

INTRODUCTION TO STATISTICS

Textbook for Class XI

Author
A.L. NAGAR

Editor
G.S. KUSHWAHA



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

CONTENTS

CHAPTER 1 : Introduction	1
CHAPTER 2 : Collection of Data	6
CHAPTER 3 : Organization of Data	15
CHAPTER 4 : Presentation of Data	24
CHAPTER 5 : Frequency Curves and Diagrams	37
CHAPTER 6 : Measures of Central Tendency	48
CHAPTER 7 : Measures of Dispersion	63
CHAPTER 8 : Coefficient of Correlation	78
CHAPTER 9 : Introduction to Index Numbers	95
CHAPTER 10: Project on Application of Statistical Methods in Economics	105
APPENDIX A. TABLE OF TWO-DIGIT RANDOM NUMBERS	107
APPENDIX B. QUESTIONNAIRE	111
APPENDIX C. SOME IMPORTANT SOURCES OF SECONDARY DATA	114
GLOSSARY OF STATISTICAL TERMS	115
ANSWERS	117

CHAPTER 1

Introduction

1. Meaning of Statistics

According to the **Oxford Dictionary** the term 'statistics', used in **plural** means:

'Numerical facts systematically collected';

and in **singular** it means

'Science of collecting, classifying, and using statistics'; or

'Statistical fact or item'.

The Random House Dictionary of the English Language gives the following meaning of 'statistics':

"**(Construed as singular)** the science that deals with the collection, classification, analysis and interpretation of numerical facts or data, and that, by use of mathematical theories of probability, imposes order and regularity on aggregates of more or less disparate elements. **(Construed as plural)** the numerical facts or data themselves".

Statistical methods of analysis of statistical data may be purely descriptive or probabilistic. The **descriptive methods** include:

- (i) collection and classification of data,
- (ii) presenting them in tabular and diagrammatic forms, and
- (iii) calculating summary indices to measure certain characteristics of data.

We will discuss them in some detail in the following chapters. However, the probabilistic methods are beyond the scope of the present book.

The statistical data may be quantitative or qualitative.

Quantitative Data

We have quantitative data, if the variables can be measured in numerical terms. For example, daily temperatures, heights and weights of individuals, prices and incomes, etc. are quantitative variables. Their values can be expressed numerically.

Qualitative Data

Sometimes, it is not possible to measure variables numerically, in the same straight forward manner as heights and weights of individuals, or, prices of commodities, or, income of individuals. For example, attitudes of people to a political system, intelligence of individuals and their aptitudes toward music and art, beauty of individuals or some objects (like flowers, gardens, etc.) cannot be numerically measured. However, we may rank them according to the quality of their attributes. We may compare two objects and say that A is

more beautiful than B, B is more beautiful than C, and so on. We may assign rank '1' to the most beautiful, rank '2' to the second best, and so on. The lowest rank may be assigned to the least beautiful. Similarly, performance of artists (musicians, painters, etc.) cannot be numerically measured, but judges may assign ranks to them. We have qualitative data in such cases. The ranks may be used as numerical measurements for purposes of statistical analysis.

2. Scope of Statistics

Statistical data are used by us in everyday life. For example, while preparing the family budget for a month, we need data on prices of various goods and services to decide what proportion of the income should be allocated for various items of consumption (like food, clothing, travel, schooling of children, etc.). The performance of students in an examination is described in terms of marks secured by them. We may compare the performance of different schools in terms of the number of students passed in Class X. How many students got a first, second and third division and how many failed?

Economists use statistical data to analyse trends in prices of various goods and services; and for analysing consumption and production patterns in the economy.

The Government is the largest data collecting agency, which collect data on various demographic characteristics of the population (like birth and death rates and, size and composition of the population, etc.) income, consumption,

industrial and agricultural production, etc.

The government and policy makers use statistical data to formulate suitable policies of economic development, and so on.

In business one uses statistical data to study relationship between sales and prices, and to determine fluctuations in the market. Statistical data may also be used to determine the magnitude of inventories to be held at a given point of time.

In industry one may be interested in analysing the relationship between inputs and outputs, etc.

Statistics plays an important role both in experimental and non-experimental sciences. In both cases, the statistical relations between variables are derived from observed data.

In experimental sciences (like physics, chemistry, etc.) one is able to generate his/her own data under controlled laboratory conditions using high precision instruments. Therefore, the data available to the experimental scientist may be relatively more accurate than those available to non-experimental scientist (like an economist). The data available to the economist are those collected and processed (classified and tabulated) by others. In some cases the economist may use primary data collected by himself.

Statistical relations between variables are invariably **inexact** and subject to error. There may be different kinds of errors committed in the process of data collection, classification, tabulation, or, interpretation. We will discuss them in the following chapters. There may also be errors in the statistical

relations due to omitted variable. As an example let us consider the relationship of quantity demanded of a certain commodity and its price. We know that the quantity demanded decreases as its price increases. But, the change in the quantity demanded may also occur, if income changes. Size of the family, tastes of people, etc. also affect the quantity demanded. If we are considering the relationship of quantity demanded with price alone (ignoring the effect of other variables), the relationship is bound to be in error. The error in the relationship is due to omitted variables.

Statistical methods help in measuring statistical relations in the presence of errors of measurement of data and errors in statistical relations due to omitted variables. It is the presence of such errors in statistical relations (both in experimental and non-experimental sciences) that makes the use of statistical methods indispensable. If there were no such errors, use of statistical methods would be redundant and mathematical methods would be adequate.

3. Importance of Statistics in Economics

In economics, we are usually required to measure changes in some variables, if some other variables change. For example, by what proportion would the demand for a certain commodity decrease, if its price increased by one per cent? As another example, by what proportion would the total net private investment in the economy increase, if the Reserve Bank of India declared half per cent cut in the rate of interest? How would the output of a firm change, if the

management decided to employ more labour, or invest more funds in technology?

Intuitively, one can say that the quantity demanded of the commodity would decrease, if its price increases. But the proportion of change in demand would depend upon whether it is a necessary good (like salt, rice, etc.) or a luxury item (like a TV or a refrigerator).

If we can measure the statistical relation of quantity demanded and price, for a given set of data, we can use this to obtain a numerical estimate of the price elasticity of demand. If the numerical value of elasticity (which is defined as the 'relative change in demand for a proportionate change in price') is close to zero, we conclude that the demand would change very little for a small change in price. Otherwise, if the elasticity is high (say more than one) a small change will have considerable effect on demand. We can also use the estimated demand-price relationship to obtain an estimate of quantity demanded corresponding to a given price.

In economics, we have single equation models describing the behaviour of individuals (like buyers, sellers, investors, etc.) These are generally inexact equations (due to omitted variables, besides errors in measurement of variables, as explained above).

Multi-equation models are used to describe the structure of a market (where buyers and sellers interact) and structure of an economy where several economic agents (like consumers, producers, labour, etc.) interact. We have models to describe various sectors of the economy – like models in agriculture, industry and foreign trade.

4. Misuse of Statistical Methods

As it is true for every science, statistical methods are based on certain well defined assumptions. While analysing and interpreting statistical results one should be very careful about the assumptions they are based on. If we ignore them our conclusions will be wrong.

The quality of data plays an important part in statistical analysis. Fallacies may arise, if we are using data which are either insufficient, or unrepresentative, or incomparable. For example, suppose we want to test the efficacy of a certain drug, or a fertilizer; we must test it in a number of cases under varying conditions. If we jump to conclusions by examining the effect in very few cases, we are likely to make mistakes. Similarly, if we want to estimate the average income of people living in a town and base our estimate on income of people living in the rich areas alone we will get an overestimate.

Any vagueness in the definitions and concepts employed in data collection may lead to wrong conclusions. For example, if we are collecting data on the income of households, we must make clear how we intend to calculate the required income. People with fixed wage or salary income may be able to give a correct answer, but a businessman does not have a fixed income. His income varies from month to month, and he may not maintain accurate records of his income regularly. Then, how would we calculate his income? Similarly, suppose we are collecting data on employment. There are many people who may be employed only

for a part of the day or month. How should we measure their employment?

Statistical methods are no substitute for common sense. They should not be used blindly. The following two examples illustrate this point.

An epidemic once broke out in some villages of a certain state. The government took immediate steps to control the epidemic. Medical assistance was sent in terms of medicines and doctors. One of the political leaders, who also claimed to be a statistician, collected data and found that the number of deaths were larger in those villages which had large number of doctors. He, therefore, concluded that doctors were responsible for deaths and they should be punished.

Another interesting story is usually 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. What was the mistake? Does the fault lie with the statistical method of calculating averages, or, with the misuse of the averages?

Many times, people talk of 'statistical fallacies', 'statistical lies', etc. It should be noted that they arise due to wrong handling of data, or, because of ignoring the underlying assumptions, or, by misusing certain summary indices (like averages) — as noted above.

EXERCISES

1. 'Statistics is defined as aggregate of numerical facts'. Give a few examples.
2. 'Statistics is defined as the 'science' which deals with the analysis of statistical data'. Give examples.
3. Distinguish between 'quantitative' and 'qualitative' data, and give some examples of both.
4. What are the kinds of errors that you envisage in the process of collection of data?
5. Explain why the statistical relations are generally inexact.
6. Use of statistical methods is indispensable both in experimental and non-experimental sciences — why?
7. Compare mathematical and statistical relations, and give examples of both.
8. 'Statistical methods are no substitute for common sense' — illustrate.

CHAPTER 2

Collection of Data

If we are planning to study an economic or social problem, we would require data on certain variables. For example, if we want to study, how does the demand for a certain commodity react to change in its price, we would require data on quantity demanded of that commodity and its price, in several markets and at different points of time. If we want to study, what proportion of income does a household spend on food, clothing, schooling of children and house rent, etc. we would require data on household budgets.

We may collect our own data by conducting a market survey, or, an enquiry into household budgets; we may also use data from published sources, or, data made available to us from office records.

1. Primary and Secondary Data

The data collected by conducting a statistical enquiry, or, a field investigation, are called **primary data**. It does not matter who collects the data. What is important is that they are based on **first hand information**. Thus, the data collected by the investigator himself are primary data. Also, the data collected by some other agency (like the National Sample Survey Organisation or the Reserve Bank of India, etc.) but made available to the

investigator in **original form** (on questionnaires or schedules) are primary data.

If the data have been collected and processed (scrutinized, tabulated and represented diagrammatically) by some one other than the investigator, they are called **secondary data**. Thus, the published data are necessarily secondary data. For some important sources of secondary data, see Appendix C at the end of this book.

The primary data may be collected either by

- (i) Census Method or the Method of Complete Enumeration, where all individual units in the region are covered or
- (ii) Sampling Method, where only a fraction of the total population is covered.

2. Collection of Primary Data

The person who plans and conducts the statistical investigation is called the **investigator**. Those who go out to the field to collect actual data are called the **enumerators**. The **respondents** provide the actual data, by answering the questions in the questionnaire.

Planning of the Fieldwork

Whether we adopt the census, or the

sample method for collecting data, the fieldwork must be carefully planned and organised.

(a) *Preparation of the Questionnaire*

Questionnaire is a list of questions, prepared by the investigator, on the subject of inquiry. While preparing the questionnaire the following points should be noted:

- (i) **The questionnaire should not be very long.** A lengthy questionnaire may tire out both the enumerator and the respondent.
- (ii) **The questions to be included in the questionnaire must be precise and short.** Vague, or, ambiguous questions must be avoided; as they would provide inaccurate information.
- (iii) **The questions should be framed such that they can be cross-checked with other questions in the questionnaire.**
- (iv) **The questions should not involve much arithmetical calculations for the enumerator or the respondent.**

(b) *Mode of Enquiry*

The data may be collected, either by

- (i) the interview method, or
- (ii) mailing questionnaire.

Advantages of the Interview Method

The advantages of the interview method are the following :

- (i) the enumerator can personally explain to the respondent the objective of the enquiry and importance of the study,
- (ii) this will help in getting better cooperation of the respondent and in obtaining accurate answers to the questions in the questionnaire,

(iii) the enumerator can help the respondent in interpreting the questions correctly and recording his/her answers.

(iv) this will save time of the respondent and will keep him/her in good humour.

Disadvantages of the Interview Method

This method is expensive. We need a large team of enumerators and spend on their training and travel, besides other expenses on food, stationery, lodging, etc.

Advantages of Mailing Questionnaires

The method of mailing questionnaires to respondents is far more convenient and less expensive.

Disadvantages of Mailing Questionnaires

- (i) The respondent may not understand or may misinterpret some questions.
- (ii) The respondent may not take enough care to answer all questions correctly.
- (iii) The respondent may ignore and not return the questionnaire at all.
- (iv) Some of the questionnaires may be lost in the mail.

The method of mailing questionnaires is suitable when it is compulsory by law to file information required in the questionnaire. For example, government agencies make it compulsory for banks and companies to supply information to the government.

(c) *Training of the Enumerators*

Training programmes for the enumerators have to be arranged, so that they can interpret the questions correctly, explain to the respondents the

of the enquiry and importance of the study. They must be trained to be polite in their presentation.

(d) *Pilot Survey*

In case, it is going to be a large scale field study, it is useful to, initially, conduct a survey on a smaller scale (called a pilot survey) before launching the large survey.

Advantages of Pilot Survey

The pilot survey will help

- (i) to pre-test the suitability of questions to be included in the questionnaire,
- (ii) to avoid any unforeseen problems that might arise during the large scale survey,
- (iii) to assess the performance of the enumerators and remove any difficulties they may be facing,
- (iv) to assess the reactions of respondents to the questions in the questionnaire,
- (v) to assess the time that the actual survey might take,
- (vi) to assess the cost of the large scale survey, and
- (vii) to get a preliminary idea about certain aspects of the data.

3. **The Census Method**

The census method is also called the '**method of complete enumeration**'. The essential feature of this method is that every individual unit in the whole population is to be covered. We do not select some and leave out others.

The Census of India is carried out once in every ten years on this basis. A house-to-house enquiry is carried out covering all households in India.

Demographic data on birth and death rates and, on size and composition of population, etc. are collected and published by the Registrar General. Most recent population census in India was carried out in February, 2001. Census of manufacturing units is also carried out periodically. The census method is also used to obtain an estimate of the total area under principal crops in India. The data are obtained by using village records maintained regularly by the village administrative officer.

The process of conducting a census involves preparing a questionnaire on the basis of the purpose of study, and then sending out a large team of well-trained enumerators to the respondents in the field to get the questionnaires filled. There has to be a fixed time frame for the enquiry.

The data collected by census method will, generally, be voluminous. We must process them into manageable form before drawing any conclusions out of the data. In other words, we must scrutinize the data for any apparent errors, tabulate them suitably and, if necessary, present them diagrammatically.

The following kinds of errors are likely to occur in data collection.

Errors in Data Collection

(i) *Errors of Measurement*

As a very simple example, suppose each student in your class is asked to measure the length of the teacher's table in the classroom. Every student is provided a separate measuring tape. If you compare the measurements taken by the students, you will surely find that the measurements are not identical. The

differences in measurements may arise because some students measured up to the nearest of the unit, while others measured to the nearest of the tenth place of decimal. Rounding off errors will always be there. The differences in measurements may also occur due to differences in measuring tapes themselves, due to manufacturing defects. Some students may also be careless.

In a household survey, suppose we want to know the household expenditure on various items of consumption. If we ask the head of the household about the monthly expenditure on food, we are bound to get only an approximate figure. 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. They also vary according to their quality. Which price do we take? We can, at best, take some kind of an average price.

In experimental sciences (like physics, chemistry, etc.) such errors of measurements will occur while taking readings on various instruments.

(ii) *Errors due to Mishandling of the Questionnaire*

The enumerator, or, the respondent may misunderstand, or, misinterpret some questions of the questionnaire. Since the scale of operations is large in such survey, a large team of enumerators has to be employed. All enumerators may not be equally efficient. Some careless ones may not take their job seriously, even after intensive training. There may also be lack of coordination in the team work.

(iii) *Recording Mistakes*

The enumerator or the respondent may commit errors in recording data; for example, he/she may record 13 instead of 31, and so on. Sometimes the handwriting is so bad and unclear that the tabulator may misread the recorded answers, while transferring data to files or computer.

(iv) *Errors of Non-Response*

These errors arise when the respondent refuses to fill up the questionnaire, or, the respondent is not available even after repeated visits by the enumerator.

The magnitude of the errors of non-response will, generally, be large, if questionnaires are mailed to the respondent and not carried to him/her personally by the enumerator. The respondent, in that case, may not care, or, feel too lazy to return the questionnaire duly filled. In many cases the questionnaires may be lost in the mail.

(v) *Arithmetical Errors*

If some questions, require a little of arithmetical calculations, there is a possibility of such errors. For example, if the question is, 'what was the expenditure on food last month'. The head of the household will have to add expenditures on rice, wheat, salt, sugar, milk, etc. and also on fruits and vegetables. The error may occur in recollecting the items, their prices and also in adding.

