.* Histogram also gives value of ***mode*** of the frequency distribution graphically as shown in Figure 4.5 and the x-coordinate of the dotted vertical line gives the mode.

Frequency Polygon

A frequency polygon is a plane bounded by straight lines, usually four or more lines. Frequency polygon is an alternative to histogram and is also derived from histogram itself. A frequency polygon can be fitted to a histogram for studying the shape of the curve. The simplest method of drawing a frequency polygon is to join the midpoints of the topside of the consecutive rectangles of the histogram. It leaves us with the two



Data source: Table 4.9

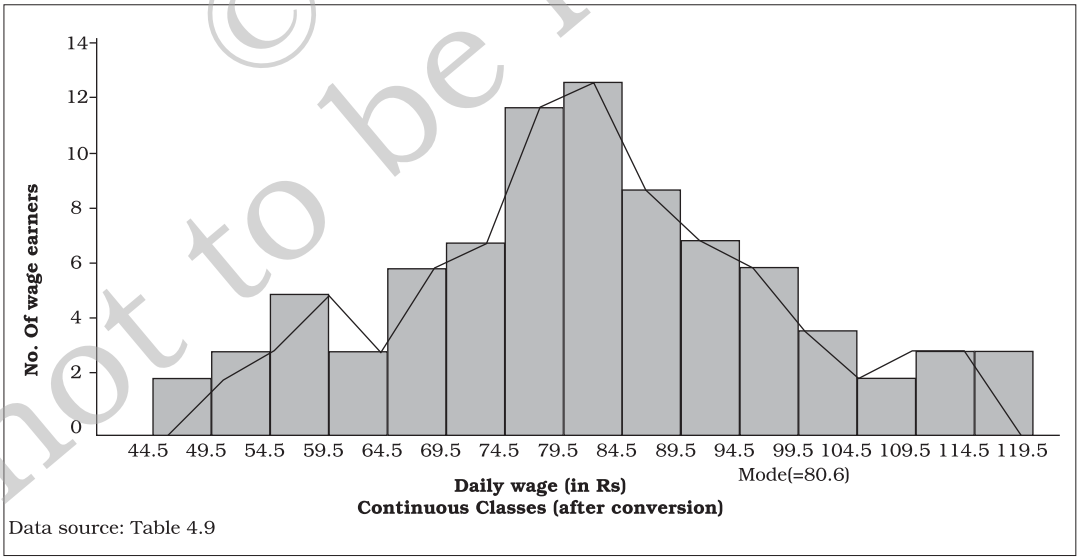
Fig. 4.5: *Histogram for the distribution of 85 daily wage earners in a locality of a town.*

ends away from the base line, denying the calculation of the area under the curve. The solution is to join the two end-points thus obtained to the base line at the mid-values of the two classes with zero frequency immediately at each end of the distribution. Broken lines or dots may join the two ends with the base line. Now the total area under the curve, like the area in the histogram, represents the total frequency or sample size.

Frequency polygon is the most common method of presenting grouped frequency distribution. Both class boundaries and class-marks can be used along the X-axis, the distances between two consecutive class marks being proportional/equal to the width of the class intervals. Plotting of data becomes easier if the class-marks fall on the heavy lines of the graph paper. No matter whether class boundaries or midpoints are used in the X-axis, frequencies (as ordinates) are always plotted against the mid-point of class intervals. When all the points have been plotted in the graph, they are carefully joined by a series of short straight lines. Broken lines join midpoints of two intervals, one in the beginning and the other at the end, with the two ends of the plotted curve (Fig.4.6). When comparing two or more distributions plotted on the same axes, frequency polygon is likely to be more useful since the vertical and horizontal lines of two or more distributions may coincide in a histogram.

Frequency Curve

The frequency curve is obtained by drawing a smooth freehand curve passing through the points of the



Data source: Table 4.9

Fig. 4.6: *Frequency polygon drawn for the data given in Table 4.9*

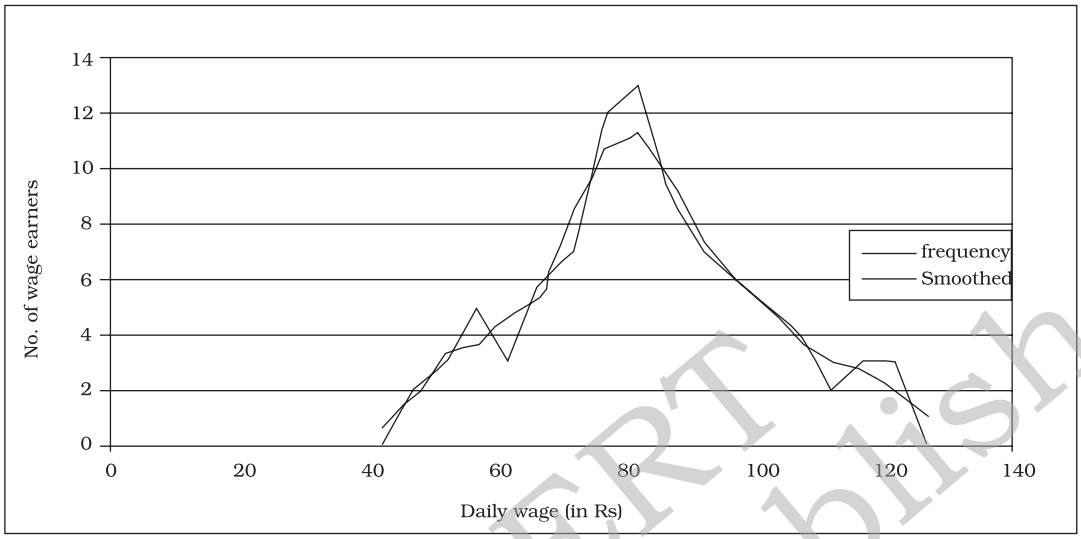


Fig. 4.7: *Frequency curve for Table 4.9*

frequency polygon as closely as possible. It may not necessarily pass through all the points of the frequency polygon but it passes through them as closely as possible (Fig. 4.7).

Ogive

Ogive is also called cumulative frequency curve. As there are two types of cumulative frequencies, for example less than type and more than type, accordingly there are two ogives for any grouped frequency distribution data. Here in place of simple frequencies as in the case of frequency polygon, cumulative frequencies are plotted along y-axis against class limits of the frequency distribution. For less than ogive the cumulative frequencies are plotted against the respective upper limits of the class intervals whereas for more than ogives the cumulative frequencies are plotted against the respective lower limits of the class interval. An interesting feature of the two ogives together is that their intersection point gives the ***median*** Fig. 4.8 (b) of the frequency distribution. As the shapes of the two ogives suggest, less than ogive is never decreasing and more than ogive is never increasing.

TABLE 4.10
Frequency distribution of marks obtained in mathematics

Marks x	*Number of students* f	*'Less than' cumulative frequency*	*'More than' cumulative frequency*
0–20	6	6	64
20–40	5	11	58
40–60	33	44	53
60–80	14	58	20
80–100	6	64	6
Total	*64*		

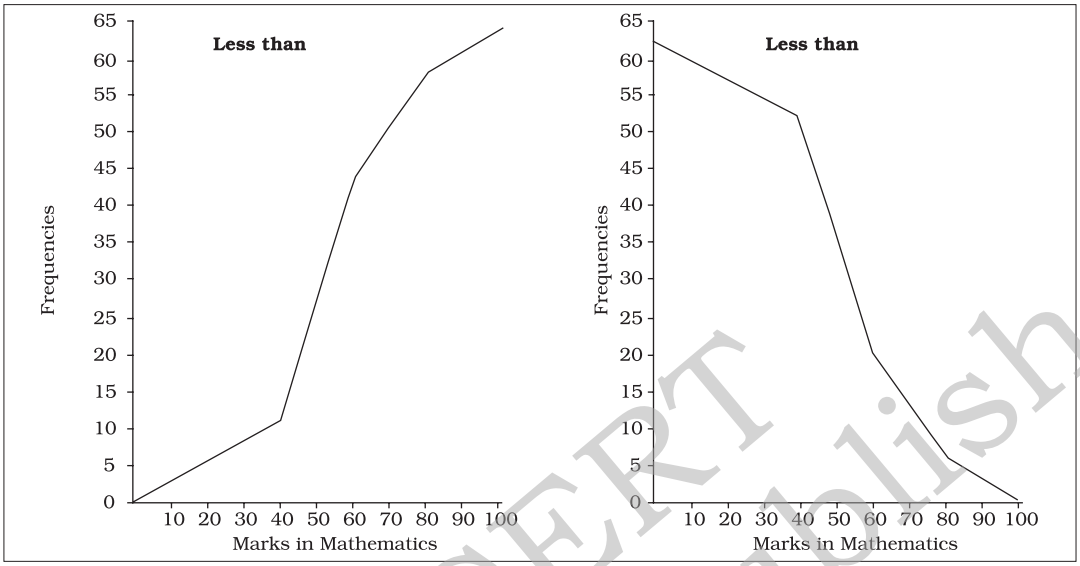


Fig. 4.8(a): *'Less than' and 'More than' ogive for data given in Table 4.10*

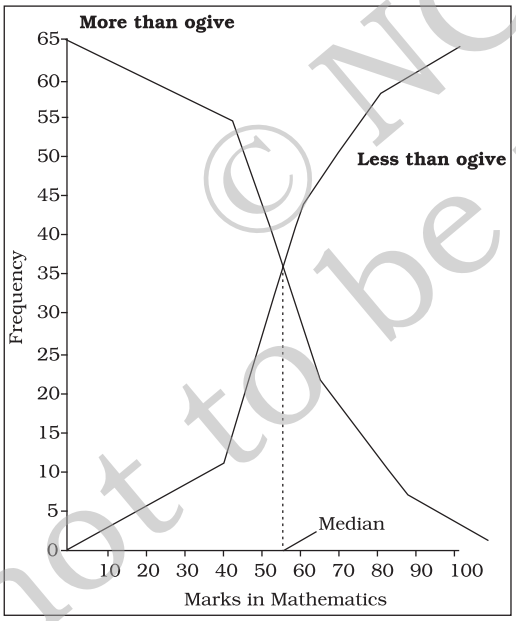


Fig. 4.8(b): *'Less than' and 'More than' ogive for data given in Table 4.10*

Arithmetic Line Graph

An arithmetic line graph is also called time series graph and is a method of diagrammatic presentation of data. In it, time (hour, day/date, week, month, year, etc.) is plotted along x-axis and the value of the variable (time series data) along y-axis. A line graph by joining these plotted points, thus, obtained is called arithmetic line graph (time series graph). It helps in understanding the trend, periodicity, etc. in a long term time series data.

Activity

- Can the ogive be helpful in locating the partition values of the distribution it represents?

TABLE 4.11
Value of Exports and Imports of India
(Rs in 100 crores)

Year	*Exports*	*Imports*
1977–78	54	60
1978–79	57	68
1979–80	64	91
1980–81	67	125
1982–83	88	143
1983–84	98	158
1984–85	117	171
1985–86	109	197
1986–87	125	201
1987–88	157	222
1988–89	202	282
1989–90	277	353
1990–91	326	432
1991–92	440	479
1992–93	532	634
1993–94	698	731
1994–95	827	900
1995–96	1064	1227
1996–97	1186	1369
1997–98	1301	1542
1998–99	1416	1761

Here you can see from Fig. 4.9 that for the period 1978 to 1999, although the imports were more than the exports all through, the rate of acceleration went on increasing after 1988–89 and the gap between the two (imports and exports) was widened after 1995.

6. Conclusion

By now you must have been able to learn how collected data could be presented using various forms of presentation – textual, tabular and diagrammatic. You are now also able to make an appropriate choice of the form of data presentation as well as the type of diagram to be used for a given set of data. Thus you can make presentation of data meaningful, comprehensive and purposeful.

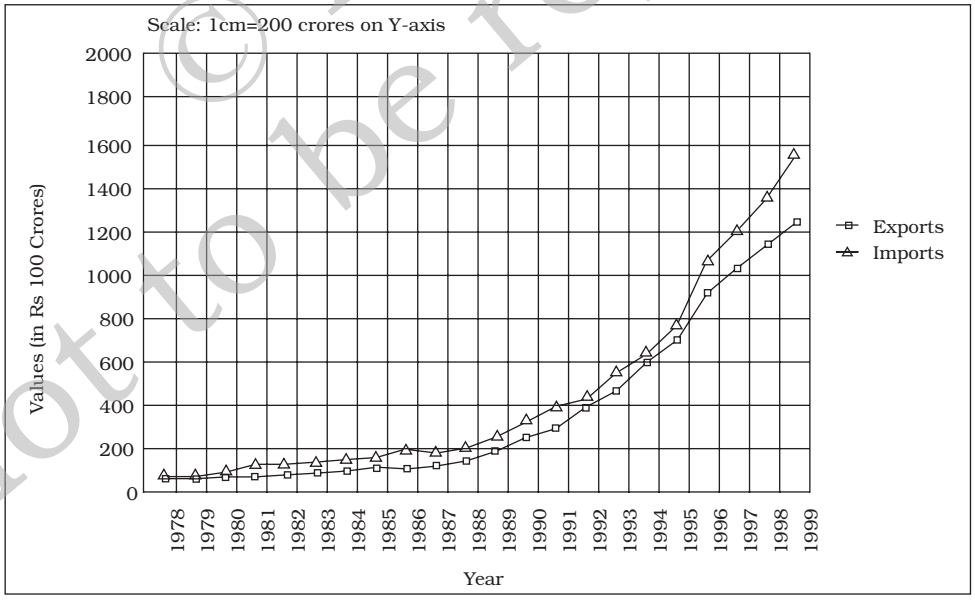


Fig. 4.9: *Arithmetic line graph for time series data given in Table 4.11*

Recap

- Data (even voluminous data) speak meaningfully through presentation.
- For small data (quantity) textual presentation serves the purpose better.
- For large quantity of data tabular presentation helps in accommodating any volume of data for one or more variables.
- Tabulated data can be presented through diagrams which enable quicker comprehension of the facts presented otherwise.

EXERCISES

Answer the following questions, 1 to 10, choosing the correct answer

1. Bar diagram is a
   (i) one-dimensional diagram
   (ii) two-dimensional diagram
   (iii) diagram with no dimension
   (iv) none of the above
2. Data represented through a histogram can help in finding graphically the
   (i) mean
   (ii) mode
   (iii) median
   (iv) all the above
3. Ogives can be helpful in locating graphically the
   (i) mode
   (ii) mean
   (iii) median
   (iv) none of the above
4. Data represented through arithmetic line graph help in understanding
   (i) long term trend
   (ii) cyclicity in data
   (iii) seasonality in data
   (iv) all the above
5. Width of bars in a bar diagram need not be equal (True/False).
6. Width of rectangles in a histogram should essentially be equal (True/False).
7. Histogram can only be formed with continuous classification of data (True/False).

8. Histogram and column diagram are the same method of presentation of data. (True/False).

9. Mode of a frequency distribution can be known graphically with the help of histogram. (True/False).

10. Median of a frequency distribution cannot be known from the ogives. (True/False).

11. What kind of diagrams are more effective in representing the following?
    (i) Monthly rainfall in a year
    (ii) Composition of the population of Delhi by religion
    (iii) Components of cost in a factory

12. Suppose you want to emphasise the increase in the share of urban non-workers and lower level of urbanisation in India as shown in Example 4.2. How would you do it in the tabular form?

13. How does the procedure of drawing a histogram differ when class intervals are unequal in comparison to equal class intervals in a frequency table?

14. The Indian Sugar Mills Association reported that, 'Sugar production during the first fortnight of December 2001 was about 3,87,000 tonnes, as against 3,78,000 tonnes during the same fortnight last year (2000). The off-take of sugar from factories during the first fortnight of December 2001 was 2,83,000 tonnes for internal consumption and 41,000 tonnes for exports as against 1,54,000 tonnes for internal consumption and nil for exports during the same fortnight last season.'
    (i) Present the data in tabular form.
    (ii) Suppose you were to present these data in diagrammatic form which of the diagrams would you use and why?
    (iii) Present these data diagrammatically.

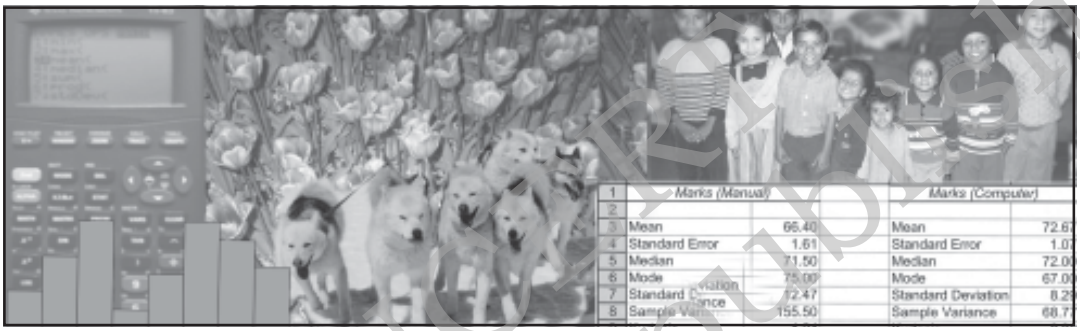
15. The following table shows the estimated sectoral real growth rates (percentage change over the previous year) in GDP at factor cost.

Year (1)	*Agriculture and allied sectors* (2)	*Industry* (3)	*Services* (4)
1994–95	5.0	9.2	7.0
1995–96	–0.9	11.8	10.3
1996–97	9.6	6.0	7.1
1997–98	–1.9	5.9	9.0
1998–99	7.2	4.0	8.3
1999–2000	0.8	6.9	8.2

Represent the data as multiple time series graphs.

CHAPTER 5

Measures of Central Tendency



Studying this chapter should enable you to:

- *understand the need for summarising a set of data by one single number;*
- *recognise and distinguish between the different types of averages;*
- *learn to compute different types of averages;*
- *draw meaningful conclusions from a set of data;*
- *develop an understanding of which type of average would be most useful in a particular situation.*

1. INTRODUCTION

In the previous chapter, you have read the tabular and graphic representation of the data. In this chapter, you will study the measures of central tendency which is a numerical method to explain the data in brief. You can see examples of summarising a large set of data in day to day life like average marks obtained by students of a class in a test, average rainfall in an area, average production in a factory, average income of persons living in a locality or working in a firm etc.

Baiju is a farmer. He grows food grains in his land in a village called Balapur in Buxar district of Bihar. The village consists of 50 small farmers. Baiju has 1 acre of land. You are interested in knowing the economic condition of small farmers of Balapur. You want to compare the economic

condition of Baiju in Balapur village. For this, you may have to evaluate the size of his land holding, by comparing with the size of land holdings of other farmers of Balapur. You may like to see if the land owned by Baiju is –

1. above average in ordinary sense (see the *Arithmetic Mean below)*
2. above the size of what half the farmers own (see the *Median below*)
3. above what most of the farmers own (see the *Mode below*)

In order to evaluate Baiju's relative economic condition, you will have to summarise the whole set of data of land holdings of the farmers of Balapur. This can be done by use of central tendency, which summarises the data in a single value in such a way that this single value can represent the entire data. The measuring of central tendency is a way of summarising the data in the form of a typical or representative value.

There are several statistical measures of central tendency or "averages". The three most commonly used averages are:

- Arithmetic Mean
- Median
- Mode

You should note that there are two more types of averages i.e. Geometric Mean and Harmonic Mean, which are suitable in certain situations. However, the present discussion will be limited to the three types of averages mentioned above.

2. ARITHMETIC MEAN

Suppose the monthly income (in Rs) of six families is given as:
1600, 1500, 1400, 1525, 1625, 1630.

The mean family income is obtained by adding up the incomes and dividing by the number of families.

$$\text{Rs } \frac{1600 + 1500 + 1400 + 1525 + 1625 + 1630}{6}$$

= Rs 1,547

It implies that on an average, a family earns Rs 1,547.

Arithmetic mean is the most commonly used measure of central tendency. It is defined as the sum of the values of all observations divided by the number of observations and is usually denoted by $\bar{X}$. In general, if there are N observations as X_1, X_2, X_3, ..., X_N, then the Arithmetic Mean is given by

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \ldots + X_N}{N}$$

$$= \frac{\Sigma X}{N}$$

Where, ΣX = sum of all observations and N = total number of observations.

How Arithmetic Mean is Calculated

The calculation of arithmetic mean can be studied under two broad categories:

1. *Arithmetic Mean for Ungrouped Data.*
2. *Arithmetic Mean for Grouped Data.*

Arithmetic Mean for Series of Ungrouped Data

Direct Method

Arithmetic mean by *direct method* is the sum of all observations in a series divided by the total number of observations.

Example 1

Calculate Arithmetic Mean from the data showing marks of students in a class in an economics test: 40, 50, 55, 78, 58.

$$\bar{X} = \frac{S X}{N}$$

$$= \frac{40 + 50 + 55 + 78 + 58}{5} = 56.2$$

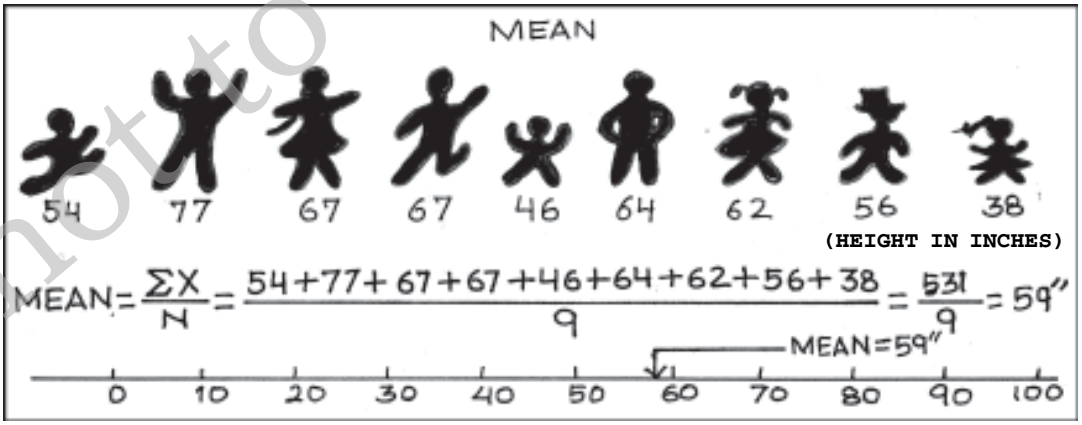
The average marks of students in the economics test are 56.2.

Assumed Mean Method

If the number of observations in the data is more and/or figures are large, it is difficult to compute arithmetic mean by direct method. The computation can be made easier by using assumed mean method.

In order to save time of calculation of mean from a data set containing a large number of observations as well as large numerical figures, you can use assumed mean method. Here you assume a particular figure in the data as the arithmetic mean on the basis of logic/experience. Then you may take deviations of the said assumed mean from each of the observation. You can, then, take the summation of these deviations and divide it by the number of observations in the data. The actual arithmetic mean is estimated by taking the sum of the assumed mean and the ratio of sum of deviations to number of observations. Symbolically,

Let, A = assumed mean
X = individual observations
N = total numbers of observations
d = deviation of assumed mean from individual observation, i.e. d = X – A



Then sum of all deviations is taken as $\Sigma d = \Sigma(X - A)$

Then find $\frac{\Sigma d}{N}$

Then add A and $\frac{\Sigma d}{N}$ to get $\bar{X}$

Therefore, $\bar{X} = A + \frac{\Sigma d}{N}$

You should remember that any value, whether existing in the data or not, can be taken as assumed mean. However, in order to simplify the calculation, centrally located value in the data can be selected as assumed mean.

Example 2

The following data shows the weekly income of 10 families.

Family
A B C D E F G H I J

Weekly Income (in Rs)
850 700 100 750 5000 80 420 2500 400 360

Compute mean family income.

TABLE 5.1
Computation of Arithmetic Mean by Assumed Mean Method

Families	*Income (X)*	*d = X – 850*	*d = (X – 850)/10*
A	850	0	0
B	700	–150	–15
C	100	–750	–75
D	750	–100	–10
E	5000	+4150	+415
F	80	–770	–77
G	420	–430	–43
H	2500	+1650	+165
I	400	–450	–45
J	360	–490	–49
	11160	*+2660*	*+266*

Arithmetic Mean using *assumed mean method*

$$\bar{X} = A + \frac{\Sigma d}{N} = 850 + (2,660)/10$$
$$= \text{Rs}1,116.$$

Thus, the average weekly income of a family by both methods is Rs 1,116. You can check this by using the direct method.

Step Deviation Method

The calculations can be further simplified by dividing all the deviations taken from assumed mean by the common factor 'c'. The objective is to avoid large numerical figures, i.e., if $d = X - A$ is very large, then find d'. This can be done as follows:

$$\frac{d}{c} = \frac{X - A}{C}.$$

The formula is given below:

$$\bar{X} = A + \frac{\Sigma d'}{N} \cdot c$$

Where $d' = (X - A)/c$, c = common factor, N = number of observations, A= Assumed mean.

Thus, you can calculate the arithmetic mean in the example 2, by the step deviation method,

$X = 850 + (266)/10 \times 10$ = Rs 1,116.

Calculation of arithmetic mean for Grouped data

Discrete Series

Direct Method

In case of *discrete series*, frequency against each of the observations is

multiplied by the value of the observation. The values, so obtained, are summed up and divided by the total number of frequencies. Symbolically,

$$\bar{X} = \frac{\Sigma fX}{\Sigma f}$$

Where, ΣfX = sum of product of variables and frequencies.
Σf = sum of frequencies.

Example 3

Calculate mean farm size of cultivating households in a village for the following data.

Farm Size (in acres):

64 63 62 61 60 59

No. of Cultivating Households:

8 18 12 9 7 6

TABLE 5.2
Computation of Arithmetic Mean by Direct Method

Farm Size (X) in acres (1)	*No. of cultivating households (f)* (2)	*X* (1 × 2) (3)	*d* (X – 62) (4)	*fd* (2 × 4) (5)
64	8	512	+2	+16
63	18	1134	+1	+18
62	12	744	0	0
61	9	549	–1	–9
60	7	420	–2	–14
59	6	354	–3	–18
	60	*3713*	*–3*	*–7*

Arithmetic mean using *direct method*,

$$\bar{X} = \frac{\Sigma fX}{\Sigma f} = \frac{3717}{60} = 61.88 \text{ acres}$$

Therefore, the mean farm size in a village is 61.88 acres.

Assumed Mean Method

As in case of individual series the calculations can be simplified by using assumed mean method, as described earlier, with a simple modification. Since frequency (f) of each item is given here, we multiply each deviation (d) by the frequency to get fd. Then we get Σfd. The next step is to get the total of all frequencies i.e. Σf. Then find out $\Sigma fd / \Sigma f$. Finally the arithmetic mean is calculated by $\bar{X} = A + \frac{\Sigma fd}{\Sigma f}$ using assumed mean method.

Step Deviation Method

In this case the deviations are divided by the common factor 'c' which simplifies the calculation. Here we estimate $d' = \frac{d}{c} = \frac{X - A}{C}$ in order to reduce the size of numerical figures for easier calculation. Then get fd′ and $\Sigma fd'$. Finally the formula for step deviation method is given as,

$$\bar{X} = A + \frac{\Sigma fd'}{\Sigma f} \cdot c$$

Activity

- Find the mean farm size for the data given in example 3, by using *step deviation* and *assumed mean* methods.

Continuous Series

Here, class intervals are given. The process of calculating arithmetic mean

in case of continuous series is same as that of a discrete series. The only difference is that the mid-points of various class intervals are taken. You should note that class intervals may be exclusive or inclusive or of unequal size. Example of exclusive class interval is, say, 0–10, 10–20 and so on. Example of inclusive class interval is, say, 0–9, 10–19 and so on. Example of unequal class interval is, say, 0–20, 20–50 and so on. In all these cases, calculation of arithmetic mean is done in a similar way.

Example 4

Calculate average marks of the following students using (a) Direct method (b) Step deviation method.

Direct Method

Marks

0–10 10–20 20–30 30–40 40–50 50–60 60–70

No. of Students

5 12 15 25 8 3 2

TABLE 5.3

Computation of Average Marks for Exclusive Class Interval by Direct Method

Mark (x)	*No. of students (f)*	*mid value (m)*	*fm (2)×(3)*	$d'=\frac{(m-35)}{10}$	*fd'*
(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
0–10	5	5	25	–3	–15
10–20	12	15	180	–2	–24
20–30	15	25	375	–1	–15
30–40	25	35	875	0	0
40–50	8	45	360	1	8
50–60	3	55	165	2	6
60–70	2	65	130	3	6
	70		*2110*		*–34*

Steps:

1. Obtain mid values for each class denoted by m.
2. Obtain Σfm and apply the direct method formula:

$$\bar{X} = \frac{\Sigma fm}{\Sigma f} = \frac{2110}{70} = 30.14 \text{ marks}$$

Step deviation method

1. Obtain $d' = \frac{m - A}{c}$
2. Take A = 35, (any arbitrary figure), c = common factor.

$$\bar{X} = A + \frac{\Sigma fd'}{\Sigma f} \times c = 35 + \frac{(-34)}{70} \times 10$$
$$= 30.14 \text{ marks}$$

An interesting property of A.M.

It is interesting to know and useful for checking your calculation that the sum of deviations of items about arithmetic mean is always equal to zero. Symbolically, $\Sigma (X - \bar{X}) = 0$.

However, arithmetic mean is affected by extreme values. Any large value, on either end, can push it up or down.

Weighted Arithmetic Mean

Sometimes it is important to assign weights to various items according to their importance, when you calculate the arithmetic mean. For example, there are two commodities, mangoes and potatoes. You are interested in finding the average price of mangoes (p_1) and potatoes (p_2). The arithmetic

mean will be $\frac{p_1 + p_2}{2}$. However, you might want to give more importance to the rise in price of potatoes (p_2). To do this, you may use as 'weights' the quantity of mangoes (q_1) and the quantity of potatoes (q_2). Now the arithmetic mean weighted by the quantities would be $\frac{q_1p_1 + q_2p_2}{q_1 + q_2}$.

In general the weighted arithmetic mean is given by,

$$\frac{w_1x_1 + w_2x_2 + \ldots + w_nx_n}{w_1 + w_2 + \ldots + w_n} = \frac{\sum wx}{\sum w}$$

When the prices rise, you may be interested in the rise in the price of the commodities that are more important to you. You will read more about it in the discussion of Index Numbers in Chapter 8.

Activities

- Check this property of the arithmetic mean for the following example:

  X: 4 6 8 10 12
- In the above example if mean is increased by 2, then what happens to the individual observations, if all are equally affected.
- If first three items increase by 2, then what should be the values of the last two items, so that mean remains the same.
- Replace the value 12 by 96. What happens to the arithmetic mean. Comment.

3. MEDIAN

The arithmetic mean is affected by the presence of extreme values in the data. If you take a measure of central tendency which is based on middle position of the data, it is not affected by extreme items. Median is that positional value of the variable which divides the distribution into two equal parts, one part comprises all values greater than or equal to the median value and the other comprises all values less than or equal to it. *The Median is the "middle" element when the data set is arranged in order of the magnitude.*

Computation of median

The median can be easily computed by sorting the data from smallest to largest and counting the middle value.

Example 5

Suppose we have the following observation in a data set: 5, 7, 6, 1, 8, 10, 12, 4, and 3.
Arranging the data, in ascending order you have:

1, 3, 4, 5, 6, 7, 8, 10, 12.

The *"middle score"* is 6, so the median is 6. Half of the scores are larger than 6 and half of the scores are smaller.

If there are even numbers in the data, there will be two observations which fall in the middle. The median in this case is computed as the

arithmetic mean of the two middle values.

Example 6

The following data provides marks of 20 students. You are required to calculate the median marks.
25, 72, 28, 65, 29, 60, 30, 54, 32, 53, 33, 52, 35, 51, 42, 48, 45, 47, 46, 33.

Arranging the data in an ascending order, you get

25, 28, 29, 30, 32, 33, 33, 35, 42, 45, 46, 47, 48, 51, 52, 53, 54, 60,

65, 72.

You can see that there are two observations in the middle, namely 45 and 46. The median can be obtained by taking the mean of the two observations:

$$\text{Median} = \frac{45+46}{2} = 45.5 \text{ marks}$$

In order to calculate median it is important to know the position of the median i.e. item/items at which the median lies. The position of the median can be calculated by the following formula:

$$\text{Position of median} = \frac{(N+1)^{th}}{2} \text{ item}$$

Where N = number of items.

You may note that the above formula gives you the position of the median in an ordered array, not the median itself. Median is computed by the formula:

$$\textit{Median} = \text{size of } \frac{(N+1)^{th}}{2} \text{ item}$$

Discrete Series

In case of discrete series the position of median i.e. (N+1)/2th item can be located through cumulative frequency. The corresponding value at this position is the value of median.

Example 7

The frequency distribution of the number of persons and their respective incomes (in Rs) are given below. Calculate the median income.

Income (in Rs):	10	20	30	40
Number of persons:	2	4	10	4

In order to calculate the median income, you may prepare the frequency distribution as given below.

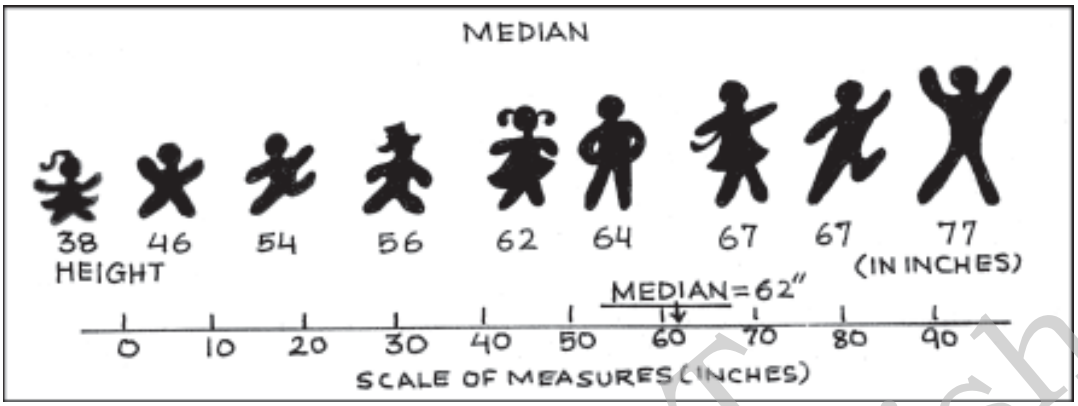
TABLE 5.4
Computation of Median for Discrete Series

Income (in Rs)	*No of persons (f)*	*Cumulative frequency (cf)*
10	2	2
20	4	6
30	10	16
40	4	20

The median is located in the (N+1)/2 = (20+1)/2 = 10.5th observation. This can be easily located through cumulative frequency. The 10.5th observation lies in the c.f. of 16. The income corresponding to this is Rs 30, so the median income is Rs 30.

Continuous Series

In case of continuous series you have to locate the median class where



N/2th item [not (N+1)/2th item] lies. The median can then be obtained as follows:

$$\text{Median} = L + \frac{(N/2 - \text{c.f.})}{f} \times h$$

Where, L = lower limit of the median class,
c.f. = cumulative frequency of the class preceding the median class,
f = frequency of the median class,
h = magnitude of the median class interval.

No adjustment is required if frequency is of unequal size or magnitude.

Example 8

Following data relates to daily wages of persons working in a factory. Compute the median daily wage.

Daily wages (in Rs):
55-60 50-55 45-50 40-45 35-40 30-35 25-30 20-25
Number of workers:
7 13 15 20 30 33 28 14

The data is arranged in ascending order here.

In the above illustration median class is the value of (N/2)th item (i.e.160/2) = 80th item of the series, which lies in 35-40 class interval. Applying the formula of the median as:

TABLE 5.5
Computation of Median for Continuous Series

Daily wages (in Rs)	*No. of Workers (f)*	*Cumulative Frequency*
20-25	14	14
25-30	28	42
30-35	33	75
35-40	30	105
40-45	20	125
45-50	15	140
50-55	13	153
55-60	7	160

$$\text{Median} = L + \frac{(N/2 - \text{c.f.})}{f} \times h$$

$$= \frac{35 + (80 - 75)}{30} \times (40 - 35)$$

$$= \text{Rs } 35.83$$

Thus, the median daily wage is Rs 35.83. This means that 50% of the

workers are getting less than or equal to Rs 35.83 and 50% of the workers are getting more than or equal to this wage.

You should remember that median, as a measure of central tendency, is not sensitive to all the values in the series. It concentrates on the values of the central items of the data.

Activities

- Find mean and median for all four values of the series. What do you observe?

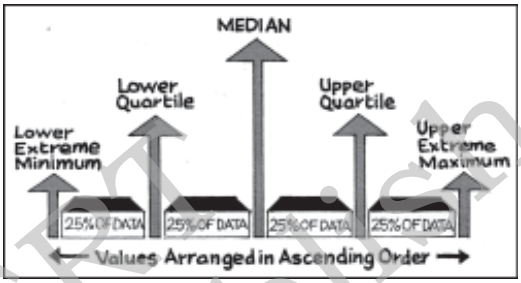
TABLE 5.6
Mean and Median of different series

Series	*X (Variable Values)*	*Mean*	*Median*
A	1, 2, 3	?	?
B	1, 2, 30	?	?
C	1, 2, 300	?	?
D	1, 2, 3000	?	?

- Is median affected by extreme values? What are outliers?
- Is median a better method than mean?

Quartiles

Quartiles are the measures which divide the data into four equal parts, each portion contains equal number of observations. Thus, there are three quartiles. The first Quartile (denoted by Q_1) or lower quartile has 25% of the items of the distribution below it and 75% of the items are greater than it. The second Quartile (denoted by Q_2) or median has 50% of items below it and 50% of the observations above it. The third Quartile (denoted by Q_3) or upper Quartile has 75% of the items of the distribution below it and 25% of the items above it. Thus, Q_1 and Q_3 denote the two limits within which central 50% of the data lies.



Percentiles

Percentiles divide the distribution into hundred equal parts, so you can get 99 dividing positions denoted by P_1, P_2, P_3, ..., P_{99}. P_{50} is the median value. If you have secured 82 percentile in a management entrance examination, it means that your position is below 18 percent of total candidates appeared in the examination. If a total of one lakh students appeared, where do you stand?

Calculation of Quartiles

The method for locating the Quartile is same as that of the median in case of individual and discrete series. The value of Q_1 and Q_3 of an ordered series can be obtained by the following formula where N is the number of observations.

$$Q_1 = \text{size of } \frac{(N+1)\text{th}}{4} \text{ item}$$

Q_3 = size of $\frac{3(N+1)th}{4}$ item.

Example 9

Calculate the value of *lower quartile* from the data of the marks obtained by ten students in an examination.
22, 26, 14, 30, 18, 11, 35, 41, 12, 32.