The magnitude of the errors mentioned above will tend to be large in a field survey by complete enumeration (or, census) method, because the errors

will tend to cumulate and it may be difficult to provide adequate training, coordinate and supervise the work of large teams of enumerators.

The cost of data collection by census method will be high as a very large team of enumerators has to be trained and their fieldwork has to be coordinated and supervised. It will require lot of expenditure on travel of enumerators, besides other costs on food, stationery, etc.

In many situations, it may not be feasible to carry out a census at all; e.g. suppose we want to estimate the total amount of timber available in a forest, or, the total amount of fish available in a river, or, the total number of birds in a sanctuary, etc.

We can, at best, obtain an estimate in such cases, based on **sample data**.

4. The Sample Method

Suppose we want to find the average income of people in a certain region.

According to the census method, we would be required to find out income of every individual in the region, add them up and divide by the number of individuals in the region. This method would require huge expenditure, as a large number of enumerators have to be employed; and the result would be contaminated by the kinds of errors described above.

Instead, we may adopt the **sample method**. Accordingly, we 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 the average income of individuals in the whole region.

In general, according to the sample

method, we select a **representative sample** of a few individual units from the population and obtain the 'estimate' of the population characteristic on the basis of sample data. For example, we may estimate the yield of wheat in Punjab by obtaining the output of wheat in a few selected fields; we may estimate the volume of timber in a certain forest, by felling a few selected trees in the forest; and so on.

Advantages of the Sample Method

The sample method has the following advantages.

- (i) The cost of the survey would be much smaller than that in complete enumeration.
- (ii) The collection of data, their tabulation and analysis would take much less time.
- (iii) The magnitude of errors described above would be much smaller.
- (iv) We need a smaller team of enumerators and it is easier to intensively train them, closely supervise their fieldwork and guard against probable errors.

When we use the sample method, to draw inferences about population characteristics, another kind of error is introduced. This is called the **sampling error**. The errors discussed above (measurement errors, recording mistakes, etc.) are called **non-sampling errors**.

For example, the **estimate** of the average income of people in a certain region, obtained on the basis of incomes of a smaller set of individuals, included in the representative sample, will not be equal to the true average income of people in the region. **The difference between**

the sample estimate and the true average income in the region is called the sampling error.

It should be noted that several sets of representative samples from the population can be drawn. Each one will give a different estimate of the population characteristic (the average income, variability of incomes, etc.). This is called **fluctuation due to sampling, or sampling fluctuation.**

Example

Consider a hypothetical case, where there are only **five** individual units in the population. The variable x has measurements 10, 15, 20, 25 and 30. We note that the **population average** is

$$\frac{10 + 15 + 20 + 25 + 30}{5} = \frac{100}{5} = 20$$

Now, suppose we want to 'estimate' the population average, on the basis of a representative sample of size two. The number of ways in which we can select a sample of size 2, out of 5 individual units in the population, is $\frac{5 \times 4}{2} = 10$

Therefore, 10 possible samples are as follows.

Sample	Values of x
1.	10, 15
2.	10, 20
3.	10, 25
4.	10, 30
5.	15, 20
6.	15, 25
7.	15, 30
8.	20, 25
9.	20, 30
10.	25, 30

The estimates of the population average, obtained from different samples, and their sampling errors are shown in the following table.

Sample	Estimate	Sampling error
1.	12.5	-7.5
2.	15	-5
3.	17.5	-2.5
4.	20	0
5.	17.5	-2.5
6.	20	0
7.	22.5	2.5
8.	22.5	2.5
9.	25	5
10.	27.5	7.5

Sampling Error of the Estimate = Estimate - True Value, where the true average = 20.

The essential requirement of the sample method is that **the sample must be representative of the population, with regard to the characteristic under consideration.** For example, suppose we want to estimate the average income of people in a certain city. We know that there are exclusive areas in the city, where very rich people live, and there are clusters of hutments of very poor people. There are middle class residents also and so on. Our sample must reflect the diversity of incomes. If our sample includes most people from a particular segment - very rich or very poor - our estimate of the average income would show an upward or downward bias. Similarly, while finding an estimate of yield of wheat in Punjab we should guard against selecting only large or small farms, and so on.

In the following sections we discuss some methods of drawing representative samples.

5. Methods of Drawing a Representative Sample

(a) *Random Sampling*

At the outset, let us warn that random sampling does not mean haphazard sampling, where no rules of selection are followed. In fact, random sampling is based on a purely scientific technique.

Random sampling requires that every individual unit in the population gets an equal chance of being included in the sample.

The selection procedure must guarantee this property. The selection procedures are illustrated below.

Lottery Method

Suppose there are 50 students in your class, and you are required to select a random sample of 5. We may adopt the following procedure:

- (i) prepare 50 slips of paper of identical shapes and size,
- (ii) write the names of students on the slips (one slip for each student),
- (iii) place the prepared slips in a box and thoroughly mix them,
- (iv) finally, without looking in the box, draw five slips.
- (v) the students whose names appear on the slips drawn, constitute the required random sample.

In lotteries, the tickets bought by the people are numbered. They are put in a box and mixed mechanically. Then a required number of tickets are drawn. People holding counterfoils of tickets, bearing the numbers drawn, are the winners.

Limitations of the Lottery Method

The following are the limitations of the

lottery method.

- (i) The lottery method may become cumbersome, if the number of the individual units in the population is too large.
- (ii) It may not be possible to use the lottery method, if the size of the population is infinite.

For example, the lottery method cannot be used, if we want to draw a random sample of 50 birds from a sanctuary, or 50 fish from a river, or, 50 leaves from a *neem* tree. The method may not be appropriate, if we want to draw a random sample of 100 households from a city, or, if we need a random sample of 50 fields growing a particular crop in a state, etc.

Use of Random Numbers

We use random numbers to draw a random sample in many cases. The random numbers have been generated by specific mathematical methods, to guarantee equal probability of selection to every individual unit in the population. The random numbers have been published in the form of books and can also be generated by using appropriate software packages. The important property of random number tables is that we may open the book at any page and start reading numbers, row-wise, or, column-wise, from any point. The numbers so obtained are random numbers. We may use 2-digit, 3-digit, or, 4-digit random numbers, as required. A specimen of random number tables is given in the Appendix A of this book.

Let us illustrate the use of random numbers with the following examples.

Example 1

In order to draw a random sample of 5 students from a class of 50 students, we assign numbers 01, 02, ..., 50 to the students. We use 2-digit random number tables. Suppose we start reading numbers in the tenth row of the random number tables, on the first page (see the Table on page 107 Appendix A). The numbers obtained are 57, 60, 86, 32 and 44. Some of these numbers have not been assigned to any student. Therefore, we frame a rule that whenever the number obtained from the table is greater than 50, we select the student with numbers $57-50=07$, $60-50=10$, $86-50=36$ i.e., the number obtained after subtracting 50 from the number obtained from the table of random numbers. If the number obtained is less than 50 we select the student bearing that number. Alternatively, we could simply skip over the numbers greater than 50 and take the next number.

Instead of assigning numbers 01, ..., 50 we could also use the enrolment numbers of students. If they are 3-digit numbers, we could use 3-digit random number tables.

Example 2

We have to select 5 fields growing wheat in Punjab.

All fields bear a number called '*khasra*

number' in village records. We may use 2-digit or 3-digit random number tables, as required.

(b) Stratified Random Sampling

The method of stratified random sampling requires that:

- (i) we sub-divide the whole population into a number of homogeneous strata, and
- (ii) draw a fixed proportion (say, 1% or $\frac{1}{2}\%$, etc.) of individual units from each stratum, by the random sampling method.

This method is used when the population is not homogeneous with regard to the characteristic under investigation. For example, to estimate the average income of people in a city, we have noted that there are exclusive areas where very rich people live and there are hutments of very poor people. There are areas where middle class people live. The population is not homogeneous with regard to levels of income. We may identify homogeneous strata (areas where very rich, middle class and poor people live). Then select a fixed proportion of individual units from each stratum.

The homogeneity of each stratum, with regard to the characteristic under investigation, is the essential feature of this method. The strata may differ significantly from each other.

EXERCISES

1. Distinguish between 'primary' and 'secondary' data. List at least three sources of secondary data.

2. Statistical tables giving district-wise birth and death rates (number of births and deaths per 1000 of population) are obtained from publications of the *Census of India 2001*. Would you call them primary or secondary data?
3. What are the main steps in the planning of a field survey?
4. What are the kinds of errors you would expect in a field survey by census method?
5. What are the advantages and disadvantages of collecting primary data by (i) personal interview, and (ii) mailing questionnaires to respondents?
6. Define the terms (a) investigator, (b) enumerator, and (c) respondent.
7. Distinguish between (a) sampling, and (b) non-sampling errors.
8. What are (a) measurement errors, and (b) recording mistakes?
9. What are the main sources of error in the collection of data?
10. What are the advantages of sampling method of collection of data over the census method?
11. Define 'random sampling'. How is it different from haphazard sampling?
12. Suppose there are 10 students in your class. You want to select three out of them. How many samples are possible?
13. Discuss how you would use the lottery method to select 3 students out of 10 in your class?
14. Does the lottery method always give you a random sample? Explain.
15. Explain the procedure of selecting a random sample of 3 students out of 10 in your class, by using random number tables.
16. Distinguish between 'random sampling' and 'stratified random sampling', clearly explaining the two methods in detail. Give examples of both.

CHAPTER 3

Organization of Data

Classification of Data

'Classification' means arranging things in appropriate order and putting them into homogeneous groups. For example, in library, the books and periodicals are classified and arranged according to subjects; students are grouped according to division they secure in a certain examination; plants and animals may be assigned to groups distinguished by structure, origin, etc.

Data may be arranged by time, space, or both. For example, we have time series data on aggregate (national) income, aggregate consumption, size of the population, etc.; and we have data on literacy rates for different states and also at different points of time, and so on.

We will discuss methods of grouping data on a single variable in some detail in the following paragraphs. However, to begin with, let us consider certain terms which are commonly used in every day life, but they have specific meanings in statistics.

(i) *Variable and Attributes*

In common language, the term 'variable' means a certain characteristic that changes from one object to the other. For

example, heights of individuals and their looks change, and therefore, they are variables. Similarly, the intelligence of people, prices of commodities and incomes are variables.

However, in statistics, the term 'variable' is used, only if, the changing characteristic can be numerically measured. Thus, heights and weights of individuals are variables, as they can be measured in numerical terms. Prices of commodities vary over time and space, and they can be numerically measured. Therefore, price is a variable. Similarly, incomes of individuals, household expenditures on various items of consumption, size of households, inputs and outputs of firms are all variables.

Although, the looks of people, their intelligence and aptitude for art and music change from one individual to the other, they cannot be measured numerically in the same way as heights and weights or prices and incomes. Therefore, they are not called variables in the statistical sense. They are called 'attributes'. We may rank individuals according to the quality of attributes. The ranks are sometimes used as their numerical values for purpose of analysis.

(ii) *Continuous and Discrete Variables*

A variable is called 'continuous', if it can take any real value in a given range. It may take integral values (whole numbers like 1, 2, 3, ...), fractional values (like $\frac{1}{2}$, $\frac{3}{4}$, $\frac{7}{8}$, ...), or, the values like $\sqrt{2} = 1.414...$, $\sqrt{3} = 1.732...$, $\sqrt{7} = 2.645...$ (which are not exact fractions). In other words, it can take all conceivable real values in the given range. For example, heights and weights of individuals, prices of commodities, income of individuals may be treated as continuous variables. (Although, in practice, the measurements are taken only approximately, up to one or two places of decimal, the true values may be anything in a certain range.)

If the variable can take only integer values (like whole numbers), it is called a 'discrete' or 'discontinuous' variable. For example, number of students in different classes, different schools or the size of households are discrete variables, as they can take integral values only.

(iii) *Population*

In common language, the word 'population' means the number of persons living in a certain region. We may count the number of persons and obtain the size of population of that region. Similarly, we may find the population of certain animals in forests in the country; or the population of certain plants in a garden; and so on. The term population implies 'head count'.

However, in statistics,

the data on any single variable, or, a set of variables, for all individual units in a region, constitute the population of that variable or variables.

If the data are on a single variable, the set of measurements constitutes a **univariate population** of that variable. Otherwise, we have a **bivariate or multivariate population** of the set of variables. For example, we may have a univariate population of prices of a particular commodity prevailing in the markets covering a particular geographical area; a population of incomes of all individuals of a defined region. We may have a bivariate population of heights and weights of all individuals in a region; or a multivariate population of expenditures on various items of consumption of all households.

We will restrict our discussion to a univariate population, and illustrate the construction of a

- (a) Frequency Array, and
- (b) Frequency Distribution

(a) *Frequency Array*

We obtain a **frequency array**, if the variable x is discrete and we have frequencies corresponding to **each** value (there are no class intervals) of the variable. Let us illustrate this with the following example.

Example

A survey of 100 households was carried out to obtain information on their size, i.e. the number of members of households. The results of the survey are classified as a **frequency array** in Table 3.1.

TABLE 3.1

Frequency Array of Size of Households	
Size of the household x	Number of households f
(1)	(2)
1	5
2	15
3	25
4	35
5	10
6	5
7	3
8	2
Total	100

The Column (1) of Table 3.1 gives the values which the variable x (size of the household) takes; and Column (2) gives the corresponding frequencies (number of households). Thus, there are 5 households each of whose size is 1, there are 15 households each of size 2, and so on.

Table 3.1 gives the frequency array of size of households.

(b) Frequency Distribution

Suppose the largest value of x is B and smallest value is A . Then $R = B - A$ is the total **range** of x . A large range indicates that the values of x are spread over a large interval, or the variation of values of x is large. A small range indicates smaller variation in the values of x . Thus, the range is a measure of variation (or dispersion) of x .

However, the range is a rather **crude** measure of variation of x . It does not say anything about the distribution of values of x within the range. Are the values of x uniformly distributed within the range, as in the Fig. 3.1;

or they are clustered about some value close to the upper or lower ends or the middle of the range, as shown in Figs. 3.2, 3.3 and 3.4?

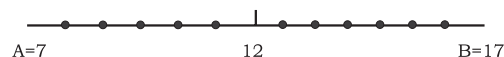


Fig.3.1 Values are uniformly distributed over the range

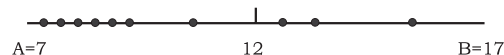


Fig.3.2 Values concentrated in the lower part of the range

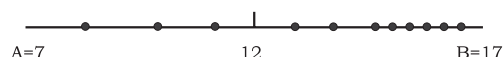


Fig.3.3 Values are concentrated in the upper part of the range

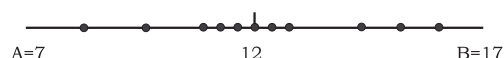


Fig.3.4 Values are concentrated in the middle of the range

For example, suppose we are considering the distribution of marks obtained by 1 lakh students in mathematics in a certain examination. The maximum of marks obtained is 100, and the minimum is zero. So the range is $R = 100$. It is possible that 70 per cent of students got marks more than 60, and 20 per cent got less than 40 marks. In another case, 70 per cent of the students got marks between 40 and 60, etc.

As another example, suppose we have data on monthly incomes of 10,000 individuals, the maximum of which is Rs 50,000 and minimum is Rs 1,000. Thus, the range is Rs 49,000. We observe that majority of individuals (say, 70 per cent) have small incomes close to

Rs 5,000 and very few (say 2 per cent) have income close to Rs 30,000.

In order to get a better idea about the distribution of values within the range, we should sub-divide the total range into a number of class intervals and find out the number of values in different classes.

The number of values in a particular class is called frequency in that class.

This leads us to the construction of a frequency distribution.

While constructing a frequency distribution one should pay attention to the following points.

- (i) How many classes should we have?
- (ii) What should be the size of each class?
- (iii) How should we choose the class limits?

Number of Classes

In fact, there is no hard and fast rule about this; but as a working rule the number of classes should lie between 5 and 15. It should be noted that the number of classes will be large, if we choose small size class intervals and it will be small, if the size of class intervals is large.

As an illustration, suppose the range is 70, and we choose classes of width 2 each. We would require $70 \div 2 = 35$ classes. However, the number of classes would be 14, if the width of each class was 5.

A frequency table with more than 15 classes will be rather bulky and hence not desirable. It does not serve the purpose of classifying data into class intervals, if we choose less than 5 classes.

Size of Class Intervals

We may choose all classes of the same width or of different widths. However, in the present text we restrict to equal class intervals only. In this case (of equal class intervals), the size of the class interval is determined as soon as we have decided about the number of classes.

Suppose n is the number of classes and all classes are of width h , then $n \times h = R$. Knowing the range R and number of classes n we can obtain $h = \frac{R}{n}$ as the width of the class interval. If the range is 70 and we choose 10 classes, the width is 7.

We require that the values of the variable are uniformly distributed within any class interval.

In that case, we can safely assume that all values are equal to the middle value of the class interval. For example, if a class interval is 10-20, all values in this class may be considered equal to the middle value 15 (obtained as $\frac{20 + 10}{2} = 15$).

Using this assumption, the error committed will be small, as the positive and negative errors will tend to cancel out. However, the error will be large, if the values are not uniformly distributed within the class interval.