Arranging the data in an ascending order,
11, 12, 14, 18, 22, 26, 30, 32, 35, 41.

Q_1 = size of $\frac{(N+1)th}{4}$ item = size of $\frac{(10+1)th}{4}$ item = size of 2.75th item
= 2nd item + .75 (3rd item – 2nd item)
= 12 + .75(14 –12) = 13.5 marks.

Activity

- Find out Q_3 yourself.

5. MODE

Sometimes, you may be interested in knowing the most typical value of a series or the value around which maximum concentration of items occurs. For example, a manufacturer would like to know the size of shoes that has maximum demand or style of the shirt that is more frequently demanded. Here, *Mode* is the most appropriate measure. The word *mode* has been derived from the French word "la Mode" which signifies the most fashionable values of a distribution, because it is repeated the highest number of times in the series.

Mode is the most frequently observed data value. It is denoted by M_o.

Computation of Mode

Discrete Series

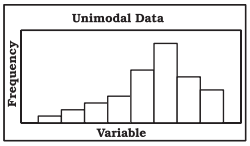
Consider the data set 1, 2, 3, 4, 4, 5. The mode for this data is 4 because 4 occurs most frequently (twice) in the data.

Example 10

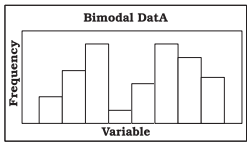
Look at the following discrete series:

Variable	10	20	30	40	50
Frequency	2	8	20	10	5

Here, as you can see the maximum frequency is 20, the value of mode is 30. In this case, as there is a unique value of mode, the data is *unimodal*. But, the mode is not necessarily unique, unlike arithmetic mean and median. You can have data with two modes (bi-modal) or more than two modes (multi-modal). It may be possible that there may be no mode if no value appears more frequent than any other value in the distribution. For example, in a series 1, 1, 2, 2, 3, 3, 4, 4, there is no mode.



Unimodal Data



Bimodal Data

Continuous Series

In case of continuous frequency distribution, modal class is the class with largest frequency. Mode can be calculated by using the formula:

$$M_O = L + \frac{D_1}{D_1 + D_2} \times h$$

Where L = lower limit of the modal class

D_1 = difference between the frequency of the modal class and the frequency of the class preceding the modal class (ignoring signs).

D_2 = difference between the frequency of the modal class and the frequency of the class succeeding the modal class (ignoring signs).

h = class interval of the distribution.

You may note that in case of continuous series, class intervals should be equal and series should be exclusive to calculate the mode. If mid points are given, class intervals are to be obtained.

Example 11

Calculate the value of modal worker family's monthly income from the following data:

Income per month (in '000 Rs)			
Below 50	Below 45	Below 40	Below 35
Below 30	Below 25	Below 20	Below 15
Number of families			
97	95	90	80
60	30	12	4

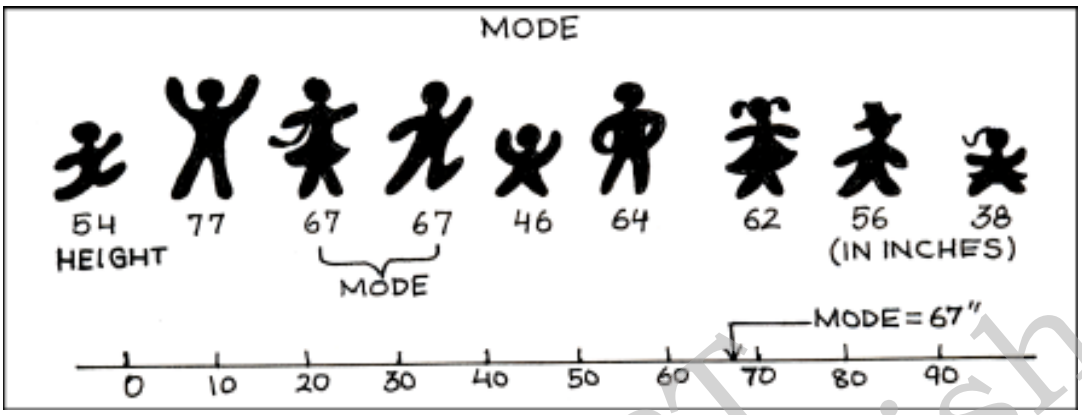
As you can see this is a case of cumulative frequency distribution. In order to calculate mode, you will have to covert it into an exclusive series. In

TABLE 5.7
Grouping Table

Income (in '000 Rs)	*Group Frequency*					
	I	*II*	*III*	*IV*	*V*	*VI*
45–50	97 – 95 = 2					
40–45	95 – 90 = 5	7		17		
35–40	90 – 80 = 10		15			
30–35	80 – 60 = 20	30			35	
25–30	60 – 30 = 30		50			60
20–25	30 – 12 = 18	48		68		
15–20	12 – 4 = 8		26		56	
10–15	4	12				30

TABLE 5.8
Analysis Table

Columns	*Class Intervals*							
	45–50	*40–45*	*35–40*	*30–35*	*25–30*	*20–25*	*15–20*	*10–15*
I					×			
II					×	×		
III				×	×			
IV				×	×	×		
V					×	×	×	
VI			×	×	×			
Total	–	–	*1*	*3*	*6*	*3*	*1*	–



this example, the series is in the descending order. Grouping and Analysis table would be made to determine the modal class.

The value of the mode lies in 25–30 class interval. By inspection also, it can be seen that this is a modal class.

Now L = 25, D_1 = (30 – 18) = 12, D_2 = (30 – 20) = 10, h = 5

Using the formula, you can obtain the value of the mode as:

M_O (in '000 Rs)

$$M = \frac{D_1}{D_1 + D_2} \times h$$

$$= 25 + \frac{12}{10+12} \times 5 = \text{Rs}\,27{,}273$$

Thus the *modal* worker family's monthly income is Rs 27,273.

Activities

- A shoe company, making shoes for adults only, wants to know the most popular size of shoes. Which average will be most appropriate for it?
- Take a small survey in your class to know the student's preference for Chinese food using appropriate measure of central tendency.
- Can mode be located graphically?

6. Relative Position of Arithmetic Mean, Median and Mode

Suppose we express,

Arithmetic Mean = M_e

Median = M_i

Mode = M_o

so that e, i and o are the suffixes. The relative magnitude of the three are $M_e > M_i > M_o$ or $M_e < M_i < M_o$ (suffixes occurring in alphabetical order). *The median is always between the arithmetic mean and the mode.*

7. Conclusion

Measures of central tendency or averages are used to summarise the data. It specifies a single most representative value to describe the data set. Arithmetic mean is the most commonly used average. It is simple

to calculate and is based on all the observations. But it is unduly affected by the presence of extreme items. Median is a better summary for such data. Mode is generally used to describe the qualitative data. Median and mode can be easily computed graphically. In case of open-ended distribution they can also be easily computed. Thus, it is important to select an appropriate average depending upon the purpose of analysis and the nature of the distribution.

Recap

- The measure of central tendency summarises the data with a single value, which can represent the entire data.
- Arithmetic mean is defined as the sum of the values of all observations divided by the number of observations.
- The sum of deviations of items from the arithmetic mean is always equal to zero.
- Sometimes, it is important to assign weights to various items according to their importance.
- Median is the central value of the distribution in the sense that the number of values less than the median is equal to the number greater than the median.
- Quartiles divide the total set of values into four equal parts.
- Mode is the value which occurs most frequently.

EXERCISES

1. Which average would be suitable in the following cases?
   (i) Average size of readymade garments.
   (ii) Average intelligence of students in a class.
   (iii) Average production in a factory per shift.
   (iv) Average wages in an industrial concern.
   (v) When the sum of absolute deviations from average is least.
   (vi) When quantities of the variable are in ratios.
   (vii) In case of open-ended frequency distribution.
2. Indicate the most appropriate alternative from the multiple choices provided against each question.
   (i) The most suitable average for qualitative measurement is
   (a) arithmetic mean
   (b) median
   (c) mode

(d) geometric mean
(e) none of the above

(ii) Which average is affected most by the presence of extreme items?
(a) median
(b) mode
(c) arithmetic mean
(d) geometric mean
(e) harmonic mean

(iii) The algebraic sum of deviation of a set of n values from A.M. is
(a) n
(b) 0
(c) 1
(d) none of the above

[Ans. (i) b (ii) c (iii) b]

3. Comment whether the following statements are true or false.
(i) The sum of deviation of items from median is zero.
(ii) An average alone is not enough to compare series.
(iii) Arithmetic mean is a positional value.
(iv) Upper quartile is the lowest value of top 25% of items.
(v) Median is unduly affected by extreme observations.
[Ans. (i) False (ii) True (iii) False (iv) True (v) False]

4. If the arithmetic mean of the data given below is 28, find (a) the missing frequency, and (b) the median of the series:

Profit per retail shop (in Rs)	0-10	10-20	20-30	30-40	40-50	50-60
Number of retail shops	12	18	27	-	17	6

(Ans. The value of missing frequency is 20 and value of the median is Rs 27.41)

5. The following table gives the daily income of ten workers in a factory. Find the arithmetic mean.

Workers	A	B	C	D	E	F	G	H	I	J
Daily Income (in Rs)	120	150	180	200	250	300	220	350	370	260

(Ans. Rs 240)

6. Following information pertains to the daily income of 150 families. Calculate the arithmetic mean.

Income (in Rs)	*Number of families*
More than 75	150
,, 85	140
,, 95	115
,, 105	95
,, 115	70
,, 125	60
,, 135	40
,, 145	25

(Ans. Rs 116.3)

7. The size of land holdings of 380 families in a village is given below. Find the median size of land holdings.

Size of Land Holdings (in acres)	Less than 100	100–200	200 – 300	300–400	400 and above. –
Number of families	40	89	148	64	39

(Ans. 241.22 acres)

8. The following series relates to the daily income of workers employed in a firm. Compute (a) highest income of lowest 50% workers (b) minimum income earned by the top 25% workers and (c) maximum income earned by lowest 25% workers.

Daily Income (in Rs)	10–14	15–19	20–24	25–29	30–34	35–39
Number of workers	5	10	15	20	10	5

(Hint: compute median, lower quartile and upper quartile.)
[Ans. (a) Rs 25.11 (b) Rs 19.92 (c) Rs 29.19]

9. The following table gives production yield in kg. per hectare of wheat of 150 farms in a village. Calculate the mean, median and mode production yield.

Production yield (kg. per hectare)	50–53	53–56	56–59	59–62	62–65	65–68	68–71	71–74	74–77
Number of farms	3	8	14	30	36	28	16	10	5

(Ans. mean = 63.82 kg. per hectare, median = 63.67 kg. per hectare, mode = 63.29 kg. per hectare)

CHAPTER 6

Measures of Dispersion

Studying this chapter should enable you to:

- *know the limitations of averages;*
- *appreciate the need of measures of dispersion;*
- *enumerate various measures of dispersion;*
- *calculate the measures and compare them;*
- *distinguish between absolute and relative measures.*

1. INTRODUCTION

In the previous chapter, you have studied how to sum up the data into a single representative value. However, that value does not reveal the variability present in the data. In this chapter you will study those measures, which seek to quantify variability of the data.

Three friends, Ram, Rahim and Maria are chatting over a cup of tea. During the course of their conversation, they start talking about their family incomes. Ram tells them that there are four members in his family and the average income per member is Rs 15,000. Rahim says that the average income is the same in his family, though the number of members is six. Maria says that there are five members in her family, out of which one is not working. She calculates that the average income in her family too, is Rs 15,000. They are a little surprised since they know that Maria's father is earning a huge salary. They go into details and gather the following data:

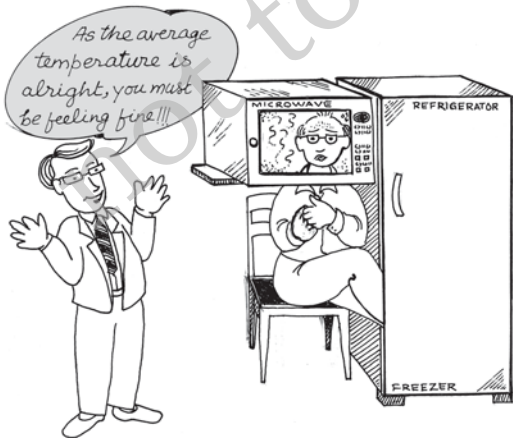
Family Incomes

Sl. No.	*Ram*	*Rahim*	*Maria*
1.	12,000	7,000	0
2.	14,000	10,000	7,000
3.	16,000	14,000	8,000
4.	18,000	17,000	10,000
5.	-----	20,000	50,000
6.	-----	22,000	------
Total income	*60,000*	*90,000*	*75,000*
Average income	*15,000*	*15,000*	*15,000*

Do you notice that although the average is the same, there are considerable differences in individual incomes?

It is quite obvious that averages try to tell only one aspect of a distribution i.e. a representative size of the values. To understand it better, you need to know the spread of values also.

You can see that in Ram's family., differences in incomes are comparatively lower. In Rahim's family, differences are higher and in Maria's family are the highest. Knowledge of only average is insufficient. If you have another value which reflects the quantum of variation in values, your understanding of a distribution improves considerably. For example, per capita income gives only the average income. A measure of dispersion can tell you about income inequalities, thereby improving the understanding of the relative standards of living enjoyed by different strata of society.



Dispersion is the extent to which values in a distribution differ from the average of the distribution.

To quantify the extent of the variation, there are certain measures namely:

(i) Range
(ii) Quartile Deviation
(iii) Mean Deviation
(iv) Standard Deviation

Apart from these measures which give a numerical value, there is a graphic method for estimating dispersion.

Range and Quartile Deviation measure the dispersion by calculating the spread within which the values lie. Mean Deviation and Standard Deviation calculate the extent to which the values differ from the average.

2. Measures Based upon Spread of Values

Range

Range (R) is the difference between the largest (L) and the smallest value (S) in a distribution. Thus,

$R = L - S$

Higher value of Range implies higher dispersion and vice-versa.

Activities

Look at the following values:
20, 30, 40, 50, 200

- Calculate the Range.
- What is the Range if the value 200 is not present in the data set?
- If 50 is replaced by 150, what will be the Range?

Range: Comments

Range is unduly affected by extreme values. It is not based on all the values. As long as the minimum and maximum values remain unaltered, any change in other values does not affect range. It can not be calculated for open-ended frequency distribution.

Notwithstanding some limitations, Range is understood and used frequently because of its simplicity. For example, we see the maximum and minimum temperatures of different cities almost daily on our TV screens and form judgments about the temperature variations in them.

Open-ended distributions are those in which either the lower limit of the lowest class or the upper limit of the highest class or both are not specified.

Activity

- Collect data about 52-week high/low of 10 shares from a newspaper. Calculate the range of share prices. Which stock is most volatile and which is the most stable?

Quartile Deviation

The presence of even one extremely high or low value in a distribution can reduce the utility of range as a measure of dispersion. Thus, you may need a measure which is not unduly affected by the outliers.

In such a situation, if the entire data is divided into four equal parts, each containing 25% of the values, we get the values of Quartiles and Median. (You have already read about these in Chapter 5).

The upper and lower quartiles (Q_3 and Q_1, respectively) are used to calculate Inter Quartile Range which is $Q_3 - Q_1$.

Inter-Quartile Range is based upon middle 50% of the values in a distribution and is, therefore, not affected by extreme values. Half of the Inter-Quartile Range is called Quartile Deviation. Thus:

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

Q.D. is therefore also called *Semi-Inter Quartile Range.*

Calculation of Range and Q.D. for ungrouped data

Example 1

Calculate Range and Q.D. of the following observations:

20, 25, 29, 30, 35, 39, 41, 48, 51, 60 and 70

Range is clearly 70 – 20 = 50

For Q.D., we need to calculate values of Q_3 and Q_1.

Q_1 is the size of $\frac{n+1}{4}$ th value.

n being 11, Q_1 is the size of 3rd value.

As the values are already arranged in ascending order, it can be seen that Q_1, the 3rd value is 29. [What will you do if these values are not in an order?]

Similarly, Q_3 is size of $\frac{3(n+1)}{4}$ th value; i.e. 9th value which is 51. Hence $Q_3 = 51$

$$Q.D. = \frac{Q_3 - Q_1}{2} = \frac{51 - 29}{2} = 11$$

Do you notice that Q.D. is the average difference of the Quartiles from the median.

Activity

- Calculate the median and check whether the above statement is correct.

Calculation of Range and Q.D. for a frequency distribution.

Example 2

For the following distribution of marks scored by a class of 40 students, calculate the Range and Q.D.

TABLE 6.1

Class intervals C I	*No. of students* (f)
0–10	5
10–20	8
20–40	16
40–60	7
60–90	4
	40

Range is just the difference between the upper limit of the highest class and the lower limit of the lowest class. So Range is 90 – 0 = 90. For Q.D., first calculate cumulative frequencies as follows:

Class-Intervals *CI*	*Frequencies* *f*	*Cumulative Frequencies* *c. f.*
0–10	5	05
10–20	8	13
20–40	16	29
40–60	7	36
60–90	4	40
	n = 40	

Q_1 is the size of $\frac{n}{4}$ th value in a continuous series. Thus it is the size of the 10th value. The class containing the 10th value is 10–20. Hence Q_1 lies in class 10–20. Now, to calculate the exact value of Q_1, the following formula is used:

$$Q_1 = L + \frac{\frac{n}{4} - cf}{f} \cdot i$$

Where L = 10 (lower limit of the relevant Quartile class)

c.f. = 5 (Value of c.f. for the class preceding the Quartile class)

i = 10 (interval of the Quartile class), and

f = 8 (frequency of the Quartile class) Thus,

$$Q_1 = 10 + \frac{10 - 5}{8} \cdot 10 = 16.25$$

Similarly, Q_3 is the size of $\frac{3n}{4}$ th

value; i.e., 30th value, which lies in class 40–60. Now using the formula for Q_3, its value can be calculated as follows:

$$Q_3 = L + \frac{\frac{3n}{4} - \text{c.f.}}{f} \quad i$$

$$Q_3 = 40 + \frac{30 - 29}{7} \quad 20$$

$$Q_3 = 42.87$$

$$Q.D. = \frac{42.87 - 16.25}{2} = 13.31$$

In individual and discrete series, Q_1 is the size of $\frac{n+1\text{th}}{4}$ value, but in a continuous distribution, it is the size of $\frac{n\ \text{th}}{4}$ value. Similarly, for Q_3 and median also, n is used in place of n+1.

If the entire group is divided into two equal halves and the median calculated for each half, you will have the median of better students and the median of weak students. These medians differ from the median of the entire group by 13.31 on an average. Similarly, suppose you have data about incomes of people of a town. Median income of all people can be calculated. Now if all people are divided into two equal groups of rich and poor, medians of both groups can be calculated. Quartile Deviation will tell you the average difference between medians of these two groups belonging to rich and poor, from the median of the entire group.

Quartile Deviation can generally be calculated for open-ended distributions and is not unduly affected by extreme values.

3. Measures of Dispersion from Average

Recall that dispersion was defined as the extent to which values differ from their average. Range and Quartile Deviation do not attempt to calculate, how far the values are, from their average. Yet, by calculating the spread of values, they do give a good idea about the dispersion. Two measures which are based upon deviation of the values from their average are Mean Deviation and Standard Deviation.

Since the average is a central value, some deviations are positive and some are negative. If these are added as they are, the sum will not reveal anything. In fact, the sum of deviations from Arithmetic Mean is always zero. Look at the following two sets of values.

Set A : 5, 9, 16
Set B : 1, 9, 20

You can see that values in Set B are farther from the average and hence more dispersed than values in Set A. Calculate the deviations from Arithmetic Mean amd sum them up. What do you notice? Repeat the same with Median. Can you comment upon the quantum of variation from the calculated values?

Mean Deviation tries to overcome this problem by ignoring the signs of deviations, i.e., it considers all deviations positive. For standard deviation, the deviations are first squared and averaged and then square root of the average is found. We shall now discuss them separately in detail.

Mean Deviation

Suppose a college is proposed for students of five towns A, B, C, D and E which lie in that order along a road. Distances of towns in kilometres from town A and number of students in these towns are given below:

Town	*Distance from town A*	*No. of Students*
A	0	90
B	2	150
C	6	100
D	14	200
E	18	80
		620

Now, if the college is situated in town A, 150 students from town B will have to travel 2 kilometers each (a total of 300 kilometres) to reach the college. The objective is to find a location so that the average distance travelled by students is minimum.

You may observe that the students will have to travel more, on an average, if the college is situated at town A or E. If on the other hand, it is somewhere in the middle, they are likely to travel less. The average distance travelled is calculated by Mean Deviation which is simply the arithmetic mean of the differences of the values from their average. The average used is either the arithmetic mean or median.

(Since the mode is not a stable average, it is not used to calculate Mean Deviation.)

Activities

- Calculate the total distance to be travelled by students if the college is situated at town A, at town C, or town E and also if it is exactly half way between A and E.
- Decide where, in you opinion, the college should be established, if there is only one student in each town. Does it change your answer?

Calculation of Mean Deviation from Arithmetic Mean for ungrouped data.

Direct Method

Steps:

(i) The A.M. of the values is calculated
(ii) Difference between each value and the A.M. is calculated. All differences are considered positive. These are denoted as $|d|$
(iii) The A.M. of these differences (called deviations) is the Mean Deviation.

i.e. $\text{M.D.} = \frac{S|d|}{n}$

Example 3

Calculate the Mean Deviation of the following values; 2, 4, 7, 8 and 9.

The A.M. $= \frac{\Sigma X}{n} = 6$

X	$\lvert d \rvert$
2	4
4	2
7	1
8	2
9	3
	12

$$M.D._{(\bar{x})} = \frac{12}{5} = 2.4$$

Assumed Mean Method

Mean Deviation can also be calculated by calculating deviations from an assumed mean. This method is adopted especially when the actual mean is a fractional number. (Take care that the assumed mean is close to the true mean).

For the values in example 3, suppose value 7 is taken as assumed mean, M.D. can be calculated as under:

Example 4

X	$\lvert d \rvert$
2	5
4	3
7	0
8	1
9	2
	11

In such cases, the following formula is used,

$$M.D._{(\bar{x})} = \frac{\Sigma |d| + (\bar{x} - A\bar{x})(\Sigma f_B - \Sigma f_A)}{n}$$

Where $\Sigma |d|$ is the sum of absolute deviations taken from the assumed mean.

$\bar{x}$ is the actual mean.

$A\bar{x}$ is the assumed mean used to calculate deviations.

Σf_B is the number of values below the actual mean including the actual mean.

Σf_A is the number of values above the actual mean.

Substituting the values in the above formula:

$$M.D._{(\bar{x})} = \frac{11 + (6 - 7)(2 - 3)}{5} = \frac{12}{5} = 2.4$$

Mean Deviation from median for ungrouped data.

Direct Method

Using the values in example 3, M.D. from the Median can be calculated as follows,

(i) Calculate the median which is 7.

(ii) Calculate the absolute deviations from median, denote them as $|d|$.

(iii) Find the average of these absolute deviations. It is the Mean Deviation.

Example 5

	[X-Median]
X	$\lvert d \rvert$
2	5
4	3
7	0
8	1
9	2
	11

M. D. from Median is thus,

$$M.D._{(median)} = \frac{\Sigma |d|}{n} = \frac{11}{5} = 2.2$$

Short-cut method

To calculate Mean Deviation by short cut method a value (A) is used to calculate the deviations and the following formula is applied.

$$M.D._{(Median)} = \frac{\Sigma |d| + (Median - A)(\Sigma f_B - \Sigma f_A)}{n}$$

where, A = the constant from which deviations are calculated. (Other notations are the same as given in the assumed mean method).

Mean Deviation from Mean for Continuous distribution

TABLE 6.2

Profits of companies (Rs in lakhs) *Class-intervals*	Number of Companies frequencies
10–20	5
20–30	8
30–50	16
50–70	8
70–80	3
	40

Steps:

(i) Calculate the mean of the distribution.

(ii) Calculate the absolute deviations $|d|$ of the class midpoints from the mean.

(iii) Multiply each $|d|$ value with its corresponding frequency to get $f|d|$ values. Sum them up to get $\Sigma f|d|$.

(iv) Apply the following formula,

$$M.D._{(\bar{x})} = \frac{\Sigma f|d|}{\Sigma f}$$

Mean Deviation of the distribution in Table 6.2 can be calculated as follows:

Example 6

C.I.	*f*	*m.p.*	*|d|*	*f|d|*
10–20	5	15	25.5	127.5
20–30	8	25	15.5	124.0
30–50	16	40	0.5	8.0
50–70	8	60	19.5	156.0
70–80	3	75	34.5	103.5
	40			*519.0*

$$M.D._{(\bar{x})} = \frac{\Sigma f|d|}{\Sigma f} = \frac{519}{40} = 12.975$$

Mean Deviation from Median

TABLE 6.3

Class intervals	*Frequencies*
20–30	5
30–40	10
40–60	20
60–80	9
80–90	6
	50

The procedure to calculate Mean Deviation from the median is the same as it is in case of M.D. from Mean, except that deviations are to be taken from the median as given below:

Example 7

C.I.	*f*	*m.p.*	$\|d\|$	$f\|d\|$
20–30	5	25	25	125
30–40	10	35	15	150
40–60	20	50	0	0
60–80	9	70	20	180
80–90	6	85	35	210
	50			*665*

$$M.D._{(Median)} = \frac{\Sigma f\,|d|}{\Sigma f}$$

$$= \frac{665}{50} = 13.3$$

Mean Deviation: Comments

Mean Deviation is based on all values. A change in even one value will affect it. It is the least when calculated from the median i.e., it will be higher if calculated from the mean. However it ignores the signs of deviations and cannot be calculated for open-ended distributions.

Standard Deviation

Standard Deviation is the positive square root of the mean of squared deviations from mean. So if there are five values x_1, x_2, x_3, x_4 and x_5, first their mean is calculated. Then deviations of the values from mean are calculated. These deviations are then squared. The mean of these squared deviations is the variance. Positive square root of the *variance* is the standard deviation.

(Note that Standard Deviation is calculated on the basis of the mean only).

Calculation of Standard Deviation for ungrouped data

Four alternative methods are available for the calculation of standard deviation of individual values. All these methods result in the same value of standard deviation. These are:

(i) Actual Mean Method
(ii) Assumed Mean Method
(iii) Direct Method
(iv) Step-Deviation Method

Actual Mean Method:

Suppose you have to calculate the standard deviation of the following values:

5, 10, 25, 30, 50

Example 8

X	*d*	d^2
5	–19	361
10	–14	196
25	+1	1
30	+6	36
50	+26	676
	0	*1270*

Following formula is used:

$$s = \sqrt{\frac{\Sigma d^2}{n}}$$

$$s = \sqrt{\frac{1270}{5}} = \sqrt{254} = 15.937$$

Do you notice the value from which deviations have been calculated in the above example? Is it the Actual Mean?

Assumed Mean Method

For the same values, deviations may be calculated from any arbitrary value

$A_{\bar{X}}$ such that $d = X - A_{\bar{X}}$. Taking $A_{\bar{X}}$ = 25, the computation of the standard deviation is shown below:

Example 9

X	d	d^2
5	–20	400
10	–15	225
25	0	0
30	+5	25
50	+25	625
	–5	*1275*

Formula for Standard Deviation

$$\sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2}$$

$$\sigma = \sqrt{\frac{1275}{5} - \left(\frac{-5}{5}\right)^2} = \sqrt{254} = 15.937$$

The sum of deviations from a value other than actul mean is not equal to zero

Direct Method

Standard Deviation can also be calculated from the values directly, i.e., without taking deviations, as shown below:

Example 10

X	x^2
5	25
10	100
25	625
30	900
50	2500
120	*4150*

(This amounts to taking deviations from zero)

Following formula is used.

$$\sigma = \sqrt{\frac{\Sigma x^2}{n} - (\bar{x})^2}$$

or $$\sigma = \sqrt{\frac{4150}{5} - (24)^2}$$

or $$\sigma = \sqrt{254} = 15.937$$

Standard Deviation is not affected by the value of the constant from which deviations are calculated. The value of the constant does not figure in the standard deviation formula. Thus, Standard Deviation is *Independent of Origin.*

Step-deviation Method

If the values are divisible by a common factor, they can be so divided and standard deviation can be calculated from the resultant values as follows:

Example 11

Since all the five values are divisible by a common factor 5, we divide and get the following values:

x	x'	d	d^2
5	1	–3.8	14.44
10	2	–2.8	7.84
25	5	+0.2	0.04
30	6	+1.2	1.44
50	10	+5.2	27.04
		0	*50.80*

(Steps in the calculation are same as in actual mean method).

The following formula is used to calculate standard deviation:

$$s = \sqrt{\frac{\Sigma d^2}{n}} \cdot c$$

$$x' = \frac{x}{c}$$

c = common factor

Substituting the values,

$$s = \sqrt{\frac{50.80}{5}} \quad 5$$

$$s = \sqrt{10.16} \cdot 5$$

$$s = 15.937$$

Alternatively, instead of dividing the values by a common factor, the deviations can be divided by a common factor. Standard Deviation can be calculated as shown below:

Example 12

x	*d*	*d'*	d^2
5	–20	–4	16
10	–15	–3	9
25	0	0	0
30	+5	+1	1
50	+25	+5	25
		–1	*51*

Deviations have been calculated from an arbitrary value 25. Common factor of 5 has been used to divide deviations.

$$s = \sqrt{\frac{\Sigma d'^2}{n} - \left(\frac{\Sigma d'}{n}\right)^2} \cdot c$$

$$s = \sqrt{\frac{51}{5} - \left(\frac{-1}{5}\right)^2} \cdot 5$$

$$s = \sqrt{10.16} \cdot 5 = 15.937$$

> Standard Deviation is *not independent of scale.* Thus, if the values or deviations are divided by a common factor, the value of the common factor is used in the formula to get the value of Standard Deviation.

Standard Deviation in Continuous frequency distribution:

Like ungrouped data, S.D. can be calculated for grouped data by any of the following methods:

(i) Actual Mean Method
(ii) Assumed Mean Method
(iii) Step-Deviation Method

Actual Mean Method

For the values in Table 6.2, Standard Deviation can be calculated as follows:

Example 13

(1) CI	*(2) f*	*(3) m*	*(4) fm*	*(5) d*	*(6) fd*	*(7) fd^2*
10–20	5	15	75	–25.5	–127.5	3251.25
20–30	8	25	200	–15.5	–124.0	1922.00
30–50	16	40	640	–0.5	–8.0	4.00
50–70	8	60	480	+19.5	+156.0	3042.00
70–80	3	75	225	+34.5	+103.5	3570.75
	40		*1620*		*0*	*11790.00*

Following steps are required:

1. Calculate the mean of the distribution.

$$\bar{x} = \frac{\Sigma fm}{\Sigma f} = \frac{1620}{40} = 40.5$$

2. Calculate deviations of mid-values from the mean so that

$$d = m - \bar{x} \text{ (Col. 5)}$$

3. Multiply the deviations with their

corresponding frequencies to get 'fd' values (col. 6) [Note that $\sum fd = 0$]

4. Calculate 'fd^2' values by multiplying 'fd' values with 'd' values. (Col. 7). Sum up these to get $\sum fd^2$.
5. Apply the formula as under:

$$\sigma = \sqrt{\frac{\sum fd^2}{n}} = \sqrt{\frac{11790}{40}} = 17.168$$

Assumed Mean Method

For the values in example 13, standard deviation can be calculated by taking deviations from an assumed mean (say 40) as follows:

Example 14

(1) *CI*	*(2)* *f*	*(3)* *m*	*(4)* *d*	*(5)* *fd*	*(6)* *fd^2*
10–20	5	15	-25	–125	3125
20–30	8	25	-15	–120	1800
30–50	16	40	0	0	0
50–70	8	60	+20	160	3200
70–80	3	75	+35	105	3675
	40			*+20*	*11800*

The following steps are required:

1. Calculate mid-points of classes (Col. 3)
2. Calculate deviations of mid-points from an assumed mean such that $d = m - A_{\bar{x}}$ (Col. 4). Assumed Mean = 40.
3. Multiply values of 'd' with corresponding frequencies to get 'fd' values (Col. 5). (note that the total of this column is not zero since deviations have been taken from assumed mean).
4. Multiply 'fd' values (Col. 5) with 'd' values (col. 4) to get fd^2 values (col. 6). Find $\sum fd^2$.
5. Standard Deviation can be calculated by the following formula.

$$\sigma = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2}$$

$$\text{or } \sigma = \sqrt{\frac{11800}{40} - \left(\frac{20}{40}\right)^2}$$

$$\text{or } \sigma = \sqrt{294.75} = 17.168$$

Step-deviation Method

In case the values of deviations are divisible by a common factor, the calculations can be simplified by the step-deviation method as in the following example.

Example 15

(1) *CI*	*(2)* *f*	*(3)* *m*	*(4)* *d*	*(5)* *d'*	*(6)* *fd'*	*(7)* *fd'^2*
10–20	5	15	–25	–5	–25	125
20–30	8	25	–15	–3	–24	72
30–50	16	40	0	0	0	0
50–70	8	60	+20	+4	+32	128
70–80	3	75	+35	+7	+21	147
	40				*+4*	*472*

Steps required:

1. Calculate class mid-points (Col. 3) and deviations from an arbitrarily chosen value, just like in the assumed mean method. In this example, deviations have been taken from the value 40. (Col. 4)
2. Divide the deviations by a common factor denoted as 'C'. C = 5 in the

above example. The values so obtained are 'd'' values (Col. 5).

3. Multiply 'd'' values with corresponding 'f' values (Col. 2) to obtain 'fd'' values (Col. 6).
4. Multiply 'fd'' values with 'd'' values to get 'fd'^{2}' values (Col. 7)
5. Sum up values in Col. 6 and Col. 7 to get $\sum fd'$ and $\sum fd'^2$ values.
6. Apply the following formula.

$$s = \sqrt{\frac{\sum fd'^2}{\sum f} - \left(\frac{\sum fd'}{\sum f}\right)^2} \cdot c$$

or $$s = \sqrt{\frac{472}{40} - \left(\frac{4}{40}\right)^2} \cdot 5$$

or $$s = \sqrt{11.8 - .01} \cdot 5$$

or $$s = \sqrt{11.79} \cdot 5$$

$$s = 17.168$$

Standard Deviation: Comments

Standard Deviation, the most widely used measure of dispersion, is based on all values. Therefore a change in even one value affects the value of standard deviation. It is independent of origin but not of scale. It is also useful in certain advanced statistical problems.

5. Absolute and Relative Measures of Dispersion

All the measures, described so far, are absolute measures of dispersion. They calculate a value which, at times, is difficult to interpret. For example, consider the following two data sets:

Set A	500	700	1000
Set B	100000	120000	130000

Suppose the values in Set A are the daily sales recorded by an ice-cream vendor, while Set B has the daily sales of a big departmental store. Range for Set A is 500 whereas for Set B, it is 30,000. The value of Range is much higher in Set B. Can you say that the variation in sales is higher for the departmental store? It can be easily observed that the highest value in Set A is double the smallest value, whereas for the Set B, it is only 30% higher. Thus absolute measures may give misleading ideas about the extent of variation specially when the averages differ significantly.

Another weakness of absolute measures is that they give the answer in the units in which original values are expressed. Consequently, if the values are expressed in kilometers, the dispersion will also be in kilometers. However, if the same values are expressed in meters, an absolute measure will give the answer in meters and the value of dispersion will appear to be 1000 times.

To overcome these problems, relative measures of dispersion can be used. Each absolute measure has a relative counterpart. Thus, for Range, there is Coefficient of Range which is calculated as follows:

$$\text{Coefficient of Range} = \frac{L - S}{L + S}$$

where L = Largest value
S = Smallest value

Similarly, for Quartile Deviation, it

is Coefficient of Quartile Deviation which can be calculated as follows:

Coefficient of Quartile Deviation

$$= \frac{Q_3 - Q_1}{Q_3 + Q_1} \text{ where } Q_3 = 3^{rd} \text{ Quartile}$$

$Q_1 = 1^{st}$ Quartile

For Mean Deviation, it is Coefficient of Mean Deviation.

Coefficient of Mean Deviation =

$$\frac{\text{M.D.}(\bar{x})}{\bar{x}} \text{ or } \frac{\text{M.D.(Median)}}{\text{Median}}$$

Thus if Mean Deviation is calculated on the basis of the Mean, it is divided by the Mean. If Median is used to calculate Mean Deviation, it is divided by the Median.

For Standard Deviation, the relative measure is called Coefficient of Variation, calculated as below:

Coefficient of Variation

$$= \frac{\text{Standard Deviation}}{\text{Arithmetic Mean}} \cdot 100$$

It is usually expressed in percentage terms and is the most commonly used relative measure of dispersion. Since relative measures are free from the units in which the values have been expressed, they can be compared even across different groups having different units of measurement.

7. LORENZ CURVE

The measures of dispersion discussed so far give a numerical value of dispersion. A graphical measure called Lorenz Curve is available for estimating dispersion. You may have heard of statements like 'top 10% of the people of a country earn 50% of the national income while top 20% account for 80%'. An idea about income disparities is given by such figures. Lorenz Curve uses the information expressed in a cumulative manner to indicate the degree of variability. It is specially useful in comparing the variability of two or more distributions.

Given below are the monthly incomes of employees of a company.

TABLE 6.4

Incomes	*Number of employees*
0–5,000	5
5,000–10,000	10
10,000–20,000	18
20,000–40,000	10
40,000–50,000	7

Example 16

Income limits	*Mid-points*	*Cumulative mid-points*	*Cumulative mid-points as percentages*	*No. of employees frequencies*	*Comulative frequencies*	*Comulative frequencies as percentages*
(1)	(2)	(3)	(4)	(5)	(6)	(7)
0–5000	2500	2500	2.5	5	5	10
5000–10000	7500	10000	10.0	10	15	30
10000–20000	15000	25000	25.0	18	33	66
20000–40000	30000	55000	55.0	10	43	86
40000–50000	45000	100000	100.0	7	50	100

Construction of the Lorenz Curve

Following steps are required.

1. Calculate class mid-points and find cumulative totals as in Col. 3 in the example 16, given above.
2. Calculate cumulative frequencies as in Col. 6.
3. Express the grand totals of Col. 3 and 6 as 100, and convert the cumulative totals in these columns into percentages, as in Col. 4 and 7.
4. Now, on the graph paper, take the cumulative percentages of the variable (incomes) on Y axis and cumulative percentages of frequencies (number of employees) on X-axis, as in figure 6.1. Thus each axis will have values from '0' to '100'.
5. Draw a line joining Co-ordinate (0, 0) with (100,100). This is called the line of equal distribution shown as line 'OC' in figure 6.1.
6. Plot the cumulative percentages of the variable with corresponding cumulative percentages of frequency. Join these points to get the curve OAC.