Choice of Class Limits

Suppose x is a **continuous** variable, such that it can take any value in a given range. In that case, it is possible to choose class limits which are not equal to any of the observed values. For example, height of individuals is a continuous variable, even though, in practice, one can

measure height to the nearest of the unit value (in centimetres) as 165, 170, 169, 171, ...; or to the nearest of tenth place of decimals as 165.3, 170.4, 168.9, 170.8, We may specify class intervals as

160.55 – 165.55, 165.55 – 175.55, ..., so that none of the observed values of x is equal to any of the class limits.

However, if x is a **discrete** variable which takes only integral values, it will not make any sense in choosing class limits other than the integral values. For example, suppose x is the size of household. We may observe $x = 5, 4, 8, 3, 7, \dots$ in a survey. In this case we may adopt one of the following methods.

Inclusive Method

We choose the class intervals such that both the upper and lower limits are part of the interval. In other words, **the class limits are inclusive**. For example, suppose we choose the class intervals as 1-5, 6-10, 11-15, 16-20, etc. and we find there is a household whose size is 5. This household will be included in the first class 1-5 as both limits are inclusive. Similarly, a household with size 6 will be included in the second class 6-10, and so on.

Exclusive Method

Either the upper or lower limit is excluded from the class interval. Suppose the class intervals are chosen as 1-5, 5-9, 9-13, 13-17, etc. and we specify that the upper limit is excluded and lower limit is included in the class interval. Now, a household whose size is 5 will be included in the second class interval 5-9; and so on.

We could, as well, specify that the

upper limit is included and lower limit is excluded. In that case, the household with size 5 will be included in the first class 1-5; and so on.

(c) *Construction of a Frequency Array and a Frequency Distribution*

Illustration 1: Frequency Array

Twenty students in your class decided to evaluate the performance of the teacher. Rating was done on a scale of 1 to 5; where the rating “1” indicates the best performance and the rating “5” indicates poorest performance. The results are shown as a frequency array given below.

Rating(x)	1	2	3	4	5
Number of students(f)	4	6	7	2	1

In this case, the variable (x) is ‘rating’. It is a discrete variable which takes values 1,2,3,4 and 5. The number of students represents the frequencies. There are 4 students who gave the rating ‘1’ to the teacher; there are 6 who gave the rating ‘2’; and so on. The total frequency is 20.

The counting of frequencies can be done more conveniently, if we use “tally marks” or “tally bars”. Let us illustrate this as follows.

TABLE 3.2
Frequency Array of Ratings

Rating x	Tally Marks	Frequency f
1	////	4
2	//// /	6
3	//// //	7
4	//	2
5	/	1
Total		20

We put a single tally mark (/) for each student. There are 4 students who gave the rating '1', therefore, four tally marks are placed against $x = 1$; 6 tally marks are placed against $x = 2$, and so on. For convenience, the fifth tally is shown across the earlier four. This helps in counting.

Illustration 2: Frequency Distribution

In Table 3.3, we are given the percentage of marks in mathematics obtained by 100 students in a certain examination.

Construct a frequency distribution for the data in Table 3.3.

As noted before, there are several alternative ways of choosing the number of classes, their size and the class limits.

Suppose, we choose 10 classes as 0-10, 10-20, ..., 90-100. All class intervals have the same width 10; and the upper class limits are equal to the lower limits of successive class intervals. Let us use the working rule that **the upper limits of class intervals are excluded but lower limits are included**. Thus, if a student

gets marks equal to the upper limit of some class interval, we classify him/her in the next class.

The counting of frequencies is done by placing tally marks against various class intervals. A tally, is put against a class for each student whose marks lie in this class. For example, if the marks obtained by a student are 57, we put a tally (/) against the class interval 50-60; if the marks are 71 a tally is put against 70-80, and so on. If someone got 40 marks, the tally is put against 40-50 as per rule.

It helps in counting the tally marks, if four of them are put as (////) and the fifth is shown by placing the tally across them, as (~~////~~). Thus, if there are 16 tallies in a class, we put them as (~~////~~ ~~////~~ /), etc.

The frequency in any class is equal to the number of tally marks against this class.

The frequency distribution is shown in Table 3.4.

TABLE 3.3

Percentage of Marks in Mathematics Obtained by 100 Students in a Certain Examination

47	45	10	60	51	56	66	96	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	09	56	14	22
66	53	46	70	43	61	59	12	30	35
45	44	57	76	82	39	32	14	90	25

TABLE 3.4

Frequency Distribution of the Percentage of Marks in Mathematics Obtained by 100 Students in a Certain Exam

Percentage of marks x Class interval	Tally marks	Frequency f
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

Illustration 3: Frequency Distribution

The data on daily wage earnings (in rupees) of 40 individuals are given below.

200	120	350	550	400	140	350	85
180	110	110	600	350	500	450	200
170	90	170	800	190	700	630	170
210	185	250	120	180	350	110	250
430	140	300	400	200	400	210	300

We may treat the variable 'daily wage earnings' as continuous. Observe that the maximum wage is Rs 800 and minimum is Rs 85. Thus, the range is $R = 715$. Let us choose the width of class intervals as $h=100$, and specify the classes as

70.5 - 170.5, 170.5 - 270.5, ..., 770.5 - 870.5

The frequency distribution is shown in Table 3.5.

**TABLE 3.5
Frequency Distribution of Daily Wage Earnings**

Wage (rupees) x	Tally marks	Frequency f
70.5-170.5	//// //// //	12
170.5-270.5	//// //// /	11
270.5-370.5	//// /	6
370.5-470.5	////	5
470.5-570.5	//	2
570.5-670.5	//	2
670.5-770.5	/	1
770.5-870.5	/	1
Total		40

(d) Error of Grouping

Once the data have been grouped into class intervals, we obtain the frequency distribution.

The frequency distribution gives the number of observations (frequencies) in different classes, **but not their actual values**. All values, in any class are assumed to be equal to the middle value of the class interval. This leads to the **error of grouping**.

For example, in Table 3.5, there are 12 values in the class 70.5 to 170.5, viz.

120	140	85	110	110	170
90	170	170	120	110	140

We assume them, all, to be equal to the middle value $\frac{1}{2}(70.5 + 170.5) = 120.5$ of the class interval. Thus, the error of grouping in each case is

-0.5	19.5	-35.5	-10.5	-10.5	49.5
-30.5	49.5	49.5	-0.5	-10.5	19.5

In Table 3.6, we give the actual values (wage earnings) in different classes.

TABLE 3.6
Actual Wage Earnings in Different Classes

Wages x	Actual wage earnings	Frequency
70.5-170.5	120, 140, 85, 110, 110, 170, 90, 170, 170, 120, 110, 140	12
170.5-270.5	200, 180, 200, 190, 210, 185, 250, 180, 250, 200, 210	11
270.5-370.5	350, 350, 350, 350, 300, 300	6
370.5-470.5	400, 450, 430, 400, 400	5
470.5-570.5	550, 500	2
570.5-670.5	600, 630	2
670.5-770.5	700	1
770.5-870.5	800	1
Total		40

We should note that the total error of grouping, in any class, will be small, if the values are uniformly distributed within the class and the class interval is not too large. In that case, the positive and negative errors will tend to cancel out. Keeping this in mind, the student may examine the groupings in Table 3.5 and Table 3.6 and if he/she finds them unsatisfactory, suggest alternative groupings which can be better justified. Would you choose equal or unequal class intervals in that case?

Finally, there is no error of grouping in a frequency array, as there are no class intervals involved.

EXERCISES

1. What do you understand by 'classification of data'?
2. Distinguish between (a) variable, and (b) attribute. Explain with examples.
3. Distinguish between (a) continuous, and (b) discrete variables. Explain with examples.
4. Explain the term 'population' as it is used in statistics. Define (a) univariate, (b) bivariate, and (c) multivariate population. Give examples.
5. What is a frequency distribution? What are the main points underlying the construction of a frequency distribution?
6. How many classes should we choose? How does one decide about the size of class intervals?
7. Distinguish between a frequency array and a frequency distribution.
8. Explain the 'exclusive' and 'inclusive' methods used in classification of data.
9. In an entrance examination, for admission to the Engineering course of a certain university, the top 50 students, who were selected, got the following percentage of marks in the aggregate:

95	92	91	90	88
82	86	87	87	88
70	72	76	77	77
79	79	79	78	78
68	67	67	66	65

63	63	63	63	64
58	59	60	60	60
60	61	62	62	62
55	55	55	56	57
54	53	52	52	50

- (i) Obtain the range of the distribution of percentage of marks.
- (ii) Classify the data in class intervals as
 - (a) 45 – 55, 55 – 65, 65 – 75, ...
 - (b) 50 – 55, 55 – 60, 60 – 65, ...
- (iii) Examine the error of grouping in each case.

10. The following data relate to monthly household expenditure in rupees of 50 households.

904	1559	3473	735	2760
2041	612	753	1855	4439
5090	1085	823	2346	1523
211	1360	1110	2152	1183
1218	1315	1105	628	2712
4248	812	264	1183	1171
1007	1180	953	137	2048
2025	1583	1324	9621	9676
1397	832	962	2177	2575
1293	365	1146	5222	1396

- (a) Obtain the range of the distribution of monthly household expenditure.
 - (b) Divide the range into appropriate number of class intervals and obtain the frequency distribution of expenditure.
 - (c) Do you think that equal or unequal class intervals would be appropriate? Give reasons for your answers.
 - (d) Find the number of households whose monthly expenditure is
 - (i) less than Rs 1,500
 - (ii) more than Rs 6,000.
- 11.** Pick up any printed page of the text of this book. By choosing appropriate class intervals, find the frequency distribution of 'number of alphabets' used in different words. You may ignore any numerical values.

CHAPTER 4

Presentation of Data

1. Descriptive Form

If the number of observations is small, we may simply describe them in words. For example, the *Times of India* of 25 October, 2001, p.3, reported that “The National Institute of Communicable Diseases received 12 cases of anthrax scare ..., pushing the total number of cases received to 93. Of these 12 cases, 6 were from Delhi, 4 from Ghaziabad and 2 from Bulandshahar. So far 52 cases have been examined ... and all of them have tested negative ...’. Another report published in the same newspaper on 26 October, 2001, p.3, was “The traffic police have challaned 16,891 autorikshaw drivers for over-charging till 15 October, this year. The figure is nine times higher than that for last year’s 1,877 ... The number of autorikshaw drivers prosecuted for refusing to take a passenger has also risen by over three times. Till 15 October, 19,960 drivers had been prosecuted for refusal of service. Last year, only 5,805 were prosecuted”.

Generally, the data collected by census or sampling method tend to be voluminous. In order to be able to draw any conclusions out of them, about some characteristics of the population, we must condense them into manageable form. We may

- (i) **tabulate data or**
- (ii) **represent them diagrammatically**

2. Tabulation of Data

The presentation of data in rows and columns, along with some explanatory notes, forms a table. As a very simple example, the data collected on the size of 255 households in a city, are tabulated as follows.

TABLE 4.1
Distribution of Size of Households

Household Size	2	3	4	5	6	7	8	9	Total
No. of households	5	25	40	65	50	35	20	15	255

We may draw the following conclusions from this table.

- (i) The largest size of the household is 9 and smallest is 2. Thus, the size of household varies from 2 to 9.
- (ii) About three-fourths of the households (i.e. 75 per cent of them) are of size larger than 3 and smaller than 8.
- (iii) The majority (65) of households consists of 5 members.

We should note that it would not have

been easy to derive even such simple conclusions from raw data.

(a) *Parts of a Table*

A table has the following parts.

(i) *Table Number*

The table number is given either, on top, or in the beginning of the title. It is used for identification of the table and is generally numerical.

Ordinarily, the table number is given in terms of whole numbers. However, sometimes, subscripted numbers like 2.1 and 3.1 are also used. In that case, the first digit refers to the chapter, or, the section, where the table appears and the second digit to its order. For instance, Table 2.1 would mean the first table in the second chapter (or, the section). Similarly, Table 3.2 would mean the second table in the third chapter (or, the section).

(ii) *Title*

The title briefly describes the contents of the table. It should be clearly worded and so formulated as to be able to describe the important features of the table. It is placed at the head of the table, either together with the table number or just below it. As an example look at the Table 4.1.

(iii) *Column Headings*

Column headings provide brief descriptions in the form of short captions like 'population', 'gross national product', etc.

Units of measurements of data in a column may be given in brackets either

along with or below the caption. In case units of measurements are the same for all the figures in different columns, we do not have to put them separately for each column. Instead, we may put them just once in the margin below the title of the table. Column numbers are indicated below the column headings.

(iv) *Row Headings*

Each row of the table also has a heading. For example, in Table 4.2, 'the states' indicated at the extreme left of each row constitute the row headings.

A brief description of the row headings (e.g. the caption 'major Indian states' in Table 4.2) is given at the upper left hand corner of the table.

(v) *Body of the Table*

This is the main part of the table that contains data. We can identify any figure by considering the row and column in which it lies. For example, in Table 4.2, we obtain the literacy rate (%) of 'males' in AP as 55 per cent, and for 'females' in West Bengal as 46 per cent.

(vi) *Footnotes*

Footnotes are given at the bottom of the table to explain some specific features of the data. In case they relate to some particular figures in the table, then those figures are superscripted by marks like (*) or numbers. Sometimes the purpose of a footnote may be just to explain some important feature of the whole table.

(vii) *Source Note*

It describes the source of the data in the table. It is given below the table.

TABLE 4.2		Table Number	
Literacy Rates by Sex of Major Indian States			Title
Major Indian States*	Literacy rates (%)#		
	Total	Male	Female
(1)	(2)	(3)	(4)
Andhra Pradesh (AP)	44	55	33
Assam (AS)	53	62	43
Bihar (BR)	39	53	23
Gujarat (GJ)	61	73	49
Haryana (HR)	56	69	41
Karnataka (KA)	56	67	44
Kerala (KE)	90	94	86
Madhya Pradesh (MP)	44	58	29
Maharashtra (MR)	65	77	52
Orissa (OR)	49	63	35
Punjab (PB)	59	66	50
Rajasthan (RJ)	39	55	20
Tamil Nadu (TN)	63	74	51
Uttar Pradesh (UP)	42	56	25
West Bengal (WB)	58	68	47
All India	52	64	39
* Major states include those with population of one crore or more as per the 1991 Census.			
# Relates to population aged 7 years and more, but excluding Jammu and Kashmir			
Source: Census of India 1991.			Source Note

(b) *Construction of a Table*

A table should generally be self explanatory. While constructing a table, the following points should be kept in mind.

- (i) **The table is compact, concise and readable at a glance.** It should not be bulky and cumbersome. Otherwise, it

defeats the basic purpose of putting data in manageable form.

- (ii) **The table should facilitate comparison of data.** The figures to be compared should be placed in adjacent rows or columns.
- (iii) **The figures to be emphasized should be printed in bold letters, or be encircled, etc.**

- (iv) **The large numbers should be approximated (e.g. up to nearest crores or lakhs, etc.).** Large numbers are hard to read and difficult to compare.
- (v) **An entry that is 'zero' should be clearly distinguished from the one which is not available.**

3. Diagrammatic Form

whereas a diagram may not be as exact as a table, it is useful because it makes visual comparisons easier. It facilitates

- (i) developing a quick idea about the relative magnitudes for comparison purposes, and
- (ii) grasping the salient features of the data.

While constructing a diagram we should note the following points.

- (i) **The diagram should bear a number and heading.** These are, generally, put either at the top of the diagram or just below the diagram. The diagram number is used for identification and for purposes of referencing; and the heading should briefly describe the contents of the diagram.
- (ii) **Both the x and y-axes must be clearly labelled.** It should be clear from the diagram as to what is being measured on the two axes. For example, we may measure 'population' or, 'gross national product' on the y-axis and 'years' on the x-axis.
- (iii) **Units of measurements of the variables should be stated along with the variable.** For example, if population is measured on the y-axis, we must state 'population (lakhs or crores of persons)' on the y-axis and state 'years' on the x-axis.

- (iv) **The scale of measurement on both axes should be stated at the top right hand corner of the diagram or just below the diagram.** For example, 1 cm = 1 lakh or 1 crore of population, and so on.

- (v) **The choice of origin should be clear from the diagram.** If there are very large numbers involved (like GNP in crores of rupees or population in lakhs or crores of persons, etc.) it may be convenient to choose the origin say, 50,000 or 5 lakhs, etc. This helps in visual comparison.

There are various kinds of diagrams in common use. Some of them are the following:

- (i) **Geometric Forms** — including
 - (a) bar diagram,
 - (b) multiple bar diagram
 - (c) pie diagram
- (ii) **Frequency Diagrams** — including
 - (a) histogram
 - (b) frequency polygon
 - (c) frequency curve
 - (d) ogive
- (iii) **Arithmetic line-graphs (Time Series Graph)**

4. Geometric Forms of Diagrams

- (a) *Bar Diagram*

A bar diagram consists of a set of bars drawn vertically (or horizontally) over time, space or groups. The height (or, the length) of the bar indicates the magnitude of data. The thickness of bars does not matter, but all bars should have the same thickness and bars should be placed at equal distances.

The bar diagram is useful for visual comparison of data over time, space or groups.

For illustration, in Fig.4.1, we use the data in Column (2) of Table 4.2 to show the literacy rates (per cent) of 15 major Indian states in the form of a bar diagram. The states are represented on the x-axis and literacy rates are measured (1 cm = 10 per cent) along the y-axis. Actual figures, for each state, are mentioned at the top of the corresponding bar.

Figure 4.1 clearly shows that the literacy rate in Kerala is way above any other state. Maharashtra and Tamil Nadu are next in order, although they have significantly lower literacy rates. Bihar and Rajasthan have the lowest rates.

(b) Multiple Bar Diagram

These diagrams display bars for more than one variable. They help in comparing different variables at the same time. In order to distinguish between



different variables we may use different colours or shades in the bars related to different variables.

In case of multiple bar diagrams, the bars should either be properly labelled, or a key index describing the bars should be given separately in a corner of the diagram.

As an illustration, we show in Fig. 4.2 the imports and exports of India for the data given below in Table 4.3.

The years are shown on the x-axis and values are measured on the y-axis (1cm = 20,000 crores of rupees).

The height of the bar indicates the value of imports/exports as in Table 4.3. Actual figures are noted at the top of each bar.

The bar shown as  indicates imports and the other  shows exports.

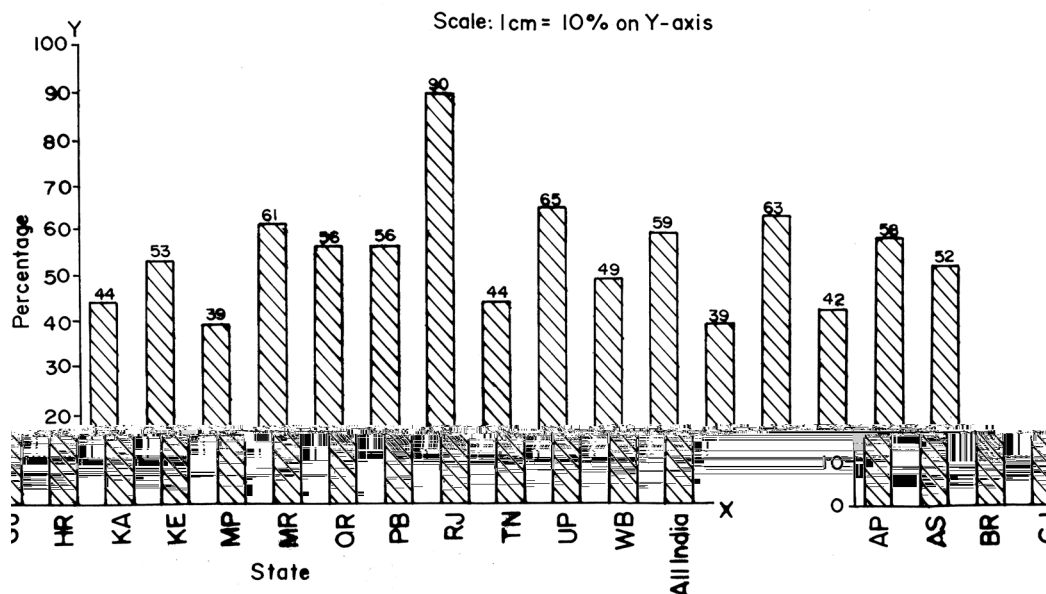


Fig.4.1 Literacy rates

TABLE 4.3
Imports and Exports of India

Year	Imports in current prices (rupees in thousand crores)	Exports in current prices (rupees in thousand crores)
(1)	(2)	(3)
1996-97	139	119
1997-98	154	130
1998-99	176(p)	142(p)
1999-00	149*	119*

* April - December, 1999

(p) provisional

Source: Economic Survey, 1999-2000

TABLE 4.4
Index Numbers of Agricultural Production
(Base 1981-82 = 100 and figures to the
nearest of unit value)

Years	Rice	Foodgrains Wheat	Pulses
(1)	(2)	(3)	(4)
1990-91	149	157	141
1993-94	162	170	131
1994-95	165	187	139
1995-96	155	176	121
1996-97	164	197	140
1997-98	166	189	127
1998-99	173	201	145

As another illustration, Fig. 4.3 shows a multiple bar diagram of index numbers of agricultural production for data given in Table 4.4.

In Fig.4.3, the years are shown on the x-axis. There are three vertical bars (for rice, wheat and pulses) drawn in different shades for each year. The height of each

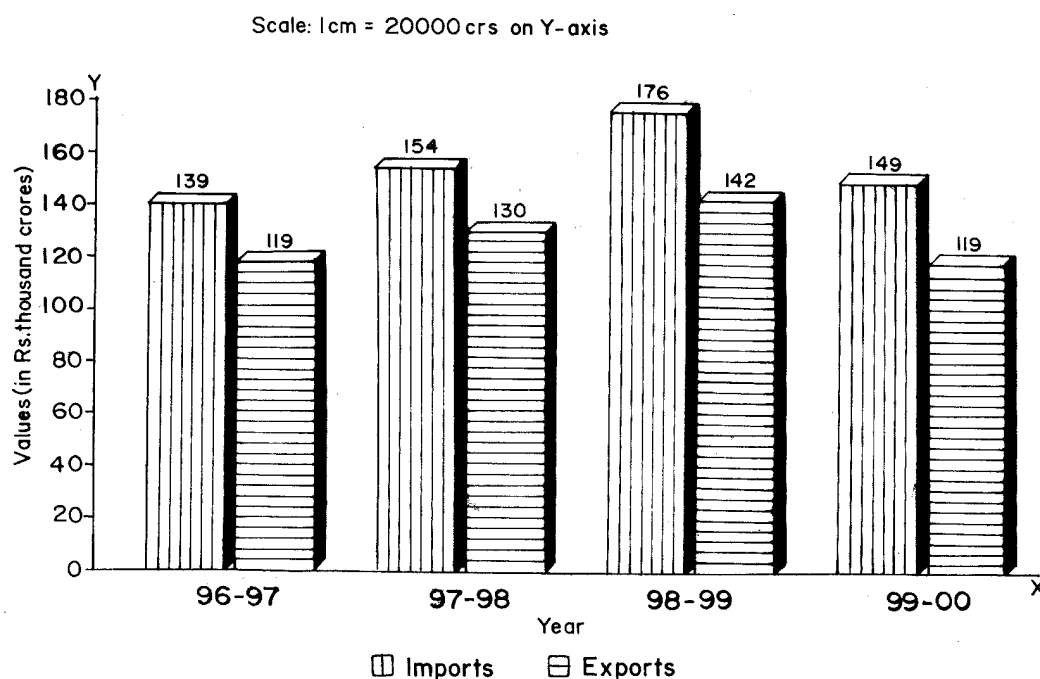


Fig.4.2 Imports and exports of India

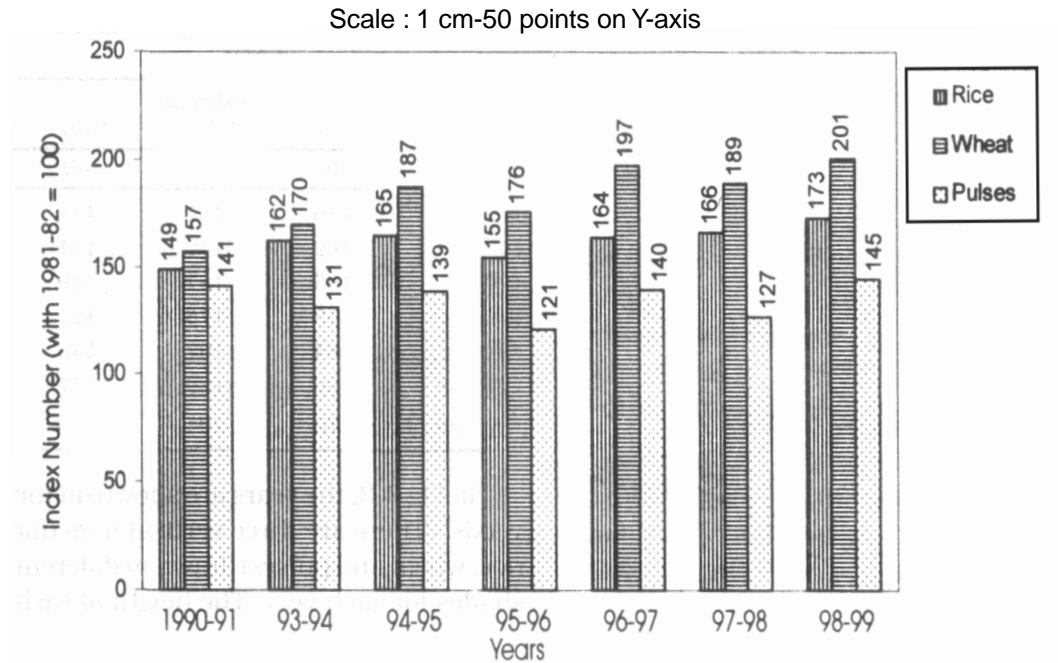


Fig.4.3 Index number of agricultural production

bar indicates the index number shown at the top of the bar. Width of each bar is the same. Scale of measurement on the y-axis is 1 cm = 50 points. (For convenience, the index numbers have been given to the nearest of their unit values.)

Broken Bar Diagram

Sometimes, the range of the variable may be too large to accommodate the whole bar corresponding to the largest value within the graph sheet. In that case, we make the bars in pieces, with each new piece starting with some jump on the numerical scale. The bar is labelled for the actual amount.

As an exercise, one may draw a bar diagram for 'plan expenditures' of successive Five Year Plans beginning with the First Five Year Plan. Also, one may

draw a double (or, multiple) bar diagram for Plan and Non-Plan expenditures of successive Five Year Plans.

Component Bar Diagram

Sometimes, the bars are broken into different components of the same variable. Each component should be shown with different colour/shade/design for identification.

For example, the performance of students in a certain examination is shown in the table below.

Year	Number	1 st Div.	2 nd Div.	3 rd Div.	Failed
1995	85	12	27	32	14
1996	105	18	47	35	5
1997	97	17	40	37	3
1998	112	20	45	40	7
1999	117	25	48	37	7

The component bar diagram is shown in Fig.4.4.

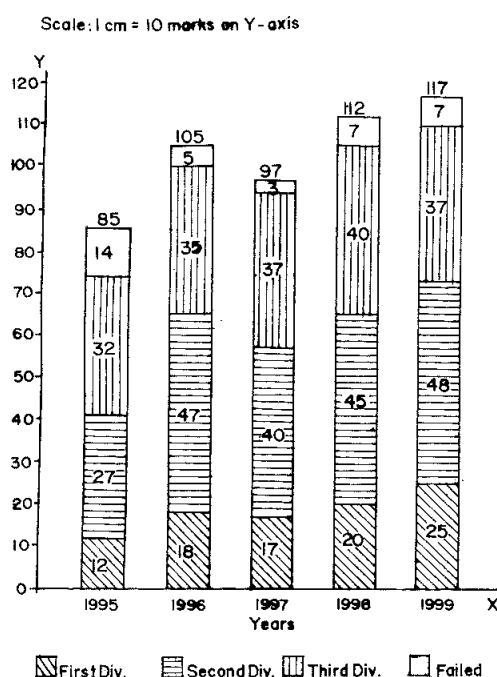


Fig.4.4 Component bar diagram

(c) Pie Diagram

It is also called 'pie chart'. The pie diagram is used to display percentage distribution of some aggregates in various categories. A circle is divided into number of segments according to given percentages. This is achieved by sub-dividing the total angle at the centre in specified proportions. The total angle at the centre is 360° or 2π . Therefore, the diagram is called a 'pie diagram'.

As an example, we consider the percentage distribution of imports of

India in 1998-99 according to a few commodity groups. The data are given in Table 4.5 and the pie diagram is shown in Fig.4.5. Since the percentage in group A is 6.1, the corresponding angle subtended at the centre is

$$\frac{6.1 \times 360}{100} = 21.96^\circ$$

which is approximated as 22° . Similarly, we work out the angles subtended at the centre for other groups.

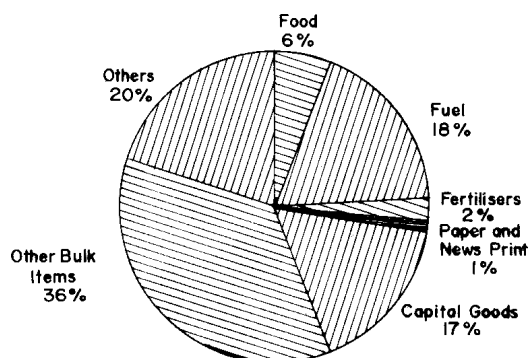
TABLE 4.5
Percentage Distribution of Imports of India
by a Few Commodity Groups (1998-99)

Commodity groups		Per cent	Angle at the centre
(1)	(2)	(3)	(4)
A	Food and allied products	6.1	22°
B	Fuel	17.6	63°
C	Fertilizers	2.3	8°
D	Paper/board manu-		
	factured and newsprint	1.1	4°
E	Capital goods	16.6	60°
F	Other bulk items	35.9	129°
G	Others	20.4	74°
Total		100.0	360°

Total Value of Imports = US \$ 4,18,579 lakh

Source: Economic Survey, 1999-2000.

As another illustration, let us consider the percentage distribution of exports of India, in 1998-99, according to a few commodity groups. The data are given in Table 4.6 and pie diagram is shown in Fig.4.6.



Note: Percentage is of the total value of imports

Fig. 4.5 Percentage distribution of major imports of India (1998-99)

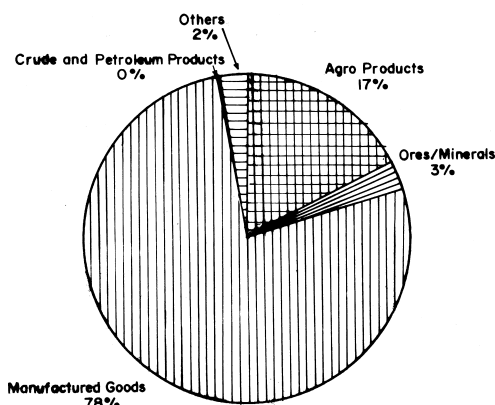


Fig. 4.6 Percentage distribution of major exports of India (1998-99)

TABLE 4.6
Percentage Distribution of Exports of India by a Few Commodity Groups (1998-99)

Commodity groups		Per cent	Angle at the centre
(1)	(2)	(3)	(4)
A	Agriculture and allied products	17.3	62°
B	Ores and minerals	2.6	9°
C	Manufactured goods	77.8	280°
D	Crude and petroleum products	0.3	1°
E	Others	2.0	8°
Total		100.0	360°

Total Value of Exports = US \$ 3,36,585 lakh

Source: Economic Survey, 1999-2000

5. Arithmetic Line-Graphs (Time Series Graph)

In time series graph we choose time unit (year, month, etc.) on the x-axis and the value of the variable is measured on the y-axis.

Table 4.7 gives index number of wholesale prices (base 1981-82 = 100) for the years 1986-87 to 1998-99. The

time series graph is shown in Fig. 4.7. The graph clearly shows a rising trend in wholesale prices.

TABLE 4.7
Index Number of Wholesale Prices (Base 1981-82 = 100)

Last week of	Index Number
1986-87	134
1987-88	149
1988-89	157
1989-90	171
1990-91	192
1991-92	218
1992-93	233
1993-94	258
1994-95	285
1995-96	300
1996-97	320
1997-98	337
1998-99	353

Source: Economic Survey, 1999-2000

As another example let us use the data, in Table 4.8, on the value (in 100 crores of rupees) of exports and imports of India from 1977-78 to 1998-99. The time series graphs for both exports and imports are shown in the same Fig. 4.8.