Studying the Lorenz Curve

OC is called the line of equal distribution, since it would imply a situation like, top 20% people earn 20% of total income and top 60% earn 60% of the total income. The farther the curve OAC from this line, the greater is the variability present in the distribution. If there are two or more curves, the one which is the farthest from line OC has the highest dispersion.

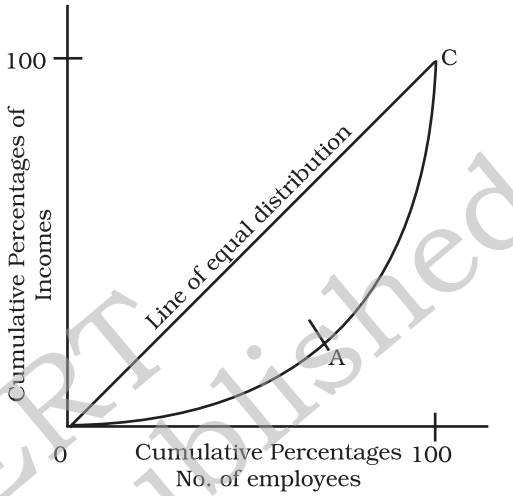


Fig. 6.1

8. CONCLUSION

Although Range is the simplest to calculate and understand, it is unduly affected by extreme values. QD is not affected by extreme values as it is based on only middle 50% of the data. However, it is more difficult to interpret M.D. and S.D. both are based upon deviations of values from their average. M.D. calculates average of deviations from the average but ignores signs of deviations and therefore appears to be unmathematical. Standard Deviation attempts to calculate average deviation from mean. Like M.D., it is based on all values and is also applied in more advanced statistical problems. It is the most widely used measure of dispersion.

Recap

- A measure of dispersion improves our understanding about the behaviour of an economic variable.
- Range and Quartile Deviation are based upon the spread of values.
- M.D. and S.D. are based upon deviations of values from the average.
- Measures of dispersion could be Absolute or Relative.
- Absolute measures give the answer in the units in which data are expressed.
- Relative smeasures are free from these units, and consequently can be used to compare different variables.
- A graphic method, which estimates the dispersion from shape of a curve, is called Lorenz Curve.

EXERCISES

1. A measure of dispersion is a good supplement to the central value in understanding a frequency distribution. Comment.
2. Which measure of dispersion is the best and how?
3. Some measures of dispersion depend upon the spread of values whereas some calculate the variation of values from a central value. Do you agree?
4. In a town, 25% of the persons earned more than Rs 45,000 whereas 75% earned more than 18,000. Calculate the absolute and relative values of dispersion.
5. The yield of wheat and rice per acre for 10 districts of a state is as under:

District	1	2	3	4	5	6	7	8	9	10
Wheat	12	10	15	19	21	16	18	9	25	10
Rice	22	29	12	23	18	15	12	34	18	12

   Calculate for each crop,
   (i) Range
   (ii) Q.D.
   (iii) Mean Deviation about Mean
   (iv) Mean Deviation about Median
   (v) Standard Deviation
   (vi) Which crop has greater variation?
   (vii) Compare the values of different measures for each crop.
6. In the previous question, calculate the relative measures of variation and indicate the value which, in your opinion, is more reliable.
7. A batsman is to be selected for a cricket team. The choice is between X and Y on the basis of their five previous scores which are:

X	25	85	40	80	120
Y	50	70	65	45	80

Which batsman should be selected if we want,

(i) a higher run getter, or

(ii) a more reliable batsman in the team?

8. To check the quality of two brands of lightbulbs, their life in burning hours was estimated as under for 100 bulbs of each brand.

Life (in hrs)	*No. of bulbs*	
	Brand A	*Brand B*
0–50	15	2
50–100	20	8
100–150	18	60
150–200	25	25
200–250	22	5
	100	*100*

(i) Which brand gives higher life?

(ii) Which brand is more dependable?

9. Averge daily wage of 50 workers of a factory was Rs 200 with a Standard Deviation of Rs 40. Each worker is given a raise of Rs 20. What is the new average daily wage and standard deviation? Have the wages become more or less uniform?

10. If in the previous question, each worker is given a hike of 10 % in wages, how are the Mean and Standard Deviation values affected?

11. Calculate the Mean Deviation about Mean and Standard Deviation for the following distribution.

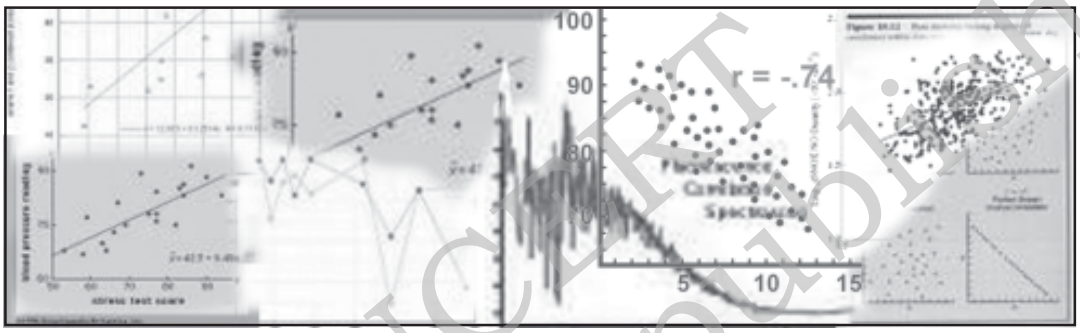
Classes	*Frequencies*
20–40	3
40–80	6
80–100	20
100–120	12
120–140	9
	50

12. The sum of 10 values is 100 and the sum of their squares is 1090. Find the Coefficient of Variation.

CHAPTER

7

Correlation



Studying this chapter should enable you to:

- *understand the meaning of the term correlation;*
- *understand the nature of relationship between two variables;*
- *calculate the different measures of correlation;*
- *analyse the degree and direction of the relationships.*

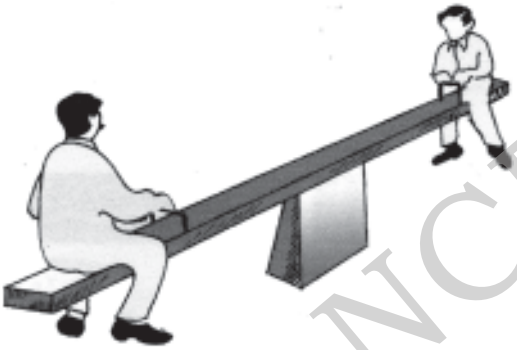
1. INTRODUCTION

In previous chapters you have learnt how to construct summary measures out of a mass of data and changes among similar variables. Now you will learn how to examine the relationship between two variables.

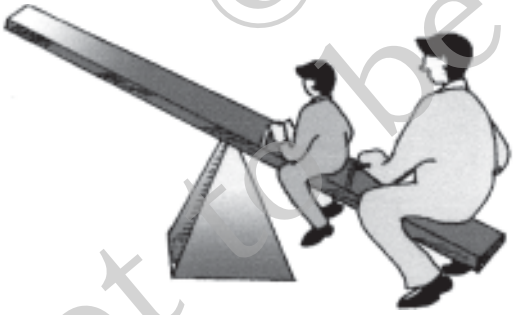
As the summer heat rises, hill stations, are crowded with more and more visitors. Ice-cream sales become more brisk. Thus, the temperature is related to number of visitors and sale of ice-creams. Similarly, as the supply of tomatoes increases in your local *mandi*, its price drops. When the local harvest starts reaching the market, the price of tomatoes drops from a princely Rs 40 per kg to Rs 4 per kg or even less. Thus supply is related to price. Correlation analysis is a means for examining such relationships systematically. It deals with questions such as:

- Is there any relationship between two variables?

- If the value of one variable changes, does the value of the other also change?



- Do both the variables move in the same direction?



- How strong is the relationship?

2. TYPES OF RELATIONSHIP

Let us look at various types of relationship. The relation between movements in quantity demanded and the price of a commodity is an integral part of the theory of demand, which you will read in class XII. Low rainfall is related to low agricultural productivity. Such examples of relationship may be given a cause and effect interpretation. Others may be just coincidence. The relation between the arrival of migratory birds in a sanctuary and the birth rates in the locality can not be given any cause and effect interpretation. The relationships are simple coincidence. The relationship between size of the shoes and money in your pocket is another such example. Even if relationship exist, they are difficult to explain it.

In another instance a third variable's impact on two variables may give rise to a relation between the two variables. Brisk sale of ice-creams may be related to higher number of deaths due to drowning. The victims are not drowned due to eating of ice-creams. Rising temperature leads to brisk sale of ice-creams. Moreover, large number of people start going to swimming pools to beat the heat. This might have raised the number of deaths by drowning. Thus temperature is behind the high correlation between the sale of ice-creams and deaths due to drowning.

What Does Correlation Measure?

Correlation studies and measures the direction and intensity of relationship among variables. Correlation measures covariation, not causation. Correlation should never be

interpreted as implying cause and effect relation. The presence of correlation between two variables X and Y simply means that when the value of one variable is found to change in one direction, the value of the other variable is found to change either in the same direction (i.e. positive change) or in the opposite direction (i.e. negative change), but in a definite way. For simplicity we assume here that the correlation, if it exists, is linear, i.e. the relative movement of the two variables can be represented by drawing a straight line on graph paper.

Types of Correlation

Correlation is commonly classified into negative and positive correlation. The correlation is said to be positive when the variables move together in the same direction. When the income rises, consumption also rises. When income falls, consumption also falls. Sale of ice-cream and temperature move in the same direction. The correlation is negative when they move in opposite directions. When the price of apples falls its demand increases. When the prices rise its demand decreases. When you spend more time in studying, chances of your failing decline. When you spend less hours in study, chances of your failing increase. These are instances of negative correlation. The variables move in opposite direction.

3. Techniques for Measuring Correlation

Widely used techniques for the study of correlation are scatter diagrams, Karl Pearson's coefficient of correlation and Spearman's rank correlation.

A scatter diagram visually presents the nature of association without giving any specific numerical value. A numerical measure of linear relationship between two variables is given by Karl Pearson's coefficient of correlation. A relationship is said to be linear if it can be represented by a straight line. Another measure is Spearman's coefficient of correlation, which measures the linear association between ranks assigned to indiviual items according to their attributes. Attributes are those variables which cannot be numerically measured such as intelligence of people, physical appearance, honesty etc.

Scatter Diagram

A scatter diagram is a useful technique for visually examining the form of relationship, without calculating any numerical value. In this technique, the values of the two variables are plotted as points on a graph paper. The cluster of points, so plotted, is referred to as a scatter diagram. From a scatter diagram, one can get a fairly good idea of the nature of relationship. In a scatter diagram the degree of closeness of the scatter points and their overall direction enable us to examine the relation-

ship. If all the points lie on a line, the correlation is perfect and is said to be unity. If the scatter points are widely dispersed around the line, the correlation is low. The correlation is said to be linear if the scatter points lie near a line or on a line.

Scatter diagrams spanning over Fig. 7.1 to Fig. 7.5 give us an idea of the relationship between two variables. Fig. 7.1 shows a scatter around an upward rising line indicating the movement of the variables in the same direction. When X rises Y will also rise. This is positive correlation. In Fig. 7.2 the points are found to be scattered around a downward sloping line. This time the variables move in opposite directions. When X rises Y falls and vice versa. This is negative correlation. In Fig.7.3 there is no upward rising or downward sloping line around which the points are scattered. This is an example of no correlation. In Fig. 7.4 and Fig. 7.5 the points are no longer scattered around an upward rising or downward falling line. The points themselves are on the lines. This is referred to as perfect positive correlation and perfect negative correlation respectively.

Activity

- Collect data on height, weight and marks scored by students in your class in any two subjects in class X. Draw the scatter diagram of these variables taking two at a time. What type of relationship do you find?

Inspection of the scatter diagram gives an idea of the nature and intensity of the relationship.

Karl Pearson's Coefficient of Correlation

This is also known as product moment correlation and simple correlation coefficient. It gives a precise numerical value of the degree of linear relationship between two variables X and Y. The linear relationship may be given by

$Y = a + bX$

This type of relation may be described by a straight line. The intercept that the line makes on the Y-axis is given by a and the slope of the line is given by b. It gives the change in the value of Y for very small change in the value of X. On the other hand, if the relation cannot be represented by a straight line as in

$Y = X^2$

the value of the coefficient will be zero. It clearly shows that zero correlation need not mean absence of any type of relation between the two variables.

Let $X_1, X_2, \ldots, X_N$ be N values of X and $Y_1, Y_2, \ldots, Y_N$ be the corresponding values of Y. In the subsequent presentations the subscripts indicating the unit are dropped for the sake of simplicity. The arithmetic means of X and Y are defined as

$$\bar{X} = \frac{\Sigma X}{N}; \qquad \bar{Y} = \frac{\Sigma Y}{N}$$

and their variances are as follows

$$\sigma^2_x = \frac{\Sigma(X - \bar{X})^2}{N} = \frac{\Sigma X^2}{N} - \bar{X}^2$$

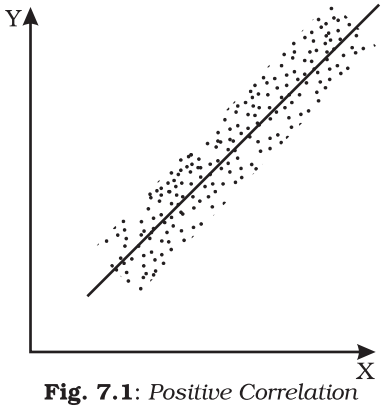


Fig. 7.1: *Positive Correlation*

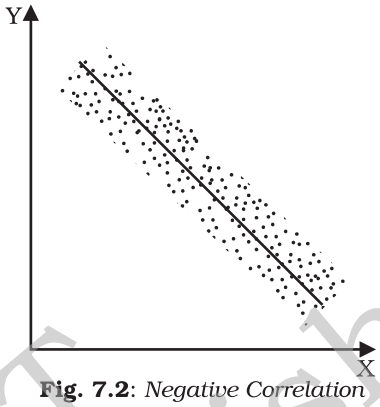


Fig. 7.2: *Negative Correlation*

Fig. 7.3: *No Correlation*

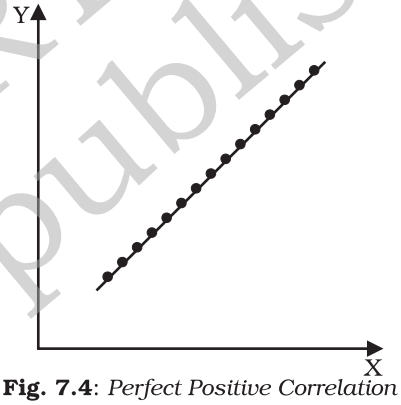


Fig. 7.4: *Perfect Positive Correlation*

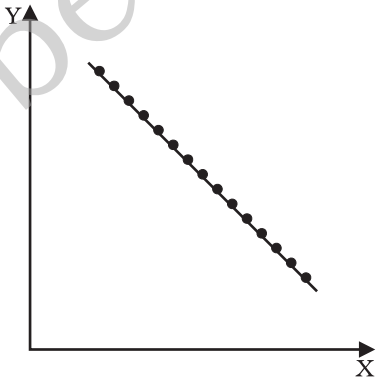


Fig. 7.5: Perfect Negative Correlation

and $\sigma^2_y = \frac{\Sigma(Y - \bar{Y})^2}{N} = \frac{\Sigma Y^2}{N} - \bar{Y}^2$

The standard deviations of X and Y respectively are the positive square roots of their variances. Covariance of X and Y is defined as

$$\text{Cov}(X,Y) = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{N} = \frac{\Sigma xy}{N}$$

Where $x = X-\bar{X}$ and $y = X - \bar{Y}$ are the deviations of the *i*th value of X and Y from their mean values respectively.

The sign of covariance between X and Y determines the sign of the correlation coefficient. The standard deviations are always positive. If the covariance is zero, the correlation coefficient is always zero. The product moment correlation or the Karl Pearson's measure of correlation is given by

$$r = \Sigma xy \Big/ N\sigma_x\sigma_y \quad ...(1)$$

or

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2}\sqrt{\Sigma(Y - \bar{Y})^2}} \quad ...(2)$$

or

$$r = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{\sqrt{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}\sqrt{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}}} \quad ...(3)$$

or

$$r = \frac{N\Sigma XY \quad (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 \quad (\Sigma X)^2}\sqrt{N\Sigma Y^2 \quad (\Sigma Y)^2}} \quad ...(4)$$

Properties of Correlation Coefficient

Let us now discuss the properties of the correlation coefficient

- *r* has no unit. It is a pure number. It means units of measurement are not part of *r*. *r* between height in feet and weight in kilograms, for instance, is 0.7.
- A negative value of *r* indicates an inverse relation. A change in one variable is associated with change in the other variable in the opposite direction. When price of a commodity rises, its demand falls. When the rate of interest rises the demand for funds also falls. It is because now funds have become costlier.

- If *r* is positive the two variables move in the same direction. When the price of coffee, a substitute of tea, rises the demand for tea also rises. Improvement in irrigation facilities is associated with higher yield. When temperature rises the sale of ice-creams becomes brisk.

- If $r = 0$ the two variables are uncorrelated. There is no linear relation between them. However other types of relation may be there.
- If $r = 1$ or $r = -1$ the correlation is perfect. The relation between them is exact.
- A high value of r indicates strong linear relationship. Its value is said to be high when it is close to +1 or –1.
- A low value of r indicates a weak linear relation. Its value is said to be low when it is close to zero.
- The value of the correlation coefficient lies between minus one and plus one, $-1 \leq r \leq 1$. If, in any exercise, the value of r is outside this range it indicates error in calculation.
- The value of r is unaffected by the change of origin and change of scale. Given two variables X and Y let us define two new variables.

$$U = \frac{X - A}{B}; \; V = \frac{Y - C}{D}$$

where A and C are assumed means of X and Y respectively, B and D are common factors. Then

$$r_{xy} = r_{uv}$$

This property is used to calculate correlation coefficient in a highly simplified manner, as in the step deviation method.

As you have read in chapter 1, the statistical methods are no substitute for common sense. Here, is another example, which highlights the need for understanding the data properly before correlation is calculated. An epidemic spreads in some villages and the government sends a team of doctors to the affected villages. The correlation between the number of deaths and the number of doctors sent to the villages is found to be positive. Normally the health care facilities provided by the doctors are expected to reduce the number of deaths showing a negative correlation. This happened due to other reasons. The data relate to a specific time period. Many of the reported deaths could be terminal cases where the doctors could do little. Moreover, the benefit of the presence of doctors becomes visible after some time. It is also possible that the reported deaths are not due to the epidemic. A tsunami suddenly hits the state and death toll rises.

Let us illustrate the calculation of r by examining the relationship between years of schooling of the farmer and the annual yield per acre.

Example 1

No. of years of schooling of farmers	*Annual yield per acre in '000 (Rs)*
0	4
2	4
4	6
6	10
8	10
10	8
12	7

Formula 1 needs the value of $\Sigma xy, \sigma_x, \sigma_y$

From Table 7.1 we get,

$\Sigma xy = 42$,

$$\sigma_x = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N}} = \sqrt{\frac{112}{7}},$$

$$\sigma_y = \sqrt{\frac{\Sigma(Y - \bar{Y})^2}{N}} = \sqrt{\frac{38}{7}}$$

Substituting these values in formula (1)

$$r = \frac{42}{7\sqrt{\frac{112}{7}}\sqrt{\frac{38}{7}}} = 0.644$$

The same value can be obtained from formula (2) also.

$$r = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma (X - \bar{X})^2}\sqrt{\Sigma (Y - \bar{Y})^2}} \quad ...(2)$$

$$r = \frac{42}{\sqrt{112}\sqrt{38}} = 0.644$$

Thus years of education of the farmers and annual yield per acre are positively correlated. The value of *r* is also large. It implies that more the number of years farmers invest in education, higher will be the yield per acre. It underlines the importance of farmers' education.