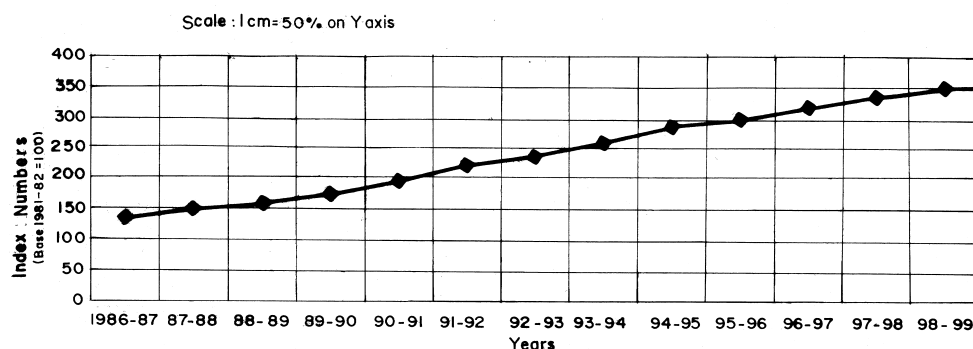


Fig.4.7 Index number of wholesale prices

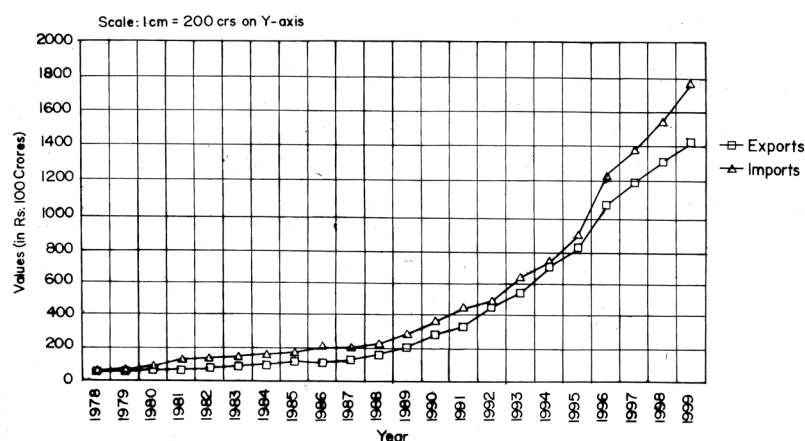


Fig.4.8 Exports and imports of India

We must clearly distinguish the two by using different colours, or, by drawing one in continuous (or bold) and the other in broken lines.

TABLE 4.8
Value of Exports and Imports of India
(in 100 crores of rupees)

Year	Exports	Imports
1977-78	54	60
1978-79	57	68
1979-80	64	91
1980-81	67	125
1981-82	78	136
1982-83	88	143

Year	Exports	Imports
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

EXERCISES

- Present the following information in tabular form:
 "According to the official statistics the incidence of crime (in a city) under all heads except murders increased during this year (2001) up to 15 December as compared to last year (2000)."
 'Sixty-six cases of dacoity were reported this year as against 19 in the corresponding period last year. The number of attempted murders rose to 256 from 200, while robbery cases rose to 636 against 324 last year. Burglaries, motor vehicles thefts and cycle thefts also increased to 3283, 2759, 5889, respectively as against 2527, 1965, 5129 last year'.
 The total number of cases registered this year was 51,809 as against 40,246 last year.
 Compare the 'descriptive' and 'tabular' methods of presentation of data.
- In Exercise 1, suppose you want to emphasize the increase in the number of murders and robberies during this period. How would you do that in the tabular form?
- The Indian Sugar Mills Association reported that, 'Sugar production during the first fortnight of December, 2001 was about 3,87,000 tons, as against 3,78,000 tons 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 tons for internal consumption and 41,000 tons for exports as against 1,54,000 tons for internal consumption and nil for exports during the same fortnight last season.'
 - Present the data in tabular form.
 - Suppose you were to present these data in diagrammatic form, which of the diagrams would you use and why?
 - Present these data diagrammatically.
- What are the requisites of a good table?
 - 'Diagrams are less accurate but more effective than tables in presenting the data'. Explain.
- What kind of diagrams are more effective in representing the following?
 - Monthly rainfall in a year
 - Composition of the population of Delhi by religion
 - Components of cost in a factory
- The following table gives absolute values (in lakhs of tons) of foodgrains production in India:

<i>Year</i>	<i>Foodgrains production (in lakhs of tons)</i>
1996-97	1994
1997-98	1923
1998-99	2030
1999-00	1091

Represent the data by an appropriate bar diagram.

7. The following table gives the data on electricity generated (in billions KWH) in India:

<i>Year</i>	<i>Electricity generated (in billions KWH)</i>
1996-97	394.5
1997-98	420.6
1998-99	448.4
April-December, 1999	355.3

Represent the data by an appropriate bar diagram.

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

<i>Year</i>	<i>Agriculture and allied sectors</i>	<i>Industry</i>	<i>Services</i>
<i>(1)</i>	<i>(2)</i>	<i>(3)</i>	<i>(4)</i>
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, clearly stating the scale of measurement.

9. The direction of India's exports (in percentage terms) in 1998-99 is shown in the following table.

	<i>Destination</i>	<i>Per cent</i>
<i>(1)</i>	<i>(2)</i>	<i>(3)</i>
A	USA	21.8
B	Germany	5.6
C	Other EU	14.7
D	UK	5.7
E	Japan	4.9
F	Russia	2.1
G	Other East Europe	0.6
H	Opec	10.5
I	Asia	19.0
J	Other LDC's	5.6
K	Others	9.5

Total Exports: US \$ 33658.5 million

Draw a pie chart to show the percentage distribution of exports to various destinations.

- 10.** The sources of imports of India (in percentage terms) in 1998-99 are shown in the following table.

<i>Sources</i>		<i>Per cent</i>
(1)	(2)	(3)
A	USA	8.7
B	Germany	5.1
C	Other EU	12.4
D	UK	6.1
E	Japan	5.7
F	Russia	1.3
G	Other East Europe	0.4
H	Opec	18.7
I	Asia	15.7
J	Other LDC's	5.6
K	Others	20.3

Total Imports: US \$ 41857 million

Draw a pie diagram to show the percentage distribution of imports from various sources.

CHAPTER 5

Frequency Curves and Diagrams

1. Graphical Presentation of Data

It is always convenient to use graph paper for drawing a curve or a diagram. The position of any point on the graph paper can be described with reference to two straight lines intersecting each other at right angles. Conventionally, one of them is drawn horizontally and the other vertically. These reference lines are called the 'axes' and the point of their intersection is called the 'origin'.

As shown in Fig. 5.1, the two axes are

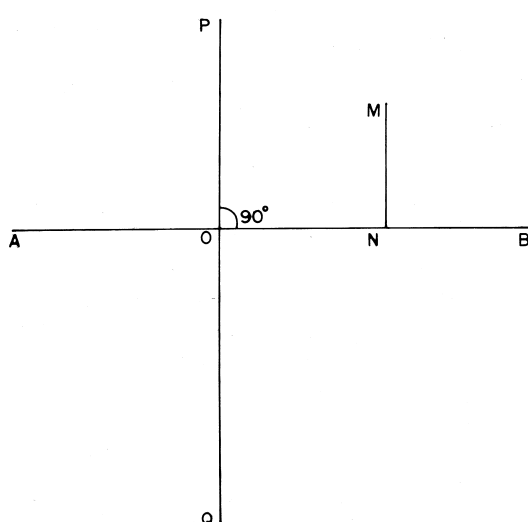


Fig.5.1 Showing coordinate axes

AB and PQ and O is the origin. Let M be any point on the graph paper. In order to determine its position, drop a perpendicular MN from M on AB. Then ON (i.e. the distance of N from the origin O) is called the 'abscissa' of M and MN is called its 'ordinate'. Thus, the abscissa is measured along the horizontal axis AB and the ordinate is measured along the vertical axis PQ. It is conventional to represent the abscissa by x and the ordinate by y ; and hence AB is called the x -axis and PQ is called the y -axis.

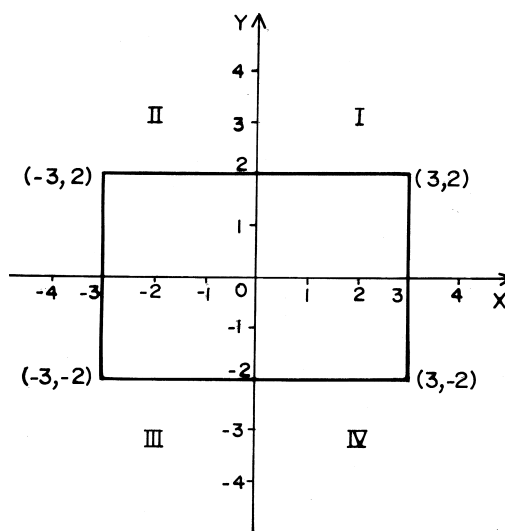


Fig.5.2 Showing four quadrants

The abscissa and ordinate, together, represent the 'coordinates' of the point and AB, PQ are the coordinate axes. Conventionally, the abscissa of the point is stated first and the ordinate next. Thus, if (2,4) are coordinates of a point, then $x=2$ is the abscissa and $y=4$ is the ordinate.

Another point to be noted in this context is that part of the x -axis is to the right of the origin 'O' and part to the left. The part of the x -axis to the right of O gives positive values of x and the part to the left of O gives negative values of x . Similarly, part of the y -axis is above the origin and part below. The part of the y -axis that is above O gives positive values of y and the part below O gives negative values of y see Fig.5.2. Thus, the graph paper is divided into four '**quadrants**' by the two axes. All points in the Quadrant I have both coordinates (abscissa and ordinate) positive. In Quadrant II the abscissa is negative and ordinate positive, in Quadrant III both, abscissa and ordinate, are negative and in Quadrant IV the abscissa is positive and ordinate negative.

Exercise

Locate the points on the graph paper whose coordinates are

(2, 8) (8, 2) (-2, 8) (-2, -8) (0,0) (-4, -6) (-6, 0) (0, -6)

2. Diagrammatic Presentation of Frequency Array - Line Graph

In Chapter 3, we have discussed that we obtain a frequency array, if the variable x is discrete and we are given frequency of each value of x . For example, a survey of 100 households in a city was carried out to ascertain the size of households.

The results of the survey were tabulated as the following frequency array.

Size of the household x	No. of households f
1	5
2	15
3	25
4	35
5	10
6	5
7	3
8	2
Total	100

The frequency array is shown in Fig. 5.3.

Scale: 1 cm = Frequency 5 on Y-axis
1 cm = Size 1 of h. h. on X-axis

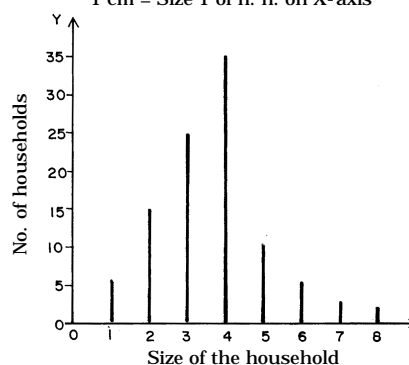


Fig.5.3 Diagrammatic presentation of frequency array

The figure consists, of a set of ordinates at different values of x (the size of the household), where the height of the ordinate is equal to the corresponding frequency (number of households). Thus, the height of the ordinate at $x=1$ is equal to 5, at $x=2$ it is 15, and so on. We have chosen the scale of measurement on the y -axis as : 1 cm=five households and on the x -axis 1 cm = size 1 of the household.

Exercise

Suppose a cubical die, with faces marked as 1, 2, ..., 6, was tossed 50 times and the following results were obtained:

x	Frequency
1	6
2	9
3	7
4	10
5	12
6	6
Total	50

Represent the frequency array diagrammatically.

3. Histogram

A frequency distribution can be represented diagrammatically, as a **histogram**. This is also called a **frequency histogram**.

We will consider two cases:

Case 1: When all class intervals are equal and have the same width, and

Case 2: When class intervals are unequal.

Let us consider Case 1 first.

Example 1: Equal Class Intervals

We have the following frequency distribution of marks obtained by 64 students in mathematics.

TABLE 5.1
Frequency Distribution of Marks in Mathematics

Marks (x) (Class intervals)	No. of students f
0-20	6
20-40	5
40-60	33
60-80	14
80-100	6
Total	64

While constructing the histogram, we note the following points.

- We measure the variable on the x -axis and show the class intervals.
- The frequencies are measured along the y -axis.
- Both the axes must be clearly labelled and the scale of measurement should be clearly shown.
- We construct rectangles on all class intervals, such that

the area of each rectangle is proportional to the frequency in the corresponding class interval.

Scale: 1 cm = Frequency 5 on Y-axis
1 cm = 10 marks on X-axis

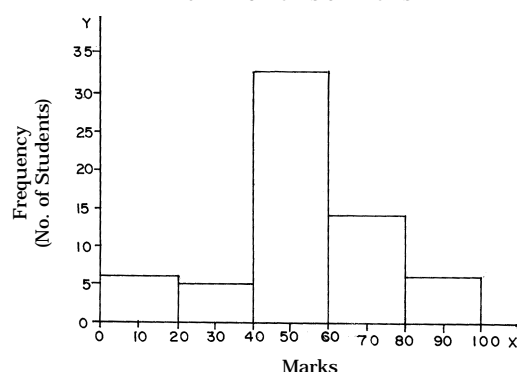


Fig.5.4 Histogram of marks in mathematics

In case of the frequency distribution in Table 5.1, shown as a histogram in Fig.5.4, the height of the first rectangle, on the class interval (0-20), is 6 (equal to the frequency in this class). The width of the class interval is 20. Therefore, the area of the rectangle is $20 \times 6 = 120$. Similarly, the height of the second rectangle is 5 and width is 20. Therefore, the area of the second rectangle is $20 \times 5 = 100$, and so on. We observe that the area of each rectangle is 20 times the frequency in the class interval. The factor '20' is the **constant of proportionality**.

The total area under the histogram is proportional to the total frequency, (i.e. 20×64).

Example 2: Equal Class Intervals

The following is the frequency distribution of monthly expenditure on food of 100 households in a certain city.

TABLE 5.2
Monthly Expenditure of Food of 100 Households

Monthly expenditure on food (in Rs) x	No. of households f
100-150	7
150-200	15
200-250	25
250-300	35
300-350	10
350-400	8
Total	100

Histogram is shown in Fig.5.5. It should be observed that on the x-axis we

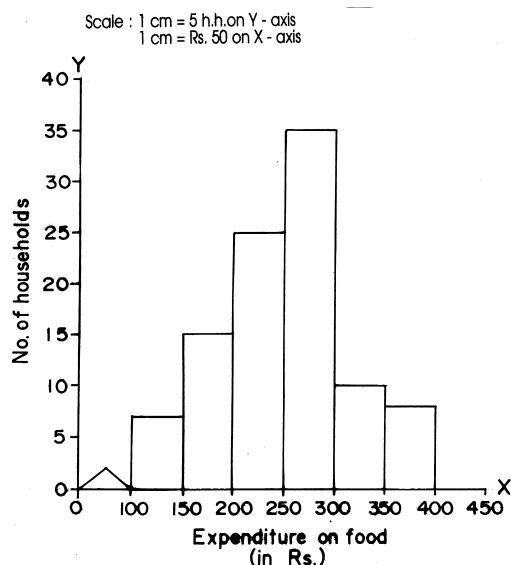


Fig.5.5 Histogram of monthly expenditure on food

show a kink, because we have a jump from 0 expenditure at the origin to 100. We could, as well, choose the origin at 100.

Example 3: Unequal Class Intervals

The following is the frequency distribution of households' per capita monthly expenditure.

TABLE 5.3
Frequency Distribution of Monthly per Capita Expenditure

Monthly per capita household expenditure (in Rs)	No. of households f	Height of the rectangle $f' = f \div \text{width of cl. Int.}$
0-15	161	≈ 11
15-25	152	≈ 15
25-50	60	≈ 2
50-100	27	≈ 1
Total	400	

Since the class intervals are unequal we may adopt any of the following two methods.

Method 1

Step 1. Divide the frequency in each class by the width of the class interval to obtain f' . This gives us the frequency per unit of the class interval.

Step 2. Measure f' on the y-axis and class intervals on the x-axis.

Since the height of the rectangle is $f' = \frac{f}{h}$ and width of the class interval is h , the area of the rectangle is f , i.e. the **area of the rectangle is equal to the frequency in the class interval.**

Method 2

We choose the class with the smallest width. Leave the frequencies of classes

with this smallest width unchanged; and obtain frequencies in other classes **proportionately** as shown below.

0-15	$\frac{10}{15} \times 161 \approx 107$
15-25	$\frac{10}{10} \times 152 \approx 152$
25-50	$\frac{10}{25} \times 60 \approx 24$
50-100	$\frac{10}{50} \times 27 \approx 5$

Draw the histogram with adjusted frequencies.

[This method is not satisfactory; as the total adjusted frequency is not equal to the total of observed frequencies.]

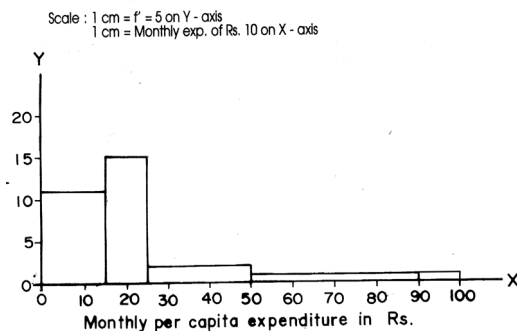


Fig.5.6 Histogram (unequal class intervals)

Example 4: Inclusive Class Intervals

The inventory - sales ratio of 124 companies are given in the table below.

TABLE 5.4
Frequency Distribution of Inventory - Sales Ratio

Inventory sales ratio (percentages)	Number of companies
0.1-5.0	2
5.1-10.0	3
10.1-15.0	8
15.1-20.0	14
20.1-25.0	38
25.1-30.0	59
Total	124

Draw a histogram.

We observe that the class intervals are given by **inclusive method** (where both the upper and lower limits are included in each class). In order to construct the histogram, we must readjust class limits to fill the gaps between two classes. The frequency distribution with adjusted class limits is given in Table 5.5.

TABLE 5.5
Frequency Distribution of Inventory - Sales Ratio with Adjusted Class Limits

Inventory sales ratio (%) (Adjusted Cl. limits)	No. of companies
0.05-5.05	2
5.05-10.05	3
10.05-15.05	8
15.05-20.05	14
20.05-25.05	38
25.05-30.05	59
Total	124

Since the upper class limit of the first class is 5.0 and the lower limit of the second class is 5.1,

$$\text{The adjusted limit} = \frac{5.0 + 5.1}{2} = 5.05,$$

and so on. In fact, we have added half of the difference (5.1 - 5.0) to the upper limits and subtracted it from the lower limits.

The histogram is shown in Fig.5.7.

Example 5 : Frequency Distribution when only the Mid-values of Class Intervals are Given.

Suppose a frequency distribution of x is given as follows.

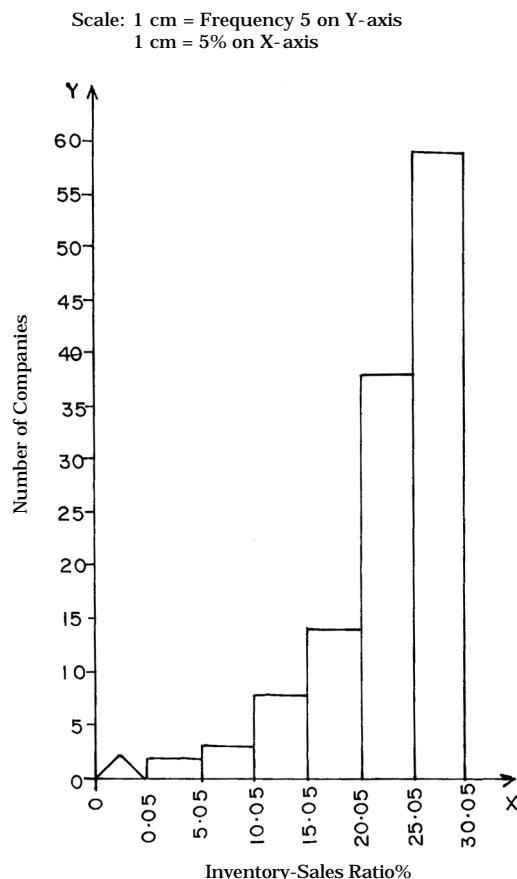


Fig. 5.7 Histogram showing distribution of inventory - sales ratio

TABLE 5.6
Frequency Distribution where Mid-values of Class Intervals are given

Mid-values of class intervals x	Frequency f
10	4
20	15
30	25
40	18
50	8
Total	70

Firstly, we should specify the class intervals as

5-15, 15-25, 25-35, 35-45, 45-55 which have mid-values as given. Then, draw the histogram as before.

4. Frequency Polygon

The frequency polygon is drawn by joining the middle points of the tops of adjacent rectangles by straight lines.

Let us illustrate this by constructing a frequency polygon for the data given in Table 5.1, for which the histogram is shown in Fig. 5.4.

For convenience we have drawn the histogram again in Fig. 5.8

While drawing the frequency polygon, we observe that some area which was under the histogram has been excluded and some area which was not under the histogram has been included under the frequency polygon. For example, see Fig. 5.8. The shaded area of the triangle RST was under the histogram, but is not under the polygon. It has been excluded from the polygon. But the shaded area of the triangle PQR has been included under the polygon. This was not under the histogram. Similarly, check other areas. There is always some area

Scale: 1 cm = Frequency 5 on Y-axis
1 cm = 20 Marks on X-axis

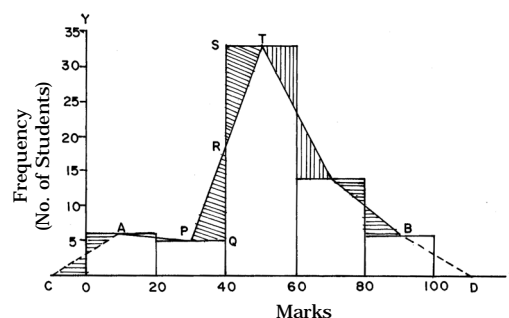


Fig. 5.8 Frequency polygon (superimposed on histogram of marks, for data in Table 5.1)

included under the polygon in lieu of area excluded from the histogram.

We require that the total area excluded from the histogram is equal to the area included under the frequency polygon. This can be achieved by a proper choice of classes.

However, note that the middle point A of the top of the rectangle on the first class interval cannot be the starting point of the frequency polygon. Because, then, the area to the left of A under the histogram is excluded without being compensated by any area included under the polygon. Similarly, B cannot be the end point. In order to solve this problem, we proceed as follows.

- (i) Extend the first class interval (0 to 20) to the left by the same width, as shown in the figure.
- (ii) Take the middle point C of the extended interval.
- (iii) Join C with A by a straight line. Similarly,
- (iv) Extend the last class interval (80-100) to the right by the same width, as shown in the figure.
- (v) Take the middle point D of the extended interval.
- (vi) Join D with B by a straight line.

Now, **the frequency polygon extends from C to D.**

5. Frequency Curve

As shown in Fig.5.9, the frequency curve is obtained by drawing a **smooth freehand curve** passing through the points of the frequency polygon (i.e, the middle points of the tops of adjacent rectangles of the histogram) as closely as possible.

The frequency curve does not necessarily pass through all points of the frequency polygon. But, it passes through them as closely as possible.

We require that

- (i) **the area under the frequency curve, over any class interval, is**

Scale: 1 cm = Frequency 5 on Y-axis
1 cm = 20 Marks on X-axis

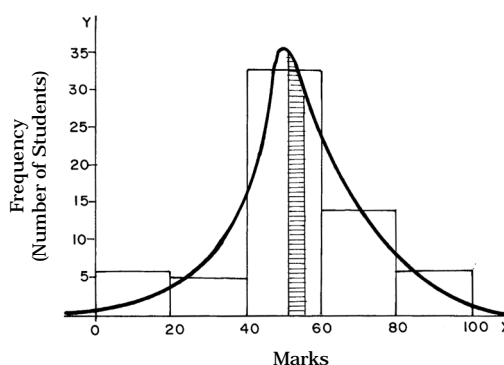


Fig.5.9 Frequency curve (superimposed on histogram of marks for data in Table 5.1)

TABLE 5.7

Cumulative Frequencies for Data in Table 5.1

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

proportional to the frequency in that class interval.

- (ii) **The total area under the frequency curve is proportional to the total frequency.**

Figure 5.9 shows the frequency curve of the distribution of marks in Table 5.1.

6. Cumulative Frequency Curve (Ogive)

Let us consider the frequency distribution of marks in mathematics in Table 5.1 as also shown in Table 5.7.

We observe that

6 students got marks **less than** 20
 11 students got marks **less than** 40
 44 students got marks **less than** 60
 58 students got marks **less than** 80
 64 students got marks **less than** 100
 Therefore, 6, 11, 44, 58 and 64 are

Scale: 1 cm = Frequency 5 on Y-axis
 1 cm = 10 Marks on X-axis

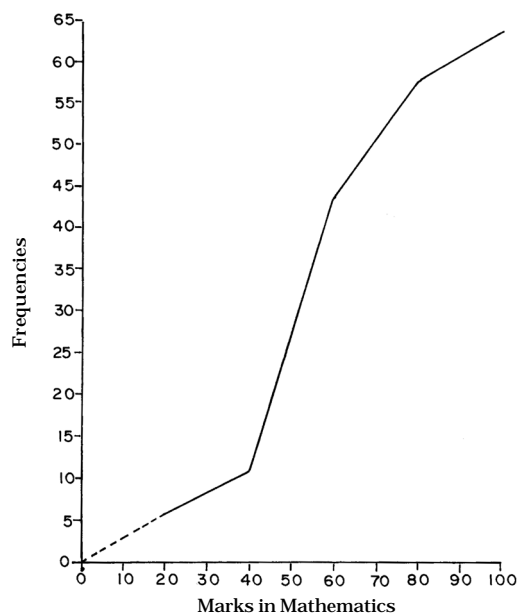


Fig.5.10 Less than cumulative frequencies

called '**less than**' frequencies, as shown in the third column of Table 5.7.

We plot these 'less than' cumulative frequencies corresponding to the upper end points of class intervals. Thus, 6 is plotted against 20, 11 against 40, 44 against 60, 58 against 80 and 64 against 100, as shown in Fig.5.10.

Similarly, we plot 'more than' cumulative frequencies against the lower ends of the class intervals, because

64 students got marks **more than** 0
 58 students got marks **more than** 20
 53 students got marks **more than** 40
 20 students got marks **more than** 60
 6 students got marks **more than** 80
 (see Fig.5.11)

Although in Figs.5.10 and 5.11 we have joined the points in the graph by straight lines, the cumulative frequency

Scale: 1 cm = Frequency 5 on Y-axis
 1 cm = 10 Marks on X-axis

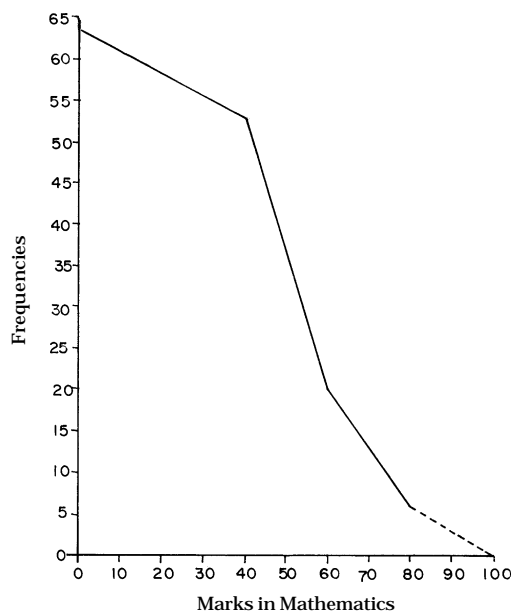


Fig.5.11 More than cumulative frequencies

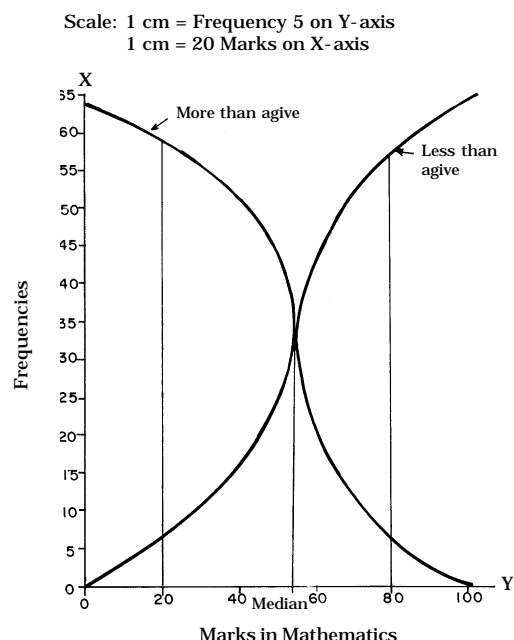


Fig.5.12 Less than and more than ogives

curve or ogive is obtained by drawing a smooth freehand curve through these points; see Fig.5.12 showing both ogives together.

We observe that

'less than' ogive is never decreasing and 'more than' ogive is never increasing.

At any value of x ,

the ordinate of less than ogive gives the number of students who got marks less than x ;

and

the ordinate of 'more than' ogive gives the number of students who got marks more than x .

At the point of intersection of the two ogives, we get the value of x (Marks in mathematics) so that the number of students who got marks more than and less than x is the same. This value of x is called the 'median'.

EXERCISES

1. Distinguish between a
 - (i) bar diagram and
 - (ii) histogram
2. Describe the procedure of drawing a histogram, when class intervals are
 - (i) equal and
 - (ii) unequal
3. Describe the procedure of drawing a
 - (i) frequency polygon and
 - (ii) frequency curve
4.
 - (i) What does the total area under a frequency curve represent?
 - (ii) What does the area, under a frequency curve, to the left of the ordinate at $x = \alpha$, represent, where α is some value of x within the total range?
 - (iii) If $x = \alpha$ and $x = \beta$, $\beta > \alpha$, are two values of x within the range, what does the area, under the frequency curve, between ordinates at $x = \alpha$ and $x = \beta$, represent?

- (iv) Draw a sketch of the frequency curve and show the areas mentioned in (i), (ii), and (iii).

5. Given a frequency distribution of x as

<i>Class intervals</i>	<i>Frequency</i>
10-20	15
20-40	35
40-80	65
80-100	5
Total	120

- (a) Obtain **less than** cumulative frequencies.
 (b) Obtain **more than** cumulative frequencies.
 (c) Draw **less than** and **more than** ogives on the same graph paper.
 (d) Where do the two ogives intersect?
 (e) Obtain the median value of x .
6. The following gives the frequency distribution of daily wage earnings, in rupees, of 450 workers.

<i>Daily wages (in Rs)</i>	<i>No. of persons</i>
70-80	44
80-90	120
90-100	80
100-110	76
110-120	50
120-130	45
130-140	25
140-150	10
Total	450

- (a) Draw a frequency histogram and then superimpose a frequency polygon and a frequency curve.
 (b) Obtain the cumulative frequencies and draw the two ogives.
 (c) Obtain the 'median' daily wage.
7. The frequency distribution of 80 public sector units' capacity utilization has been found to be as follows:

<i>Capacity utilization (%)</i>	<i>No. of units</i>
Less than 50	14
50.1-60	14
60.1-70	5
70.1-80	10
80.1-90	10
90.1-100	7
More than 100	20
Total	80

- (i) Draw a histogram and superimpose a frequency polygon and a frequency curve.
(Note that there is a gap between the end points of successive class intervals. We may modify class intervals as — 50.05, 50.05 – 60.05, 60.05 – 70.05, 70.05 – 80.05 – to eliminate the gap.)
- (ii) Would it matter, if the first and the last class intervals are open? How would you remedy this situation?

CHAPTER 6

Measures of Central Tendency

We have discussed classification and tabulation of data and their diagrammatic presentation in the earlier chapters. They help in visual comparison and interpretation of data and facilitate drawing conclusions about the salient features of the population.

However, we should like to know more about the distribution of values of the variable on which the data have been collected. For example,

- (a) what are the largest and the smallest values of the variable, and what is the range of the distribution of values,
- (b) what is the central value of the distribution about which other values are distributed,
- (c) what is the amount of variation in values about the central value, and
- (d) are the values of the variable uniformly distributed over the range, or are most of them clustered about some value in the upper or lower part of the range, or in the middle of the range?

In this chapter, we propose to discuss the measures of central tendency and postpone discussion of measures of dispersion to the next chapter.

The **measures of central tendency** are also called **measures of location**, or, briefly **averages**.

1. Averages

The term 'average value' is used by us in every day life. It is that 'single value' which represents all values in a given set of values. For example, the average height of students in your class represents heights of all students in the class. The average income of people in a certain region represents income of all people in that region.

The following averages are suitable in different situations :

- (i) **arithmetic mean**
- (ii) **median**
- (iii) **mode**
- (iv) **geometric mean**
- (v) **harmonic mean.**

Before we discuss their properties, let us discuss the requirements of a satisfactory average. (We omit discussion of geometric and harmonic means as they are not in the syllabus for Class XI students.)

2. Requirements of a Satisfactory Average

A satisfactory average should be

- (i) **rigidly defined**
- (ii) **based on all values**
- (iii) **simple and easy to interpret**
- (iv) **easy to calculate**

(v) **amenable to algebraic treatment**

If the average is not rigidly defined, it can be interpreted differently by different persons. The calculations may lead to different results. Further, if it does not use all values of the set it cannot be representative of the whole set of values. Algebraic treatment of averages is useful for further work.

3. Arithmetic Mean

Suppose ten students of your class secured the following percentage of marks in the Class X examination.

65	59	75	79	85
45	55	70	77	72

The arithmetic mean of marks is obtained by adding up the marks and dividing the total by the number of students, i.e.

$$\text{A.M.} = \frac{682}{10} = 68.2$$

Thus, the average percentage of marks is 68.2.

An interesting property of the arithmetic mean is that

the sum of deviations of values from the arithmetic mean is zero.

It is in this sense that the arithmetic mean is the central value of the distribution.

In the above example, the deviations of marks from the arithmetic mean are as shown in Table 6.1.

We note that the sum of positive deviations from the arithmetic mean is equal to the sum of negative deviations. In other words, positive and negative deviations balance each other. In this sense the arithmetic mean is analogous

TABLE 6.1
Deviations of Marks from the Arithmetic Mean

Percentage of marks	Deviations from A.M. (= 68.2)
65	-3.2
59	-9.2
75	+ 6.8
79	+ 10.8
85	+ 16.8
45	-23.2
55	-13.2
70	+ 1.8
77	+ 8.8
72	+ 3.8
Total	0

to the concept of 'centre of gravity' in physics.