To use formula (3)

$$r = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{\sqrt{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}\sqrt{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}}} \quad ...(3)$$

the value of the following expressions have to be calculated i.e. $\Sigma XY, \Sigma X^2, \Sigma Y^2$.

Now apply formula (3) to get the value of *r*.

Let us know the interpretation of different values of *r*. The correlation coefficient between marks secured in English and Statistics is, say, 0.1. It means that though the marks secured in the two subjects are positively correlated, the strength of the relationship is weak. Students with high marks in English may be getting relatively low marks in statistics. Had the value of *r* been, say, 0.9, students with high marks in English will invariably get high marks in Statistics.

TABLE 7.1
Calculation of *r* between years of schooling of farmers and annual yield

Years of Education (X)	$(X-\bar{X})$	$(X-\bar{X})^2$	*Annual yield per acre in '000 Rs (Y)*	$(Y-\bar{Y})$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
0	–6	36	4	–3	9	18
2	–4	16	4	–3	9	12
4	–2	4	6	–1	1	2
6	0	0	10	3	9	0
8	2	4	10	3	9	6
10	4	16	8	1	1	4
12	6	36	7	0	0	0
$\Sigma X=42$		$\Sigma (X-\bar{X})^2=112$	$\Sigma Y=49$		$\Sigma (Y-\bar{Y})^2=38$	$\Sigma (X-\bar{X})(Y-\bar{Y})=42$

An example of negative correlation is the relation between arrival of vegetables in the local mandi and price of vegetables. If r is –0.9, vegetable supply in the local mandi will be accompanied by lower price of vegetables. Had it been –0.1 large vegetable supply will be accompanied by lower price, not as low as the price, when *r* is –0.9. The extent of price fall depends on the absolute value of *r*. Had it been zero there would have been no fall in price, even after large supplies in the market. This is also a possibility if the increase in supply is taken care of by a good transport network transferring it to other markets.

Activity

- Look at the following table. Calculate r between annual growth of national income at current price and the Gross Domestic Saving as percentage of GDP.

Step deviation method to calculate correlation coefficient.

When the values of the variables are large, the burden of calculation can be considerably reduced by using a property of *r*. It is that *r* is independent of change in origin and scale. It is also known as step deviation method. It involves the transformation of the variables X and Y as follows:

TABLE 7.2

Year	*Annual growth of National Income*	*Gross Domestic Saving as percentage of GDP*
1992–93	14	24
1993–94	17	23
1994–95	18	26
1995–96	17	27
1996–97	16	25
1997–98	12	25
1998–99	16	23
1999–00	11	25
2000–01	8	24
2001–02	10	23

Source: *Economic Survey, (2004–05) Pg. 8,9*

a property of *r*. It is that *r* is independent of change in origin and scale. It is also known as step deviation method. It involves the transformation of the variables X and Y as follows:

$$U = \frac{X - A}{h}; V = \frac{Y - B}{k}$$

where A and B are assumed means, h and k are common factors.

Then $r_{UV} = r_{XY}$

This can be illustrated with the exercise of analysing the correlation between price index and money supply.

Example 2

Price index (X)	120	150	190	220	230
Money supply in Rs crores (Y)	1800	2000	2500	2700	3000

The simplification, using step deviation method is illustrated below. Let A = 100; h = 10; B = 1700 and k = 100

The table of transformed variables is as follows:

Calculation of *r* between price index and money supply using step deviation method

TABLE 7.3

U $\left(\frac{X-100}{10}\right)$	V $\left(\frac{Y-1700}{100}\right)$	U^2	V^2	UV
2	1	4	1	2
5	3	25	9	15
9	8	81	64	72
12	10	144	100	120
13	13	169	169	169

$\Sigma U = 41$; $\Sigma V = 35$; $\Sigma U^2 = 423$;

$\Sigma V^2 = 343$; $\Sigma UV = 378$

Substituting these values in formula (3)

$$r = \frac{\Sigma UV - \frac{(\Sigma U)(\Sigma V)}{N}}{\sqrt{\Sigma U^2 - \frac{(\Sigma U)^2}{N}}\ \sqrt{\Sigma V^2 - \frac{(\Sigma V)^2}{N}}} \quad (3)$$

$$= \frac{378 - \frac{41 \times 35}{5}}{\sqrt{423 - \frac{(41)^2}{5}}\ \sqrt{343 - \frac{(35)^2}{5}}}$$

$= 0.98$

This strong positive correlation between price index and money supply is an important premise of monetary policy. When the money supply grows the price index also rises.

Activity

- Take some examples of India's population and national income. Calculate the correlation between them using step deviation method and see the simplification.

Spearman's rank correlation

Spearman's rank correlation was developed by the British psychologist C.E. Spearman. It is used when the variables cannot be measured meaningfully as in the case of price, income, weight etc. Ranking may be more meaningful when the measurements of the variables are suspect. Consider the situation where we are required to calculate the correlation between height and weight of students in a remote village. Neither measuring rods nor weighing scales are available. The students can be easily ranked in terms of height and weight without using measuring rods and weighing scales.

There are also situations when you are required to quantify qualities such as fairness, honesty etc. Ranking may be a better alternative to quantification of qualities. Moreover, sometimes the correlation coefficient between two variables with extreme values may be quite different from the coefficient without the extreme values. Under these circumstances rank correlation provides a better alternative to simple correlation.

Rank correlation coefficient and simple correlation coefficient have the same interpretation. Its formula has

been derived from simple correlation coefficient where individual values have been replaced by ranks. These ranks are used for the calculation of correlation. This coefficient provides a measure of linear association between ranks assigned to these units, not their values. It is the Product Moment Correlation between the ranks. Its formula is

$$r_k = 1 - \frac{6\Sigma D^2}{n^3 - n} \quad \ldots(4)$$

where *n* is the number of observations and D the deviation of ranks assigned to a variable from those assigned to the other variable. When the ranks are repeated the formula is

$$r_k = 1 - \frac{6\left[\Sigma D^2 + \frac{(m^3_1 - m_1)}{12} + \frac{(m^3_2 - m_2)}{12} + \ldots\right]}{n(n^2 - 1)}$$

where m_1, m_2, ..., are the number of repetitions of ranks and $\frac{m^3_1 - m_1}{12}$ ..., their corresponding correction factors. This correction is needed for every repeated value of both variables. If three values are repeated, there will be a correction for each value. Every time m_1 indicates the number of times a value is repeated.

All the properties of the simple correlation coefficient are applicable here. Like the Pearsonian Coefficient of correlation it lies between 1 and –1. However, generally it is not as accurate as the ordinary method. This is due the fact that all the information concerning the data is not utilised. The first differences of the values of the items in the series, arranged in order of magnitude, are almost never constant. Usually the data cluster around the central values with smaller differences in the middle of the array. If the first differences were constant then *r* and r_k would give identical results. The first difference is the difference of consecutive values. Rank correlation is preferred to Pearsonian coefficient when extreme values are present. In general r_k is less than or equal to *r*.

The calculation of rank correlation will be illustrated under three situations.

1. The ranks are given.
2. The ranks are not given. They have to be worked out from the data.
3. Ranks are repeated.

Case 1: When the ranks are given

Example 3

Five persons are assessed by three judges in a beauty contest. We have to find out which pair of judges has the nearest approach to common perception of beauty.

	Competitors				
Judge	1	2	3	4	5
A	1	2	3	4	5
B	2	4	1	5	3
C	1	3	5	2	4

There are 3 pairs of judges necessitating calculation of rank correlation thrice. Formula (4) will be used —

$$r_s = 1 - \frac{6\Sigma D^2}{n^3 - n} \quad \ldots(4)$$

The rank correlation between A and B is calculated as follows:

A	*B*	*D*	D^2
1	2	–1	1
2	4	–2	4
3	1	2	4
4	5	–1	1
5	3	2	4
Total			14

Substituting these values in formula (4)

$$r_s = 1 - \frac{6\Sigma D^2}{n^3 - n} \quad \ldots(4)$$

$$= 1 - \frac{6 \times 14}{5^3 - 5} = 1 - \frac{84}{120} = 1 - 0.7 = 0.3$$

The rank correlation between A and C is calculated as follows:

A	*C*	*D*	D^2
1	1	0	0
2	3	–1	1
3	5	–2	4
4	2	2	4
5	4	1	1
Total			*10*

Substituting these values in formula (4) the rank correlation is 0.5. Similarly, the rank correlation between the rankings of judges B and C is 0.9. Thus, the perceptions of judges A and C are the closest. Judges B and C have very different tastes.

Case 2: When the ranks are not given

Example 4

We are given the percentage of marks, secured by 5 students in Economics and Statistics. Then the ranking has to be worked out and the rank correlation is to be calculated.

Student	*Marks in Statistics (X)*	*Marks in Economics (Y)*
A	85	60
B	60	48
C	55	49
D	65	50
E	75	55

Student	*Ranking in Statistics* (R_x)	*Ranking in Economics* (R_y)
A	1	1
B	4	5
C	5	4
D	3	3
E	2	2

Once the ranking is complete formula (4) is used to calculate rank correlation.

Case 3: When the ranks are repeated

Example 5

The values of X and Y are given as

X	25	45	35	40	15	19	35	42
Y	55	60	30	35	40	42	36	48

In order to work out the rank correlation, the ranks of the values are worked out. Common ranks are given to the repeated items. The

common rank is the mean of the ranks which those items would have assumed if they were slightly different from each other. The next item will be assigned the rank next to the rank already assumed. The formula of Spearman's rank correlation coefficient when the ranks are repeated is as follows

$$r_s = 1 - \frac{6\left[\Sigma D^2 + \frac{(m^3_1 - m_1)}{12} + \frac{(m^3_2 - m_2)}{12} + \ldots\right]}{n(n^2 - 1)}$$

where m_1, m_2, ..., are the number of repetitions of ranks and $\frac{m^3_1 - m_1}{12}$ ..., their corresponding correction factors.

X has the value 35 both at the 4th and 5th rank. Hence both are given the average rank i.e.,

$$\frac{4+5}{2}\text{th} = 4.5\text{th rank}$$

X	*Y*	*Rank of XR′*	*Rank of YR″*	*Deviation in Ranking D=R′–R″*	D^2
25	55	6	2	4	16
45	80	1	1	0	0
35	30	4.5	8	3.5	12.25
40	35	3	7	–4	16
15	40	8	5	3	9
19	42	7	4	3	9
35	36	4.5	6	–1.5	2.25
42	48	2	3	–1	1
Total			$\Sigma D = 65.5$		

The necessary correction thus is

$$\frac{m^3 - m}{12} = \frac{2^3 - 2}{12} = \frac{1}{2}$$

Using this equation

$$r_s = 1 - \frac{6\left[\Sigma D^2 + \frac{(m^3 - m)}{12}\right]}{n^3 - n} \quad \ldots(5)$$

Substituting the values of these expressions

$$r_s = 1 - \frac{6(65.5 + 0.5)}{8^3 - 8} = 1 - \frac{396}{504}$$
$$= 1 - 0.786 = 0.214$$

Thus there is positive rank correlation between X and Y. Both X and Y move in the same direction. However, the relationship cannot be described as strong.

Activity

- Collect data on marks scored by 10 of your classmates in class IX and X examinations. Calculate the rank correlation coefficient between them. If your data do not have any repetition, repeat the exercise by taking a data set having repeated ranks. What are the circumstances in which rank correlation coefficient is preferred to simple correlation coefficient? If data are precisely measured will you still prefer rank correlation coefficient to simple correlation? When can you be indifferent to the choice? Discuss in class.

4. Conclusion

We have discussed some techniques for studying the relationship between

two variables, particularly the linear relationship. The scatter diagram gives a visual presentation of the relationship and is not confined to linear relations. Measures of correlation such as Karl Pearson's coefficient of correlation and Spearman's rank correlation are strictly the measures of linear relationship. When the variables cannot be measured precisely, rank correlation can meaningfully be used. These measures however do not imply causation. The knowledge of correlation gives us an idea of the direction and intensity of change in a variable when the correlated variable changes.

Recap

- Correlation analysis studies the relation between two variables.
- Scatter diagrams give a visual presentation of the nature of relationship between two variables.
- Karl Pearson's coefficient of correlation r measures numerically only linear relationship between two variables. r lies between –1 and 1.
- When the variables cannot be measured precisely Spearman's rank correlation can be used to measure the linear relationship numerically.
- Repeated ranks need correction factors.
- Correlation does not mean causation. It only means covariation.

EXERCISES

1. The unit of correlation coefficient between height in feet and weight in kgs is
   (i) kg/feet
   (ii) percentage
   (iii) non-existent
2. The range of simple correlation coefficient is
   (i) 0 to infinity
   (ii) minus one to plus one
   (iii) minus infinity to infinity
3. If r_{xy} is positive the relation between X and Y is of the type
   (i) When Y increases X increases
   (ii) When Y decreases X increases
   (iii) When Y increases X does not change

4. If $r_{xy} = 0$ the variable X and Y are
   (i) linearly related
   (ii) not linearly related
   (iii) independent
5. Of the following three measures which can measure any type of relationship
   (i) Karl Pearson's coefficient of correlation
   (ii) Spearman's rank correlation
   (iii) Scatter diagram
6. If precisely measured data are available the simple correlation coefficient is
   (i) more accurate than rank correlation coefficient
   (ii) less accurate than rank correlation coefficient
   (iii) as accurate as the rank correlation coefficient
7. Why is r preferred to covariance as a measure of association?
8. Can *r* lie outside the –1 and 1 range depending on the type of data?
9. Does correlation imply causation?
10. When is rank correlation more precise than simple correlation coefficient?
11. Does zero correlation mean independence?
12. Can simple correlation coefficient measure any type of relationship?
13. Collect the price of five vegetables from your local market every day for a week. Calculate their correlation coefficients. Interpret the result.
14. Measure the height of your classmates. Ask them the height of their benchmate. Calculate the correlation coefficient of these two variables. Interpret the result.
15. List some variables where accurate measurement is difficult.
16. Interpret the values of *r* as 1, –1 and 0.
17. Why does rank correlation coefficient differ from Pearsonian correlation coefficient?
18. Calculate the correlation coefficient between the heights of fathers in inches (X) and their sons (Y)

X	65	66	57	67	68	69	70	72
Y	67	56	65	68	72	72	69	71

(Ans. r = 0.603)

19. Calculate the correlation coefficient between X and Y and comment on their relationship:

X	–3	–2	–1	1	2	3
Y	9	4	1	1	4	9

(Ans. r = 0)

20. Calculate the correlation coefficient between X and Y and comment on their relationship

X	1	3	4	5	7	8
Y	2	6	8	10	14	16

(Ans. r = 1)

Activity

- Use all the formulae discussed here to calculate r between India's national income and export taking at least ten observations.

CHAPTER

8

Index Numbers

Studying this chapter should enable you to:

- *understand the meaning of the term index number;*
- *become familiar with the use of some widely used index numbers;*
- *calculate an index number;*
- *appreciate its limitations.*

1. INTRODUCTION

You have learnt in the previous chapters how summary measures can be obtained from a mass of data. Now you will learn how to obtain summary measures of change in a group of related variables.

Rabi goes to the market after a long gap. He finds that the prices of most commodities have changed. Some items have become costlier, while others have become cheaper. On his return from the market, he tells his father about the change in price of the each and every item, he bought. It is bewildering to both. The industrial sector consists of many subsectors. Each of them is changing. The output of some subsectors are rising, while it is falling in some subsectors. The changes are not uniform. Description of the individual rates of change will be difficult to understand. Can a single figure summarise these changes? Look at the following cases:

Case 1

An industrial worker was earning a salary of Rs 1,000 in 1982. Today, he

earns Rs 12,000. Can his standard of living be said to have risen 12 times during this period? By how much should his salary be raised so that he is as well off as before?

Case 2

You must be reading about the sensex in the newspapers. The sensex crossing 8000 points is, indeed, greeted with euphoria. When, sensex dipped 600 points recently, it eroded investors' wealth by Rs 1,53,690 crores. What exactly is sensex?

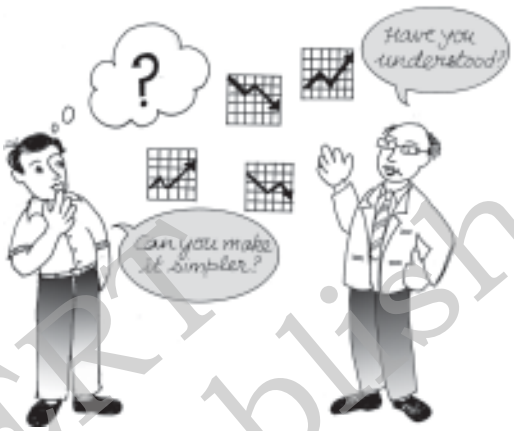
Case 3

The government says inflation rate will not accelerate due to the rise in the price of petroleum products. How does one measure inflation?

These are a sample of questions you confront in your daily life. A study of the index number helps in analysing these questions.

2. WHAT IS AN INDEX NUMBER

An index number is a statistical device for measuring changes in the magnitude of a group of related variables. It represents the general trend of diverging ratios, from which it is calculated. It is a measure of the average change in a group of related variables over two different situations. The comparison may be between like categories such as persons, schools, hospitals etc. An index number also measures changes in the value of the variables such as prices of specified list of commodities, volume of production in different sectors of an industry, production of various agricultural crops, cost of living etc.



Conventionally, index numbers are expressed in terms of percentage. Of the two periods, the period with which the comparison is to be made, is known as the base period. The value in the base period is given the index number 100. If you want to know how much the price has changed in 2005 from the level in 1990, then 1990 becomes the base. The index number of any period is in proportion with it. Thus an index number of 250 indicates that the value is two and half times that of the base period.

Price index numbers measure and permit comparison of the prices of certain goods. Quantity index numbers measure the changes in the physical volume of production, construction or employment. Though price index numbers are more widely used, a production index is also an important indicator of the level of the output in the economy.

3. CONSTRUCTION OF AN INDEX NUMBER

In the following sections, the principles of constructing an index number will be illustrated through price index numbers.

Let us look at the following example:

Example 1

Calculation of simple aggregative price index

TABLE 8.1

Commodity	*Base period price (Rs)*	*Current period price (Rs)*	*Percentage change*
A	2	4	100
B	5	6	20
C	4	5	25
D	2	3	50

As you observe in this example, the percentage changes are different for every commodity. If the percentage changes were the same for all four items, a single measure would have been sufficient to describe the change. However, the percentage changes differ and reporting the percentage change for every item will be confusing. It happens when the number of commodities is large, which is common in any real market situation. *A price index represents these changes by a single numerical measure.*

There are two methods of constructing an index number. It can be computed by the *aggregative method* and by the *method of averaging relatives.*

The Aggregative Method

The formula for a *simple aggregative price index* is

$$P_{01} = \frac{\Sigma P_1}{\Sigma P_0} \times 100$$

Where P_1 and P_0 indicate the price of the commodity in the current period and base period respectively. Using the data from example 1, the *simple aggregative price index* is

$$P_{01} = \frac{4+6+5+3}{2+5+4+2} \times 100 = 138.5$$

Here, price is said to have risen by 38.5 percent.