Example 1

- Verify that the sum of deviations from any other value is not equal to zero.
- What is the sum of deviations from 'zero'? (Deviations from zero are also called deviations from the origin.)

In general, suppose x_1, x_2, \dots, x_n are n values of a certain variable x ; then, their arithmetic mean is defined as

$$\frac{x_1 + x_2 + \dots + x_n}{n}$$

which is symbolically expressed as \bar{x} . Briefly, we describe this as

$$\bar{x} = \frac{1}{n} \sum x_i$$

where $\sum x_i$ means the sum of n values.

\sum is a 'Greek' alphabet called 'sigma'.

In fact, this is capital sigma.

Example 2

Verify that

$$\sum (x_i - \bar{x}) = 0,$$

i.e. the sum of deviations from the arithmetic mean is zero.

Let us write.

$$\begin{aligned}\sum (x_i - \bar{x}) &= (x_1 - \bar{x}) + (x_2 - \bar{x}) + \dots + (x_n - \bar{x}) \\ &= (x_1 + x_2 + \dots + x_n) - n\bar{x} \\ &= 0\end{aligned}$$

$$\text{because } \bar{x} = \frac{1}{n} \sum x_i$$

Example 3

The monthly expenditures, in rupees, of 5 households are

1550	1715	1690	820	1150
------	------	------	-----	------

The arithmetic mean is 1385, i.e. the average monthly expenditure of these households is Rs 1385.

Note that the arithmetic mean may not be equal to any one of the values of the set.

4. Characteristics of Arithmetic Mean

Let us now consider the special features of the arithmetic mean, which will help in comparing it with other measures of central tendency.

- (i) As stated above, the sum of deviations of values from the arithmetic mean is zero.
- (ii) Suppose we are considering the monthly incomes of some households. Then the arithmetic mean will give us that income which each household would get, if the total income was equally distributed among all households.
- (iii) The arithmetic mean is a simple function of the values of the variable. This has two advantages:
 - (a) it is easy to derive the

mathematical properties of the arithmetic mean, and

- (b) it can be calculated unambiguously from the given data.

- (iv) In calculating the arithmetic mean we use all values of the variable. This means that even a single very large or very small value of the variable will unduly affect the arithmetic mean.

For example, the daily expenditure (in rupees) of 5 households in a locality are

25	28	32	27	33
----	----	----	----	----

The arithmetic mean of these values is Rs 29.

Now, suppose a rich household moves in this locality, whose daily expenditure is Rs 125. If we recalculate the arithmetic mean, we get its value as Rs 45. Since, the arithmetic mean has increased by more than 50 per cent one might think that the locality has become richer. However, we observe that 5 out of 6 households are spending exactly the same amount as before. The increase in the value of the arithmetic mean is entirely because of the new arrival.

5. Median

Arithmetic mean is the central value of the distribution in the sense that positive and negative deviations from the arithmetic mean balance each other. On the other hand, 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.

Suppose there are 17 students in your class. They secured the following percentage of marks in the Class X examination last year.

35	63	61	29	54	46	48	57	43
60	39	40	51	68	38	56	53	

If we arrange the marks in ascending order of magnitude, we get

29	35	38	39	40	43	46	48	51
53	54	56	57	60	61	63	68	

It follows that 8 students secured less than 51 per cent marks and 8 secured more than this percentage. Thus,

Median = 51 per cent.

Median is the central value in a sense different from the arithmetic mean. In case of the arithmetic mean it is the **numerical magnitude** of the deviations that balances. But, for the median it is the **number of values** greater than the median which balances against the number of values less than the median.

In general, if we have n values of x , they can be arranged in ascending order as

$$x_1 < x_2 < \dots < x_n.$$

Suppose **n is odd**, then

Median = the $\left(\frac{n+1}{2}\right)$ -th value.

However, if **n is even**, we have two middle values, viz., the

$$\left(\frac{n}{2}\right)\text{-th and } \left(\frac{n}{2} + 1\right)\text{-th.}$$

In this case,

$$\text{Median} = \frac{\left(\frac{n}{2}\right)\text{-th value} + \left(\frac{n}{2} + 1\right)\text{-th value}}{2}$$

i.e. the arithmetic mean of the two middle most values.

Example 4

The heights (in cm) of six students in your class are

160, 157, 156, 161, 159, and 162.

What is the median height?

If we arrange heights in ascending order, we get

156	157	159	160	161	162
-----	-----	-----	-----	-----	-----

The two middle most values are the 3rd and 4th. The median = $\frac{1}{2}(159 + 160) = 159.5$. Compare this with the arithmetic mean = 159.2 (approx.).

6. Characteristics of Median

The main features of the median are as follows.

- It is simple and easy to understand.
- Like the arithmetic mean, median is rigidly defined.
- It is not affected by very large or very small values.
- The median cannot be expressed as a simple function of the values of the given set.

7. Mode

Mode is defined as the value of the variable which occurs most frequently.

The following is the frequency array of ages of students in a cricket team in your school.

TABLE 6.2
Frequency Array of Ages of Students

Age	Number of students
14	2
15	2
16	4
17	2
18	1
Total	11

There are 4 students of age 16. Since, this is the most common age, with largest frequency, modal value is 16 years.

Sometimes, mode is more meaningful average than the arithmetic mean or the median. Consider the following example.

Example 5

A company which manufactures shoes for adult men would like to concentrate on the production of shoes of only one size. What should this size be?

The company conducted a survey of 2000 customers regarding their shoe size and obtained the following distribution.

TABLE 6.3
Frequency Distribution of Shoe Size

Shoe size	Number of customers
5	55
6	164
7	360
8	528
9	410
10	260
11	137
12	86
Total	2000

In this example, neither the arithmetic mean nor the median would be an appropriate average. It would be better to use mode. Since the largest frequency is 528 the mode is 8. It would be advisable for the company to specialize in the manufacture of shoes of size 8.

8. Characteristics of Mode

- Like the arithmetic mean and median, mode too has an easy interpretation.
- Mode is not affected by extreme values because it depends only on the largest frequency.
- Mode, like the median, cannot be

expressed as a simple function of the given values.

9. Calculation of the Arithmetic Mean for Ungrouped Data

Direct Method

Let x_1, x_2, \dots, x_n be n values of some variable x . The values have not been grouped into class intervals. Then, by direct method, the arithmetic mean of x is obtained as

$$\bar{x} = \frac{1}{n} \sum x_i;$$

i.e. we simply add up all values of x and divide the sum by their number.

Assumed Mean Method

Let A be the assumed mean. We obtain the deviations of x from A , i.e.

$$d_1 = x_1 - A, d_2 = x_2 - A, \dots, d_n = x_n - A.$$

Next, obtain the arithmetic mean of deviations

$$\bar{d} = \frac{\sum d_i}{n}$$

and

$$\bar{x} = A + \bar{d}.$$

In other words, obtain the arithmetic mean of the deviations (d) and add A , which was earlier subtracted.

Step Deviation Method

Firstly, obtain deviations (d) from an assumed mean.

Divide the deviations by a common factor, say h , to obtain

$$d_1' = \frac{d_1}{h}, d_2' = \frac{d_2}{h}, \dots, d_n' = \frac{d_n}{h}$$

Then

$$\bar{x} = A + h\bar{d}'$$

where

$$\bar{d}' = \frac{1}{n} \sum d'_i$$

In other words, obtain the arithmetic mean of d'_1, d'_2, \dots, d'_n ; then multiply it by h (because we had originally divided by h) and add A (because we had earlier subtracted A).

Example 6

Suppose the monthly incomes (x), in rupees, of 5 individuals, are given as follows:

6550, 7550, 9550, 4550 and 8000.

Obtain their arithmetic mean.

Direct Method

By **direct method**, we simply add all incomes and divide by the number of individuals. Thus, we obtain

$$\bar{x} = \frac{36200}{5} = 7240,$$

i.e. the arithmetic mean of incomes of 5 individuals is Rs 7240.

Calculation of the Arithmetic Mean by Assumed Mean Method

Let the assumed mean be $A = 5550$.

Obtain deviations (d) from the assumed mean as

$d_1 = 1000, d_2 = 2000, d_3 = 4000, d_4 = -1000$ and $d_5 = 2450$,

So that $\sum d_i = 8450$ and $\bar{d} = 1690$.

Then

$$\bar{x} = A + \bar{d}$$

gives

$$\bar{x} = 5550 + 1690 = 7240,$$

as before.

Step Deviation Method

We divide the deviations (d) by a common factors, say 1000; and obtain

$d'_1=1, d'_2=2, d'_3=4, d'_4=-1$ and $d'_5=2.45$

Now

$$\bar{d}' = \frac{\sum d'_i}{5} = \frac{8.45}{5} = 1.69;$$

and

$$\begin{aligned}\bar{x} &= A + 1000 \bar{d}' \\ &= 5550 + 1000 \times 1.69 \\ &= 7240\end{aligned}$$

as before.

We should note that taking deviations from an assumed mean, as well as the step deviation method, help in reducing the burden of calculations a great deal. They are particularly useful, if the numbers are large.

10. Calculation of Arithmetic Mean for a Frequency Array

The variable x is discrete and takes values x_1, x_2, \dots, x_n with frequencies f_1, f_2, \dots, f_n . (No class intervals are specified.)

Let us write the frequency array as

TABLE 6. 4
Frequency Array

Values of x	Frequency f
x_1	f_1
x_2	f_2
\vdots	\vdots
x_n	f_n
Total	$N = \sum f_i$

Thus, x takes the value x_1, f_1 times
 x takes the value x_2, f_2 times
 x takes the value x_n, f_n times

Therefore, by the **Direct Method**, the arithmetic mean of x is

$$\begin{aligned}\bar{x} &= \frac{\overbrace{X_1 + \dots + X_1}^{f_1 \text{ times}} + \overbrace{X_2 + \dots + X_2}^{f_2 \text{ times}} + \dots + \overbrace{X_n + \dots + X_n}^{f_n \text{ times}}}{N} \\ &= \frac{f_1 x_1 + f_2 x_2 + \dots + f_n x_n}{N} \\ &= \frac{\sum f_i x_i}{N}\end{aligned}$$

where $N = \sum f_i$

Calculation of Arithmetic Mean by Assumed Mean Method

Let A be the assumed mean.

Obtain the deviations of the values of x from the assumed mean A :

$$d_1 = x_1 - A, \dots, d_n = x_n - A$$

Next, obtain

$$\bar{d} = \frac{1}{N} \sum f_i d_i$$

and
 $\bar{x} = A + \bar{d}$.

Calculations of Arithmetic Mean by Step Deviation Method

Let h be the common factor.

Divide each deviation (d) by h , and obtain

$$d'_1 = \frac{d_1}{h}, \dots, d'_n = \frac{d_n}{h}$$

Then

$$\bar{d}' = \frac{1}{N} \sum f_i d'_i$$

and

$$\bar{x} = A + h\bar{d}'$$

In words, we multiply \bar{d}' by h (because we had divided by h earlier), and add A (because earlier we had subtracted A).

Example 7

A survey of 100 households was carried out to obtain information on their family size. The results of the survey are classified as a frequency array in the table below.

TABLE 6.5
Frequency Array of Size of Households

Size of the household x	Number of households f
1	5
2	15
3	25
4	35
5	10
6	5
7	3
8	2
Total	100

Calculate the arithmetic mean of size of households.

Direct Method

Let us arrange the calculations as in the following table.

TABLE 6.6
**Calculation of Arithmetic Mean
(By Direct Method)**

x	f	$f \times x$
1	5	5
2	15	30
3	25	75
4	35	140
5	10	50
6	5	30
7	3	21
8	2	16
Total	100	367

Therefore, the arithmetic mean is

$$\bar{x} = \frac{367}{100} = 3.67$$

Assumed Mean Method

Let the assumed mean of x be 4.

The calculations are arranged as in the table below.

TABLE 6.7
Calculation of Arithmetic Mean
(By Assumed Mean Method)

x	$d = x - 4$	f	$f \times d$
1	-3	5	-15
2	-2	15	-30
3	-1	25	-25
4	0	35	0
5	1	10	10
6	2	5	10
7	3	3	9
8	4	2	8
Total		100	$-70 + 37 = -33$

Therefore,

$$\bar{d} = -\frac{33}{100} = -0.33 \text{ and}$$

$$\bar{x} = 4 + \bar{d} = 4 - .33 = 3.67$$

as before.

In this example there is no extra advantage in using the step deviation method, because the values of x are already small and deviations from assumed mean are convenient.

Let us consider another example.

Example 8

Let the frequency array of monthly incomes of individuals be as follows.

TABLE 6.8
Frequency Array of Incomes

Income (in rupees)	Number of individuals
11,500	629
12,000	1705
12,500	1800
13,000	725
13,500	420
14,000	127
14,500	85
15,000	9
Total	5500

Direct Method

Let us arrange the calculations in the following table.

TABLE 6.9
Calculation of Arithmetic Mean
(By Direct Method)

Income x	No. of Individuals f	$f \times x$
11,500	629	7,233,500
12,000	1705	20,460,000
12,500	1800	22,500,000
13,000	725	9,425,000
13,500	420	5,670,000
14,000	127	1,778,000
14,500	85	1,232,500
15,000	9	135,000
Total	5500	68,434,000

Therefore, the arithmetic mean of incomes

$$\bar{x} = \frac{\sum f x_i}{N} = \frac{68,434,000}{5,500} \approx 12,442.55$$

Assumed Mean Method

Let the assumed mean be 12,500. We arrange calculations in the table below.

TABLE 6.10
Calculation of Arithmetic Mean
(By Assumed Mean Method)
(Assumed Mean = 12,500)

Income x	Deviations $d = x - 12500$	f	$f \times d$
11,500	-1000	629	-629,000
12,000	-500	1705	-852,500
12,500	0	1800	0
13,000	500	725	362,500
13,500	1000	420	420,000
14,000	1500	127	190,500
14,500	2000	85	170,000
15,000	2500	9	22,500
Total		5,500	-316,000

Therefore,

$$\bar{d} = \frac{1}{N} \sum f_i d_i = -\frac{316000}{5500} \approx -57.45$$

and

$$\bar{x} = A + \bar{d} = 12442.55.$$

Step Deviation Method

Let us now calculate the arithmetic mean of incomes by step deviation method. The calculations are shown in the following table.

TABLE 6.11
Calculation of Arithmetic Mean
(By Step Deviation Method)
(Assumed Mean = 12500 and Common Factor 500)

Income x	Deviations $d = x - 12500$	$d' = \frac{d}{500}$	f	$f \times d'$
11500	-1000	-2	629	-1258
12000	-500	-1	1705	-1705
12500	0	0	1800	0
13000	500	1	725	725
13500	1000	2	420	840
14000	1500	3	127	381
14500	2000	4	85	340
15000	2500	5	9	45
Total			5500	-632

Therefore,

$$\bar{d}' = \frac{1}{N} \sum f_i d'_i = -\frac{632}{5500} \approx -0.1149$$

and

$$\begin{aligned}\bar{x} &= A + h\bar{d}' \\ &= 12500 - 500 \times 0.1149 \\ &= 12442.55\end{aligned}$$

We should note that the step deviation method simplifies large calculations considerably.

11. Calculation of Arithmetic Mean - for Grouped Data

We assume that x is a continuous variable and the values of x are grouped into **equal** class intervals.

Let us illustrate calculations by using the following example.

Example 9

The frequency distribution of the monthly expenditure of 100 rural households is given below.

TABLE 6.12
Frequency Distribution of Monthly
Expenditure of 100 Households

Monthly expenditure in rupees (class intervals)	Number of households
75 - 125	10
125 - 175	22
175 - 225	38
225 - 275	16
275 - 325	8
325 - 375	4
375 - 425	2
Total	100

Calculate the arithmetic mean of monthly expenditures by

- direct method
- assumed mean method, and
- step deviation method.

(a) Direct Method

Let us arrange the calculations in the following table.

TABLE 6.13
Calculation of Arithmetic Mean
(By Direct Method)

Class interval	Mid-value of class interval x	f	$f \times x$
(1)	(2)	(3)	(4)
75 - 125	100	10	1000
125 - 175	150	22	3300
175 - 225	200	38	7600
225 - 275	250	16	4000
275 - 325	300	8	2400
325 - 375	350	4	1400
375 - 425	400	2	800
Total		100	20500

Mid-values are obtained as

$$\frac{75+125}{2}=100, \frac{125+175}{2}=150, \frac{175+225}{2}=200,$$

and so on.

$$\text{The arithmetic mean} = \frac{\sum f_i x_i}{\sum f_i} = \frac{20500}{100} = 205.$$

(b) Assumed Mean Method

Suppose the assumed mean = 200. Then the calculations of arithmetic mean are as shown in the following table.