Do you know that such an index is of limited use? The reason is that the units of measurement of prices of various commodities are not the same. It is unweighted, because the relative importance of the items has not been properly reflected. The items are treated as having equal importance or weight. But what happens in reality? In reality the items purchased differ in order of importance. Food items occupy a large proportion of our expenditure. In that case an equal rise in the price of an item with large weight and that of an item with low weight will have different implications for the overall change in the price index.

The formula for a *weighted aggregative price index* is

$$P_{01} = \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1} \times 100$$

An index number becomes a weighted index when the relative

importance of items is taken care of. Here weights are quantity weights. To construct a weighted aggregative index, a well specified basket of commodities is taken and its worth each year is calculated. It thus measures the changing value of a fixed aggregate of goods. Since the total value changes with a fixed basket, the change is due to price change. Various methods of calculating a weighted aggregative index use different baskets with respect to time.

Example 2

Calculation of *weighted aggregative price index*

TABLE 8.2

Commodity	*Base period*		*Current period*	
	Price P_0	*Quantity* q_0	*Price* p_1	*Quality* q_1
A	2	10	4	5
B	5	12	6	10
C	4	20	5	15
D	2	15	3	10

$$P_{01} = \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1} \times 100$$

$$= \frac{4\times10 + 6\times12 + 5\times20 + 3\times15}{2\times10 + 5\times12 + 4\times20 + 2\times15} \times 100$$

$$= \frac{257}{190} \times 100 = 135.3$$

This method uses the base period quantities as weights. A weighted aggregative price index using base period quantities as weights, is also known as *Laspeyre's price index*. It provides an explanation to the question that if the expenditure on base period basket of commodities was Rs 100, how much should be the expenditure in the current period on the same basket of commodities? As you can see here, the value of base period quantities has risen by 35.3 per cent due to price rise. Using base period quantities as weights, the price is said to have risen by 35.3 percent.

Since the current period quantities differ from the base period quantities, the index number using current period weights gives a different value of the index number.

$$P_{01} = \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1} \times 100$$

$$= \frac{4\times5 + 6\times10 + 5\times15 + 3\times10}{2\times5 + 5\times10 + 4\times15 + 2\times15} \times 100$$

$$= \frac{185}{140} \times 100 = 132.1$$

It uses the current period quantities as weights. A weighted aggregative price index using current period quantities as weights is known as *Paasche's price index.* It helps in answering the question that, if the

the current period basket of commodities was consumed in the base period and if we were spending Rs 100 on it, how much should be the expenditure in current period on the same basket of commodities. A Paasche's price index of 132.1 is interpreted as a price rise of 32.1 percent. Using current period weights, the price is said to have risen by 32.1 per cent.

Method of Averaging relatives

When there is only one commodity, the price index is the ratio of the price of the commodity in the current period to that in the base period, usually expressed in percentage terms. The method of averaging relatives takes the average of these relatives when there are many commodities. The price index number using *price relatives* is defined as

$$P_{01} = \frac{1}{n}\Sigma\frac{p_1}{p_0}\times 100$$

where P_1 and P_o indicate the price of the *i*th commodity in the current period and base period respectively. The ratio $(P_1/P_0) \times 100$ is also referred to as price relative of the commodity. *n* stands for the number of commodities. In the current example

$$P_{01} = \frac{1}{4}\left(\frac{4}{2}+\frac{6}{5}+\frac{5}{4}+\frac{3}{2}\right)\times 100 = 149$$

Thus the prices of the commodities have risen by 49 percent.

The *weighted index of price relatives* is the weighted arithmetic mean of price relatives defined as

$$P_{01} = \frac{\Sigma W\left(\frac{P_1}{P_0}\times 100\right)}{\Sigma W}$$

where W = Weight.

In a *weighted price relative index* weights may be determined by the proportion or percentage of expenditure on them in total expenditure during the base period. It can also refer to the current period depending on the formula used. These are, essentially, the value shares of different commodities in the total expenditure. In general the base period weight is preferred to the current period weight. It is because calculating the weight every year is inconvenient. It also refers to the changing values of different baskets. They are strictly not comparable. Example 3 shows the type of information one needs for calculating *weighted price index.*

Example 3

Calculation of *weighted price relatives index*

TABLE 8.3

Commodity	*Base year price (in Rs.)*	*Current year price (in Rs)*	*Price relative*	*Weight in %*
A	2	4	200	40
B	5	6	120	30
C	4	5	125	20
D	2	3	150	10

The weighted price index is

$$P_{01} = \frac{\Sigma W\left(\frac{P_1}{P_0} \times 100\right)}{\Sigma W}$$

$$= \frac{40 \times 200 + 30 \times 120 + 20 \times 125 + 10 \times 150}{100}$$

= 156

The *weighted price index* is 156. The price index has risen by 56 percent. The values of the *unweighted price index* and the *weighted price index* differ, as they should. The higher rise in the weighted index is due to the doubling of the most important item A in example 3.

Activity

- Interchange the current period values with the base period values, in the data given in example 2. Calculate the price index using Laspeyre's, and Paasche's formula. What difference do you observe from the earlier illustration?

4. Some Important Index Numbers

Consumer price index

Consumer price index (CPI), also known as *the cost of living index*, measures the average change in retail prices. The *CPI* for industrial workers is increasingly considered the appropriate indicator of general inflation, which shows the most accurate impact of price rise on the cost of living of common people. Consider the statement that the *CPI*

Consumer Price Index

In India three CPI's are constructed. They are CPI for industrial workers (1982 as base), CPI for urban non manual employees (1984–85 as base) and CPI for agricultural labourers (base 1986–87). They are routinely calculated every month to analyse the impact of changes in the retail price on the cost of living of these three broad categories of consumers. The CPI for industrial workers and agricultural labourers are published by Labour Bureau, Shimla. The Central Statistical Organisation publishes the CPI number of urban non manual employees. This is necessary because their typical consumption baskets contain many dissimilar items.

The weight scheme in CPI for industrial workers (1982=100) by major commodity groups is given in the following table. In this scheme food has the largest weight. Food being the most important category, any rise in the food price will have a significant impact on CPI. This also explains the government's frequent statement that oil price hike will not be inflationary.

Major Group	*Weight in %*
Food	57.00
Pan, supari, tobacco etc.	3.15
Fuel & light	6.28
Housing	8.67
Clothing, bedding & footwear	8.54
Misc. group	16.36
General	*100.00*

Source: *Economic Survey, Government of India.*

for industrial workers(1982=100) is 526 in January 2005. What does this statement mean? It means that if the industrial worker was spending Rs 100 in 1982 for a typical basket of commodities, he needs Rs 526 in January 2005 to be able to buy an identical basket of commodities. It is not necessary that he/she buys the basket. What is important is whether he has the capability to buy it.

Example 4

Construction of *consumer price index* number.

TABLE 8.4

Item	*Weight in % W*	*Base period price (Rs)*	*Current period price (Rs)*	$R=P_1/P_0 \times 100$ *(in%)*	*WR*
Food	35	150	145	96.67	3883.45
Fuel	10	25	23	92.00	920.00
Cloth	20	75	65	86.67	1733.40
Rent	15	30	30	100.00	1500.00
Misc.	20	40	45	112.50	2250.00
					9786.85

$$CPI = \frac{\Sigma WR}{\Sigma W} = \frac{9786.85}{100} = 97.86$$

This exercise shows that the cost of living has declined by 2.14 per cent. What does an index larger than 100 indicate? It means a higher cost of living necessitating an upward adjustment in wages and salaries. The rise is equal to the amount, it exceeds 100. If the index is 150, 50 percent upward adjustment is required. The salaries of the employees have to be raised by 50 per cent.

Wholesale price index

The *wholesale price index number indicates the change in the general price level.* Unlike the CPI, it does not have any reference consumer category. It does not include items pertaining to services like barber charges, repairing etc.

What does the statement "WPI with 1993-94 as base is 189.1 in March, 2005" mean? It means that the general price level has risen by 89.1 percent during this period.

Industrial production index

The *index number of industrial production measures changes in the level of industrial production* comprising many industries. It includes the production of the public and the private sector. It is a weighted average of quantity relatives. The formula for the index is

$$IIP_{01} = \frac{\Sigma q_1 \times W}{\Sigma W} \times 100$$

In India, it is currently calculated every month with 1993–94 as the base. In table 8.6, you can see the index number of some industrial groupings along with their weights.

Wholesale Price Index

The commodity weights in the WPI are determined by the estimates of the commodity value of domestic production and the value of imports inclusive of import duty during the base year. It is available on a weekly basis. Commodities are broadly classified into three categories viz primary articles, fuel, power, light and lubricants and manufactured products. The weight scheme is given below. The low weight of fuel,power,light and lubricants explains how the government can get away with such a statement that the oil price hike will not be inflationary at least in the short run.

TABLE 8.5

Category	*Weight in %*	*No. of items*
Primary articles	22.0	98
Fuel, power, light & lubricants	14.2	19
Manufactured products	63.8	318

Source: *Economic Survey 2004–2005, Govt. of India, p–89*

TABLE 8.6
Broad industrial grouping and their weights

Broad groupings	*Weight in %*	*Index no. in May, 2005*
Mining and quarrying	10.47	155.2
Manufacturing	79.36	222.7
Electricity	10.17	196.7
General index		213.0

As the table shows, the growth performances of the broad industrial categories differ. The general index represents the average performance of these categories. Why does a comparatively lower performance of mining and quarrying not pull down the general index?

Index number of agricultural production

Index number of agricultural production is a weighted average of quantity relatives. Its base period is the triennium ending 1981-82. In 2003–04 the index number of agricultural production was 179.5. It means that agricultural production has increased by 79.5 percent over the average of the three years 1979–80, 1980–81 and 1981–82. Foodgrains have a weight of 62.92 percent in this index.

SENSEX

You ofen come across a news item in a newspaper,

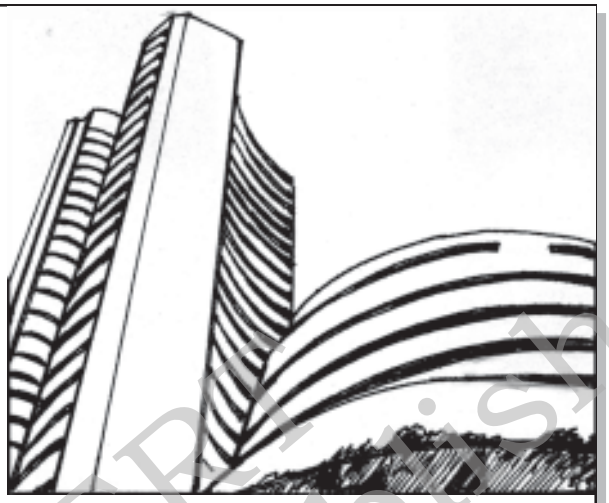
"Sensex breaches 8700 mark. BSE closes at 8650 points. Investor wealth rises by Rs 9,000 crore. The sensex broke the 8700 mark for the first time in its history but ended off the mark at 8650, also a new record closing level".

The rise in sensex was at the highest level till date, which reflects the good health of the economy in general. As the share prices increase, reflected by the rise in sensex, the value of wealth of the shareholders also rises.

Look at another news item,

"Sensex dips 600 in 30 days flat. Rs 1,53,690 crore investor wealth eroded. While the sensex has lost 338

Bombay Stock Exchange



Sensex is the short form of Bombay Stock Exchange Sensitive Index with 1978–79 as base. The value of the sensex is with reference to this period. It is the benchmark index for the Indian stock market. It consists of 30 stocks which represent 13 sectors of the economy and the companies listed are leaders in their respective industries. If the sensex rises, it indicates that the market is doing well and investors expect better earnings from companies. It also indicates a growing confidence of investors in the basic health of the economy.

points in two consecutive days, it has eroded 6.8% or 598 points since October 4 when it hit an all time high at 8800 points. Investor wealth eroded by a staggering Rs 1,53,690 crore or 6.7% during the period."

It shows that all is not well with the health of the economy. The investors may find it hard to decide whether to invest or not.

Another useful index in recent years is the human development index. Very soon producers price index number will replace wholesale price index.

Producer Price Index

Producer price index number measures price changes from the producers' perspective. It uses only basic prices including taxes, trade margins and transport costs. A Working Group on Revision of Wholesale Price Index (1993–94=100) is inter alia examining the feasibility of switching over from WPI to a PPI in India as in many countries.

5. ISSUES IN THE CONSTRUCTION OF AN INDEX NUMBER

You should keep certain important issues in mind, while constructing an index number.

- You need to be clear about the purpose of the index. Calculation of a volume index will be inappropriate, when one needs a value index.

• Besides this, the items are not equally important for different groups of consumers when a consumer price index is constructed. The rise in petrol price may not directly impact the living condition of the poor agricultural labourers. Thus the items to be included in any index have to be selected carefully to be as representative as possible. Only then you will get a meaningful picture of the change.

• Every index should have a base. This base should be as normal as possible. Extreme values should not be selected as base period. The period should also not belong to too far in the past. The comparison between 1993 and 2005 is much more meaningful than a comparison between 1960 and 2005. Many items in a 1960 typical consumption basket have disappeared at present. Therefore, the base year for any index number is routinely updated.

• Another issue is the choice of the formula, which depends on the nature of question to be studied. The only difference between the Laspeyres' index and Paasche's index is the weights used in these formulae.

• Besides, there are many sources of data with different degrees of reliability. Data of poor reliability will give misleading results. Hence, due care should be taken in the collection of data. If primary data are not being used, then the most reliable source of secondary data should be chosen.

Activity

• Collect data from the local vegetable market over a week for, at least 10 items. Try to construct the daily price index for the week. What problems do you encounter in applying both methods for the construction of a price index?

6. INDEX NUMBER IN ECONOMICS

Why do we need to use the index numbers? Wholesale price index number (WPI), consumer price index number (CPI) and industrial production index (IIP) are widely used in policy making.

• Consumer index number (CPI) or cost of living index numbers are helpful in wage negotiation, formulation of income policy, price policy, rent control, taxation and general economic policy formulation.

• The wholesale price index (WPI) is used to eliminate the effect of changes in prices on aggregates such as national income, capital formation etc.

• The WPI is widely used to measure the rate of inflation. Inflation is a general and continuing increase in prices. If inflation becomes sufficiently large, money may lose its traditional function as a medium of exchange and as a unit of account. Its primary impact lies in lowering the value of money. The weekly inflation rate is given by

$$\frac{X_t - X_{t-1}}{X_{t-1}} \times 100$$ where X_t and X_{t-1}

refer to the WPI for the t th and (t-1) th weeks.

• CPI are used in calculating the purchasing power of money and real wage:

(i) Purchasing power of money = 1/ Cost of living index

(ii) Real wage = (Money wage/Cost of living index) × 100

If the CPI (1982=100) is 526 in January 2005 the equivalent of a rupee in January, 2005 is given by

$\text{Rs}\ \frac{100}{526} = 0.19$. It means that it is worth 19 paise in 1982. If the money wage of the consumer is Rs 10,000, his real wage will be

$$\text{Rs}\ 10,000 \times \frac{100}{526} = \text{Rs}\ 1,901$$

It means Rs 1,901 in 1982 has the same purchasing power as Rs 10,000 in January, 2005. If he/she was getting Rs 3,000 in 1982, he/she is worse off due to the rise in price. To maintain the 1982 standard of living the salary should be raised to Rs 15,780 obtained by multiplying the base period salary by the factor 526/100.

• Index of industrial production gives us a quantitative figure about the change in production in the industrial sector.

• Agricultural production index provides us a ready reckoner of the performane of agricultural sector.

• Sensex is a useful guide for investors in the stock market. If the sensex is rising, investors are optimistic of the future performance of the economy. It is an appropriate time for investment.

Where can we get these index numbers?

Some of the widely used index numbers are routinely published in the Economic Survey, an annual publication of the Government of India are WPI, CPI, Index Number of Yield of Principal Crops, Index of Industrial Production, Index of Foreign Trade.

Activity

• Check from the newspapers and construct a time series of sensex with 10 observations. What happens when the base of the consumer price index is shifted from 1982 to 2000?

7. Conclusion

Thus, the method of the index number enables you to calculate a single measure of change of a large number of items. Index numbers can be calculated for price, quantity, volume etc.

It is also clear from the formulae that the index numbers need to be interpreted carefully. The items to be included and the choice of the base period are important. Index numbers are extremely important in policy making as is evident by their various uses.

Recap

- An index number is a statistical device for measuring relative change in a large number of items.
- There are several formulae for working out an index number and every formula needs to be interpreted carefully.
- The choice of formula largely depends on the question of interest.
- Widely used index numbers are wholesale price index, consumer price index, index of industrial production, agricultural production index and sensex.
- The index numbers are indispensable in economic policy making.

EXERCISES

1. An index number which accounts for the relative importance of the items is known as
   (i) weighted index
   (ii) simple aggregative index
   (iii) simple average of relatives
2. In most of the weighted index numbers the weight pertains to
   (i) base year
   (ii) current year
   (iii) both base and current year
3. The impact of change in the price of a commodity with little weight in the index will be
   (i) small
   (ii) large
   (iii) uncertain
4. A consumer price index measures changes in
   (i) retail prices
   (ii) wholesale prices
   (iii) producers prices
5. The item having the highest weight in consumer price index for industrial workers is
   (i) Food
   (ii) Housing
   (iii) Clothing
6. In general, inflation is calculated by using
   (i) wholesale price index
   (ii) consumer price index
   (iii) producers' price index

7. Why do we need an index number?
8. What are the desirable properties of the base period?
9. Why is it essential to have different CPI for different categories of consumers?
10. What does a consumer price index for industrial workers measure?
11. What is the difference between a price index and a quantity index?
12. Is the change in any price reflected in a price index number?
13. Can the CPI number for urban non-manual employees represent the changes in the cost of living of the President of India?
14. The monthly per capita expenditure incurred by workers for an industrial centre during 1980 and 2005 on the following items are given below. The weights of these items are 75,10, 5, 6 and 4 respectively. Prepare a weighted index number for cost of living for 2005 with 1980 as the base.

Items	*Price in 1980*	*Price in 2005*
Food	100	200
Clothing	20	25
Fuel & lighting	15	20
House rent	30	40
Misc	35	65

15. Read the following table carefully and give your comments.

INDEX OF INDUSTRIAL PRODUCTION BASE 1993–94

Industry	*Weight in %*	*1996–97*	*2003–2004*
General index	100	130.8	189.0
Mining and quarrying	10.73	118.2	146.9
Manufacturing	79.58	133.6	196.6
Electricity	10.69	122.0	172.6

16. Try to list the important items of consumption in your family.
17. If the salary of a person in the base year is Rs 4,000 per annum and the current year salary is Rs 6,000, by how much should his salary rise to maintain the same standard of living if the CPI is 400?
18. The consumer price index for June, 2005 was 125. The food index was 120 and that of other items 135. What is the percentage of the total weight given to food?
19. An enquiry into the budgets of the middle class families in a certain city gave the following information;

Expenses on items	*Food 35%*	*Fuel 10%*	*Clothing 20%*	*Rent 15%*	*Misc. 20%*
Price (in Rs) in 2004	1500	250	750	300	400
Price (in Rs) in 1995	1400	200	500	200	250

What is the cost of living index of 2004 as compared with 1995?

20. Record the daily expenditure, quantities bought and prices paid per unit of the daily purchases of your family for two weeks. How has the price change affected your family?

21. Given the following data-

Year	*CPI of industrial workers (1982 =100)*	*CPI of urban non-manual employees (1984–85 = 100)*	*CPI of agricultural labourers (1986–87 = 100)*	*WPI (1993–94=100)*
1995–96	313	257	234	121.6
1996–97	342	283	256	127.2
1997–98	366	302	264	132.8
1998–99	414	337	293	140.7
1999–00	428	352	306	145.3
2000–01	444	352	306	155.7
2001–02	463	390	309	161.3
2002–03	482	405	319	166.8
2003–04	500	420	331	175.9

Source: *Economic Survey, Government of India.2004–2005*

(i) Calculate the inflation rates using different index numbers.
(ii) Comment on the relative values of the index numbers.
(iii) Are they comparable?

Activity

- Consult your class teacher to make a list of widely used index numbers. Get the most recent data indicating the source. Can you tell what the unit of an index number is?
- Make a table of consumer price index for industrial workers in the last 10 years and calculate the purchasing power of money. How is it changing?

CHAPTER

9

Use of Statistical Tools

Studying this chapter should enable you to:

- *be familiar with steps in designing a project;*
- *apply various statistical tools in analysing a problem.*

INTRODUCTION

You have studied about the various statistical tools. These tools are important for us in daily life and are used in the analysis of data pertaining to economic activities such as production, consumption, distribution, banking and insurance, trade, transport, etc. In this chapter, you will learn the method of developing a project. This will help in understanding how statistical tools and methods can be used for various types of analysis. For example, you may have to collect information about a product from the consumer or about a new product or service to be launched in the market by the producer or analyse the spread of information technology in schools and so on. Developing a project by conducting a survey and preparing a report will help in analysing relevant information and suggesting improvements in a product or system.

Steps Towards Making a Project

Identifying a problem or an area of study

At the outset, you should be clear about what you want to study. On the basis

of your objective, you will proceed with the collection and processing of the data. For example, production or sale of a product like car, mobile phone, shoe polish, bathing soap or a detergent, may be an area of interest to you. You may like to address certain water or electricity problems relating to households of a particular area. You may like to study about consumer awareness among households, i.e., awareness about rights of consumers.

Choice of Target Group

The choice or identification of the target group is important for framing appropriate questions for your questionnaire. If your project relates to cars, then your target group will mainly be the middle income and the higher income groups. For the project studies relating to consumer products like soap, you will target all rural and urban consumers. For the availability of safe drinking water your target group can be both urban and rural population. Therefore, the choice of target groups, to identify those persons on whom you focus your attention, is very important while preparing the project report.

Collection of Data

The objective of the survey will help you to determine whether the data collection should be undertaken by using primary method, secondary method or both the methods. As you have read in Chapter 2, a first hand collection of data by using primary method can be done by using a questionnaire or an interview schedule, which may be obtained by personal interviews, mailing/postal surveys, phone, email, etc. Postal questionnaire must have a covering letter giving details about the purpose of inquiry. Your objective will be to determine the size and characteristics of your target group. For example, in a survey pertaining to the primary and secondary level female literacy or consumption of a particular brand or soap, you will have to go to each and every family or household to collect the information.

Secondary data will provide information through published or unpublished sources (internal record of any organisation), provided it suits your requirement. Secondary sources of data are usually used when there is paucity of time, money and manpower resources and the information is easily available. If sampling is used in your method of data collection, then the care has to be taken about the suitability of the method of sampling.

Organisation and Presentation of Data

After collecting the data, you need to process the information so received, by organising and presenting with the help of tabulation and suitable diagrams, e.g. bar diagrams, pie diagrams, etc. about which you have studied in chapter 3 and 4.

Analysis and Interpretation

Measures of Central Tendency (e.g. mean), Measures of Dispersion (e.g.

Standard deviation), and Correlation will enable you to calculate the average, variability and relationship, if it exists among the variables. You have acquired the knowledge related to above-mentioned measures in chapters 5, 6 and 7.

Conclusion

The last step will be to draw meaningful conclusions after Analysing and Interpreting the results. If possible you must try to predict the **future prospects** and suggestions relating to growth and government policies, etc. on the basis of the information collected.

Bibliography

In this section, you need to mention the details of all the secondary sources, i.e., magazines, newspapers, research reports used for developing the project.

SUGGESTED LIST OF PROJECTS

These are a few suggested projects. You are free to choose any topic that deals with an economic issue.

1. Consider yourself as an advisor to Transport Minister who aims to bring about a better and coordinated system of transportation. Prepare a project report.
2. You may be working in a village cottage industry. It could be a unit manufacturing *dhoop, agarbatti,* candles, jute products, etc. You want to start a new unit of your own. Prepare a project proposal for getting a bank loan.
3. Suppose you are a marketing manager in a company and recently you have put up advertisements about your consumer product. Prepare a report on the effect of advertisements on the sale of your product.
4. You are a District Education Officer, who wants to assess the literacy levels and the reasons for dropping out of school children. Prepare a report.
5. Suppose you are a Vigilance Officer of an area and you receive complaints about overcharging of goods by traders i.e., charging a higher price than the Maximum Retail Price (MRP). Visit a few shops and prepare a report on the complaint.
6. Consider yourself to be a Mukhiya (head of Gram Panchayat) of a particular village who wants to improve amenities like safe drinking water to your people. Address your issues in a report form.
7. As a representative of a local government, you want to assess the participation of women in various employment schemes in your area. Prepare a project report.
8. You are the Chief Health Officer of a rural block. Identify the issues to be addressed through a project study. This may include health and sanitation problems in the area.
9. As the Chief Inspector of Food and Civil Supplies department, you have received a complaint about food adulteration in the area of

your duty. Conduct a survey to find the magnitude of the problem.

10. Prepare a report on Polio immunisation programme in a particular area.
11. You are a Bank Officer and want to survey the saving habits of the people by taking into consideration income and expenditure of the people. Prepare a report.
12. Suppose you are a part of a group of students who wants to study farming practices and the problems facing farmers in a village. Prepare a project report.

SAMPLE PROJECT

This is a sample project for your guidance. The question can vary depending upon the subject of the study.

You are a young entrepreneur who wants to setup a new retail shop and want to choose a variety of toothpaste brands to sell. A sample project based on primary source of data collection could be prepared for toothpaste.

You have to start by assuring the concerned person or party, that the information required is for survey and will not be used for any other purpose. This is done through a covering letter. All the information shall be kept confidential.

Data Analysis and Interpretation

After collecting the entire information you now have to organise and classify data for the purpose of choosing brands of toothpaste which you want

to sell. Hypothetical data is given below for your reference where you will now use the statistical tools such as pie diagrams, bar diagrams, mean, standard deviation, etc.

Area Distribution

Urban users67%
Rural users33%

Observation: Majority of users belonged to urban area.

Age distribution

Age in years	*No. of Persons*
Below 10	74
10–20	56
20–30	91
30–40	146
40–50	93
Above 50	40
Total	*500*

QUESTIONNAIRE

1.Name

Age (in years)No. of persons

(a)Below 10 ☐
(b)10–20 ☐
(c)20–30 ☐
(d)30–40 ☐
(e)40–50 ☐
(f)Above 50 ☐

3.Gender: Male/Female ☐

4.Number of members in the family:

(a)1–2 ☐
(b)3–4 ☐
(c)5–6 ☐
(d)Above 6 ☐

5.How many earning members are there in your family? ☐

6.Monthly family income:

(a)Below 10,000 ☐
(b)10,000–20,000 ☐
(c)20,000–30,000 ☐
(d)Above 30,000 ☐

7.Resident of: Urban/Rural area ☐

8.Major occupation of the main bread-earner:

(a)Service ☐
(b)Professional ☐
(c)Manufacturer ☐
(d)Trader ☐
(e)Any other (please specify) ☐

9.What do you use to clean your teeth:

(a)Toothpaste ☐
(b)Toothpowder ☐
(c)Anyother ☐

10.Which brand of toothpaste do you use?

(a)Aquafresh ☐ (b)Anchor ☐
(c)Cibaca ☐ (d)Babool ☐
(e)Close-up ☐ (f)Promise ☐
(g)Colgate ☐ (h)Forhans ☐

(i)Meswak	☐	(j)Tea Tree Oil & Neem	☐
(k)Pepsodent	☐	(l)Oral B	☐
(m)Pearl 32	☐	(n)True Dent	☐
(o)Homeodent	☐	(p)Sensodyne	☐
(q)Any other	☐		

11.The price paid for each 100 gram pack of the toothpaste: ☐

12.Do you find the product costly?Y es/No

13.Do you examine the date of manufacturing and expiry of the product?Y es/No

14.Do you check the standardisation mark (like – ISI)?Y es/No

15.Do you check the ingredients used?Y es/No

16.Are you satisfied with the quality of the product?Y es/No

17.Do you complain to the shopkeeper in case of dissatisfaction?Y es/No

18.Has your complaint been timely attended?Y es/No

19.Did you ever go to a consumer court in case of dissatisfaction regarding the product?Y es/No

20.Was your complaint attended to your satisfaction?Y es/No

21.How did you come to know about the product?

Advertisement	*Families Influenced*
Television	☐
Newspaper	☐
Magazine	☐
Cinema	☐
Sales representative	☐
Exhibits - stall	☐
Radio	☐

22.Is the advertisement of the product persuasive?Y es/No

23.Were you attracted by promotional offers like rebates, free tooth brush, buy one get one free, etc.?Y es/No

24.Do the children influence purchase of particular toothpaste?Y es/No

25.If a new toothpaste is launched in the market will you buy it?Y es/No
If yes, then with what considerations? Kindly mention.

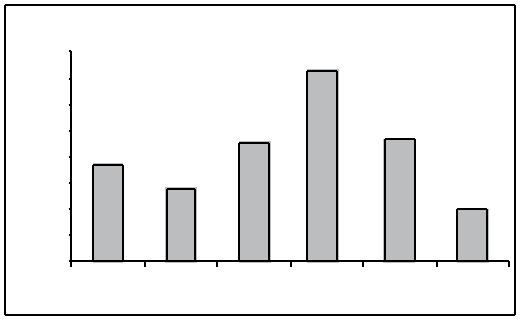


Fig. 9.1: *Bar diagram*

Observation: Majority of the persons surveyed belonged to age group 20–50.

Family Size

Family size	*No. of families*
1–2	20
3–4	40
5–6	30
Above 6	10
Total	*100*

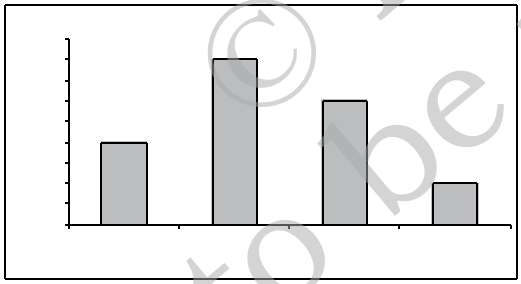


Fig. 9.2: *Bar diagram*

Observation: Majority of the families surveyed have 3–6 members.

Family monthly Income status

Income	*No. of Households*
Below 10,000	20
10, 000–20,000	40
20, 000–30,000	30
Above 30,000	10

Bar Diagram and Histogram respectively are indicating the level of families income.

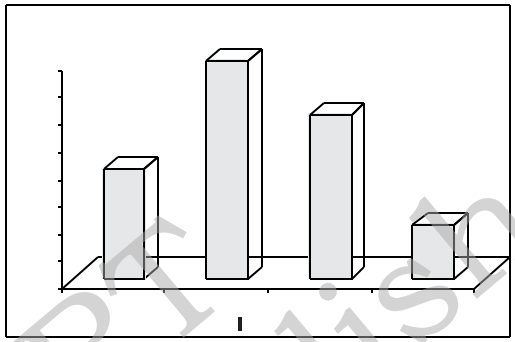


Fig. 9.3: *Bar diagram*

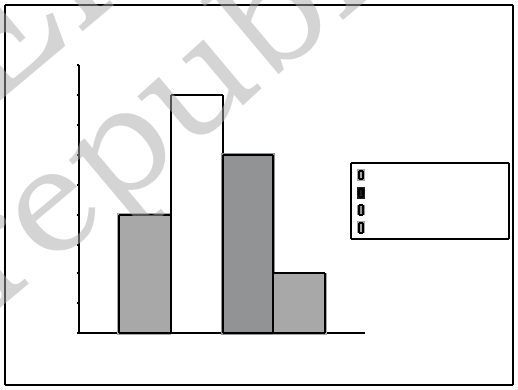


Fig. 9.4: *Histogram*

Observation: Majority of the families surveyed have monthly income between 10,000 to 30,000.

Monthly Family budget on toiletries

Items	*Expense (in Rs)*
Toothpaste	60
Soap	45
Shampoo	140
Shaving cream	25

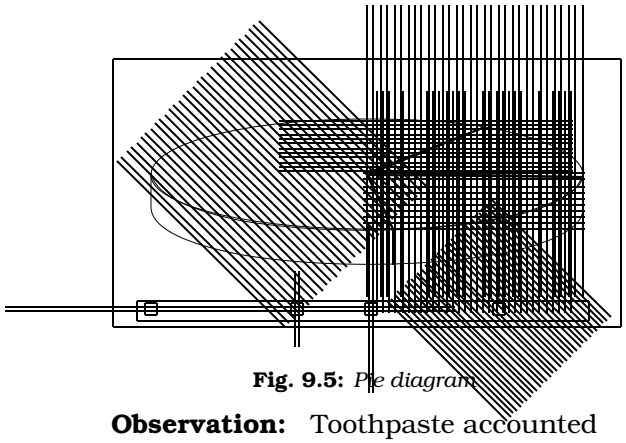


Fig. 9.5: *Pie diagram*

Observation: Toothpaste accounted for significant expenditure in family budget amongst toiletries.

Major Occupational status

Family Occupation	*No. of families*
Service	30
Professional	5
Manufacturer	10
Trader	40
Any other (please specify)	15

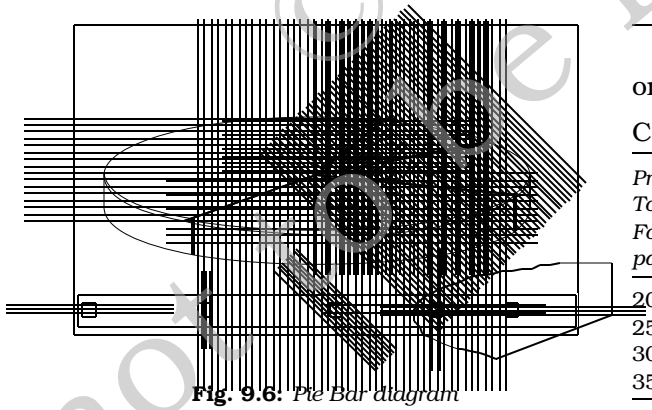


Fig. 9.6: *Pie Bar diagram*

Observation: Majority of the families surveyed were either service class or traders.

Preferred use of toothpaste

Brand	*Units*	*Brand*	*Units*
Aquafresh	5	Anchor	5
Cibaca	10	Babool	2
Close-up	15	Promise	10
Colgate	20	Forhans	0
Meswak	5	Tea tree oil & Neem	8
Pepsodent	25	Oral B	11
Pearl 32	4	True Dent	10
Homeodent	6	Sensodyne	8
Any other	0		

Observation: Pepsodent, Colgate and Close-up were the most preferred brands.

Price of the toothpaste

Prices of Toothpaste For 100 gram pack (Rs)	*No. of Households*
20–25	20
25–30	40
30–35	30
35–40	10
Total	*100*

Calculate the mean and dispersion on the basis of the above information.

Calculation of Mean,

Price of Toothpaste For 100 gm pack (Rs)	*No. of Households* f	*Mid Points* m	*fm*
20–25	20	22.5	450
25–30	40	27.5	
30–35	30	32.5	975
35–40	10	37.5	375
Total	*100*		*2900*

$$\bar{X} = \frac{\Sigma f}{\Sigma f} = \frac{m}{\quad} =$$

Observation: The average price of toothpaste across all brands is Rs 29.

Use of other statistical tools,

Prices of ToothpasteFor 100 gmfm/5 pack (Rs.)	*No. ofHouseholds(m*	*Midd'=Points*	*fd'*	*fd'*
20–25	2	2	2.5	–1
25–30	4	2	7.5	0
30–35	3	3	2.5	1
35–40	1	3	7.5	2
Total	*10*	*03*	*09*	*0*

Applying the formula of SD

$$\sigma = \frac{fd'}{N} \quad \frac{\sum fd'}{N} \times C$$

$$= \frac{}{} \times$$

$$= \times = \pm \times =$$

Observation: Price of the most toothpaste ranged between Rs 25–35

Basis of selection

Features	*Family members*
Liked the advertisement	15
Persuaded by the Dentist	5
Price	35
Quality	45
Taste	20
Ingredients	10
Standardised marking	50
Tried new product	10
Company's brand name	35

Observation: Majority of the people choose to buy the toothpaste for Standardised markings, quality, price and company's brand name.

Taste and Preferences

Brand	*Satisfied*	*Unsatisfied*
Aquafresh	5	15
Cibaca	10	5
Close up	15	10
Colgate	20	10
Meswak	5	15
Pepsodent	25	5
Anchor	5	10
Babool	2	0
Promise	10	14
Forhans	0	0
Tea tree oil and Neem	8	10
Oral B	11	15
True Dent	10	5
Sensodyne	8	3
Pearl 3	24	5
Homeodent	6	2

Observation: Amongst the most used toothpastes the percentage of dissatisfaction was relatively less.

Ingredients Preference

Plain toothpaste	15
Gel toothpaste	5
Antiseptic toothpaste	35
Flavoured toothpaste	25
Caries protective toothpaste	40
Gum toothpaste	10

Observation: Majority of the people preferred caries protective and antiseptic based toothpastes over the others.

Media Influence

Advertisement	*Families Influenced*
Television	47
News paper	30
Magazine	20
Cinema	25
Sales representative	15
Exhibits - stall	10
Radio	18

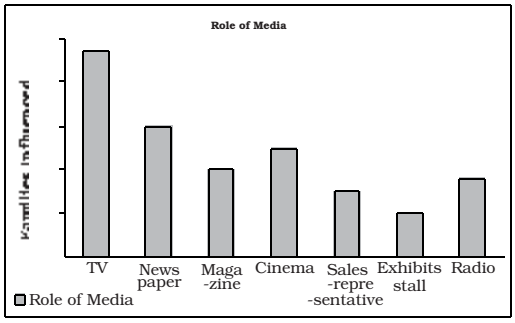


Fig. 9.7: *Bar diagram*

Observation: Majority of people came to know about the product either through television or through newspaper.

CONCLUSION/PROJECT REPORT

Majority of the users belonged to urban area. Most of the people who were surveyed belonged to age group 25 years to 50 years and had an average 3–6 members in a family. The monthly income of these families ranged between Rs 10,000 to Rs 30,000 and their main occupations were services and trading. Expenditure on toothpaste accounted for a major share in their family budget amongst toiletries. Pepsodent, Colgate and Close-up were the most preferred brands in the households surveyed. By calculating the mean it was found that the price of an average toothpaste would be Rs 29 approximately for 100 grams. People preferred those brands of toothpaste which has either a caries protection or antiseptic base. A lot of people get influenced with advertisement and the most popular medium to get across through people is television.

Recap

- The objective of the study should be clearly identified.
- The population and sample has to be chosen carefully.
- The objective of survey will indicate the type of data to be used.
- A questionnaire/interview schedule is prepared.
- Collected data can be analysed by using various statistical tools.
- Results are interpreted to draw a meaningful conclusion.