TABLE 6.14
Calculation of Arithmetic Mean
(By Assumed Mean Method)
(Assumed Mean = 200)

Class interval	Mid-value x	Deviations $d = x - 200$	f	$f \times d$
(1)	(2)	(3)	(4)	(5)
75 - 125	100	-100	10	-1000
125 - 175	150	-50	22	-1100
175 - 225	200	0	38	0
225 - 275	250	50	16	800
275 - 325	300	100	8	800
325 - 375	350	150	4	600
375 - 425	400	200	2	400
Total			100	500

Therefore,

$$\bar{d} = \frac{\sum f_i d_i}{N} = \frac{500}{100} = 5$$

and

$$\bar{x} = A + \bar{d} = 200 + 5 = 205.$$

(c) Step Deviation Method

Looking at the Column (2) of the above Table 6.14, we see that the common factor is 50. In fact, this is the width of class intervals.

The calculations are as shown in the following table.

TABLE 6.15
Calculation of Arithmetic Mean
(By Step Deviation Method)
(Assumed Mean = 200 and
Common Factor = 50)

Class interval	Mid-value x	$d = x - 200$	$d' = \frac{d}{50}$	f	$f \times d'$
(1)	(2)	(3)	(4)	(5)	(6)
75 - 125	100	-100	-2	10	-20
125 - 175	150	-50	-1	22	-22
175 - 225	200	0	0	38	0
225 - 275	250	50	1	16	16
275 - 325	300	100	2	8	16
325 - 375	350	150	3	4	12
375 - 425	400	200	4	2	8
Total				100	10

Therefore,

$$\bar{d}' = \frac{\sum f_i d'_i}{N} = \frac{10}{100} = \frac{1}{10} = 0.1$$

and

$$\bar{x} = A + h\bar{d}' = 200 + 50 \times 0.1 = 205.$$

where $A = 200$ and the common factor $h = 50$.

12. Calculation of Median

Ungrouped Data

Given a set of values of the variable, we arrange them in ascending (or, descending) order of magnitude. Median is the **middle most** value.

If the number of values is odd, say 101, median is the $\frac{101+1}{2} = 51$ -st value. In general, if $N = 2m+1$, the median is the $(m+1)$ -th value. Otherwise, if $N = 2m$ is even, there are two middle most values, viz. the m -th and $(m+1)$ -th values. Median is the arithmetic mean of the two.

We have considered examples before.

Example 10

The following data give the number of passengers carried by a bus in ten trips.

22	26	14	30	18
11	35	41	12	32

Verify that the arithmetic mean of number of passengers carried is 24.1.

In order to find the median number, we arrange the numbers in ascending order as

11 12 14 18 22 26 30 32 35 41

There are two middle values, viz., the 5th and the 6th. Median is the arithmetic mean of the two, i.e.

$$\text{median} = \frac{22 + 26}{2} = \frac{48}{2} = 24.$$

Grouped Data**Example 11**

Suppose x is a continuous variable, and the values have been grouped into class intervals. We have the frequency distribution shown in the table below.

TABLE 6.16
Frequency Distribution

Class interval x	Frequency f
75 - 125	15
125 - 175	22
175 - 225	38
225 - 275	16
275 - 325	8
325 - 375	5
375 - 425	3
Total	107

Calculate the median value of x .

In order to calculate the median, first of all, calculate the cumulative frequencies as shown in the following table.

TABLE 6.17
Cumulative Frequencies

Class Interval (1)	Frequency (2)	Less than cumulative frequencies (3)
75 - 125	15	15=15
125 - 175	22	37=15+22
175 - 225	38	75=15+22+38
225 - 275	16	91=15+22+38+16
275 - 325	8	99=15+22+38+16+8
325 - 375	5	104=15+22+38+16+8+5
375 - 425	3	107=15+22+38+16+8+5+3
Total	107	

The values of x are already arranged in ascending order in the frequency table. We observe that 15 values are less than 125; 15+22=37 values are less than 175; 15+22+38=75 values are less than 225, and so on. These are called 'less than' cumulative frequencies.

Since x is a continuous variable, the $\frac{N}{2}$ -th value (i.e. $\frac{107}{2} = 53.5$)-th is the the middle most value, i.e. the median.

Since 37 values are less than 175 and 75 values are less than 225, the $\frac{N}{2}$ -th value (i.e. 53.5 -th value) will lie in the class (175-225).

Thus, (175 -225) is the **median class**.

We assume that all values are uniformly distributed in any class.

Now, use the formula given below to interpolate the median, i.e. the 53.5-th value:

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

where

L = Lower Limit of the median class (i.e. = 175)

N = total frequency (i.e.107)

c.f.=number of values less than the lower limit L (i.e. 37)

f = frequency in the median class (i.e. 38)

h = width of the median class (i.e. 50)

Therefore, substituting values in the above formula

$$\text{Median} = 175 + \frac{53.5 - 37}{38} \times 50 = 196.7$$

In order to compare this value with the arithmetic mean, see that the arithmetic mean in this case is 203.27.

13. Calculation of Mode

Mode is that value of x for which the frequency is maximum. For example, if x takes the values 70, 69, 69, 75, 75, 75, 60 and 80, the mode (or, the modal value) is clearly 75.

If the values of x are grouped into classes (such that they are uniformly distributed within any class) and we have a frequency distribution,

(i) identify the class which has the largest frequency, and

(ii) calculate the mode as

$$\text{Mode} = L + \frac{f_0 - f_1}{2f_0 - f_1 - f_2} \times h$$

where

L = Lower limit of the modal class

f_0 = Largest frequency

f_1 = frequency in the class preceding the modal class

f_2 = frequency in the class next to the modal class

h = width of the modal class.

Observe that if $f_1 = f_2$,

$$\text{Mode} = L + \frac{1}{2} h$$

i.e. mode is equal to the middle value of the modal class.

Example 12

Data relating to the height of 352 school students are given in the following frequency distribution.

Calculate the modal height.

Height (in feet)	No. of students
3.0 - 3.5	12
3.5 - 4.0	37
4.0-4.5	79
4.5-5.0	152
5.0-5.5	65
5.5-6.0	7
Total	352

Since 152 is the largest frequency, the modal class is (4.5 - 5.0). Thus, $L = 4.5$, $f_0 = 152$, $f_1 = 79$, $f_2 = 65$ and $h = 0.5$, so that

$$\begin{aligned} \text{Mode} &= 4.5 + \frac{152 - 79}{2 \times 152 - 79 - 65} \times 0.5 \\ &= 4.73 \text{ ft. (approx.).} \end{aligned}$$

14. Relative Position of the Arithmetic Mean, Median and Mode

Suppose we use the following notations

Arithmetic Mean = M_e

Median = M_i

Mode = M_o

where e , i and o are the suffixes. Then the relative magnitude of the three are either of the following :

(i) $M_e > M_i > M_o$,

(ii) $M_e < M_i < M_o$

suffixes occurring in alphabetical order.

The median is always between the arithmetic mean and the mode.

15. Partition Values — Quartiles

Median divides the total set of values of the variable in two equal parts. Similarly, the **quartiles** divide the total set of values into four equal parts.

There are three quartiles Q_1 , Q_2 and Q_3 .

Q_1 is called the lower quartile.

The number of values less than Q_1 is one-fourth and more than Q_1 is

three-fourths of the total number of values in the set.

Q_3 is called the upper quartile

Three-fourths of the total number of values in the set are less than Q_3 and one-fourth are greater than Q_3 .

Q_2 is the median or the second quartile.

Thus, the number of values between Q_1 and Q_2 is one-fourth of the total number of values; and so is the number between Q_2 and Q_3 .

The number of values between Q_1 and Q_3 is one-half of the total number of values in the entire set.

While calculating quartiles we assume that the values of the variable are uniformly distributed in any class.

Example 13

Calculate the quartiles Q_1 and Q_3 for the following frequency distribution.

Class interval	Frequency	Cum frequency
0-25	23	23
25-50	64	87
50-75	115	202
75-100	51	253
100-125	30	283
125-150	17	300
Total	300	

The total frequency = $N = 300$,

therefore, $\frac{N}{4} = 75$.

We find that 23 values are less than 25 and 87 are less than 50. Therefore, the 75th value (Q_1) must lie in the class interval 25-50.

Q_1 is calculated by using the formula

$$Q_1 = L + \frac{\frac{N}{4} - \text{c.f.}}{f} \times h$$

where

L = lower limit of the class in which Q_1 lies,

N = total frequency

c.f. = number of values less than L

f = frequency in the class in which Q_1 lies,

h = width of the class in which Q_1 lies.

In our example

$$L = 25, \frac{N}{4} = 75, \text{c.f.} = 23, f = 64 \text{ and } h = 25.$$

Therefore, substituting in the above formula $Q_1 = 45.3125$.

We follow a similar procedure to interpolate Q_3 .

Q_3 is the $\left(\frac{3}{4}N\right)$ -th value

$$\text{Since } N = 300, \frac{3}{4}N = 225.$$

We have to locate the 225-th value.

From the frequency distribution, we see that 202 values are less than 75 and 253 values are less than 100. Therefore, the 225-th value must lie in the class 75-100.

Q_3 is calculated by using a similar formula as above for Q_1 .

$$Q_3 = L + \frac{\frac{3}{4}N - \text{c.f.}}{f} \times h$$

where

L = lower limit of the class in which Q_3 lies,

N = total frequency

c.f. = number of values less than L

f = frequency in the class in which Q_3 lies

h = size of the class in which Q_3 lies.

In the present case, Q_3 lies in the class interval 75-100.

We have

$$L = 75, \frac{3}{4}N = 225, \text{c.f.} = 202, f = 51 \text{ and } h = 25.$$

Therefore, substituting in the above formula $Q_3 = 86.2745$.

EXERCISES

1. Compare the arithmetic mean, median and mode as measures of central tendency. Describe situations where one is more suitable than the others.
2. The arithmetic mean is described as the centre of gravity of the distribution of values of the variable. Explain.
3. Show that the sum of deviations of the values of the variable from their arithmetic mean is equal to zero.
4. If x_1, \dots, x_n are n values of the variable x and A is any arbitrary value, so that $\sum(x_i - A)$ is the sum of deviations from A ; show that $\sum(x_i - A) = n(\bar{x} - A)$ where $\bar{x} = \frac{1}{n} \sum x_i$ is the arithmetic mean of x . Hence $\sum(x_i - A) = 0$ if and only if, $A = \bar{x}$.
5. 'Arithmetic mean is affected by very large and very small values, but median and mode are not affected by them. Explain?
6. Daily expenditures on vegetables (in rupees) of 20 households, in a certain locality, are given below.

25.00	26.50	30.25	28.00	23.00
31.40	34.00	33.00	30.50	27.20
28.00	35.00	38.60	34.00	22.50
24.00	23.70	28.00	29.00	32.20

Show that the value of the median lies between the arithmetic mean and the mode.

7. Calculate the arithmetic mean, median and mode for the following frequency distribution.

<i>Class Interval</i>	<i>Frequency</i>
1-3	3
3-5	10
5-7	16
7-9	13
9-11	8
11-13	3
13-15	1
Total	54

Examine their relative positions. Does the median lie between the arithmetic mean and mode?

8. Out of 50 questionnaires distributed among 50 manufacturing firms only 35 replies were received. In the column inquiring about 'capacity utilization' the figures read.

(In percentages)

54.2	96.8	74.2	57.8	99.7	84	81.2
94.3	93.7	61.6	100.0	110.1	68	71
95.2	98.3	84	91.9	84	94.1	73
84	52.6	49.2	45	62.6	75.1	84
69.2	90.7	85.4	87.3	66.1	73.9	89.3

- (a) Calculate the arithmetic mean, median and mode, and examine their relative position.
- (b) Group the data into appropriate classes, and again calculate the arithmetic mean, median and mode from grouped data. How do your results differ from those obtained in (a)? Explain.
- 9.** In a certain examination there were 100 candidates of whom 21 failed, 6 secured distinction, 43 were placed in the third division and 18 in the second division. It is known that at least 75 per cent marks are required for distinction, at least 40 per cent for passing, at least 50 per cent for second division and at least 60 per cent for first division.
Calculate the median of the distribution of marks.
- 10.** Calculate the upper and lower quartiles for the following frequency distribution.

<i>Class interval</i>	<i>Frequency</i>
13-25	6
25-37	11
37-49	23
49-61	7
61-73	3
Total	50

- 11.** Calculate the median for the distribution in the Question No.10. Also, derive the median geometrically, by drawing the two ogives and locating the point of intersection.
- 12.** The daily expenditure, in rupees, of 50 households is given as follows.

<i>Daily expenditure (Rs)</i>	<i>Number of households</i>
100-150	3
150-200	9
200-300	21
300-500	10
500-1000	5
Above 1000	2
Total	50

[In this example the last class is open. Therefore, we cannot determine the mid-value of that class. The arithmetic mean cannot be calculated for the given data. However, we may set an upper limit on some basis and then obtain the arithmetic mean. But, this must be regarded as an approximation. Calculation of median, mode and quartiles do not pose any difficulty.]

- (i) Calculate the arithmetic mean, median and mode. Compare the values.
- (ii) Also, compute the upper and lower quartiles.

CHAPTER 7

Measures of Dispersion

Whereas, the measures of central tendency provide a central value of the distribution, the measures of dispersion are required to measure the amount of variation (dispersion or scatter) of values about the central value. For example, suppose the monthly incomes (in rupees) of five households are

MONTHLY INCOMES IN RUPEES

4500	6000	5500	3750	4700
------	------	------	------	------

The arithmetic mean of income is Rs 4,890 and the median income Rs 4,700.

The amount of variation in incomes is shown by deviations from the central value.

DEVIATIONS FROM THE ARITHMETIC MEAN

-390	1110	610	-1140	-190
------	------	-----	-------	------

DEVIATIONS FROM THE MEDIAN

-200	1300	800	-950	0
------	------	-----	------	---

We observe that some deviations are positive and others are negative. Also, some deviations are large and others are small. We require an over all summary measure of variation in all values about the central value. This summary measure is called the **measure of dispersion**.

In this chapter we will discuss the following measures of dispersion

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

While the mean deviation and standard deviation are defined in terms of deviations from a central value, the range and quartile deviation are not based on deviations from any particular value.

1. The Range

The range is defined as the difference between the largest and the smallest value of the variable in the given set of values. Suppose, the values of x are arranged in ascending order as

$$x_1 < x_2 < \dots < x_n,$$

so that x_n is the largest and x_1 is the smallest value; then the range is defined as

$$R = x_n - x_1$$

All values of x lie within the range.

If the range is large, the spread of values is large and hence the variation of values of x is large. On the other hand, if the range is small, the spread of values of x is small, and the variation of x is small.

2. Characteristics of the Range

Let us note the following characteristics of the range.

- (i) **It is rigidly defined.**
- (ii) **It is easy to calculate and simple to interpret.**
- (iii) **It does not depend on all values of the variable.**

The range does not take into account the distribution of values between the smallest and largest values.

- (iv) **It is unduly affected by extreme values.**

For example, consider the monthly incomes of five households given on page 63. Since the largest income is Rs 6,000 p.m. and smallest is Rs 3,750 p.m., the range is Rs 2,250. If we include the income Rs 15,000 p.m. of a sixth household, the range increases by five times. It is equal to Rs 11,250 (i.e. $5 \times 2,250$).

- (iv) **The range depends on the units of measurement of the variables.**

It has the same unit as of the variable considered. The range of incomes is in terms of rupees, the range of distances may be in kilometres, etc.

3. Quartile Deviation

Given a set of values of the variable (grouped or ungrouped data) we can calculate the upper quartile Q_3 and lower quartile Q_1 , as defined in the earlier chapter. Then the **quartile deviation** is defined as:

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

This is also called the **semi-inter-quartile range**. In fact, $Q_3 - Q_1$ is the 'inter-quartile range'.

The advantage of using the quartile deviation is that, unlike the range, it is not affected by extreme values. In other

respects, quartile deviation has the same features as those of the range.

- (i) **It is rigidly defined.**
- (ii) **It is easy to calculate and simple to interpret.**
- (iii) **It does not depend on all values of the variable.**
- (iv) **The units of measurement of the quartile deviation are the same as those of the variable.**

Example 1

We are given the following frequency distribution of a certain variable x .

Class Interval x	Frequency f
10-20	4
20-40	10
40-70	26
70-120	8
120-200	2
Total	50

Calculate the range and quartile deviation and compare the two as measures of dispersion.

It is clear that the range is 190, as the largest value of x is 200 and the smallest value is 10.

In order to calculate the quartile deviation, first of all calculate the cumulative frequencies as shown in the following table.

TABLE 7.1
Calculation of Cumulative Frequencies

Class interval x	Frequency f	Cumulative frequencies
10-20	4	4
20-40	10	14
40-70	26	40
70-120	8	48
120-200	2	50
Total	50	

Calculation of Q_1

By definition, the number of values less than Q_1 is $\frac{1}{4} \times 50 = 12.5$.

We assume that the values of x are uniformly distributed in any class. Since the values of x are already arranged in ascending order, locate the class in which the 12.5-th value lies.

The class in which the 12.5-th value lies is (20-40), as 4 values are less than 20 and 14 are less than 40.

Use the formula

$$L + \frac{\frac{N}{4} - \text{c.f.}}{f} \times h$$

where

$L = 20$, $\frac{N}{4} = 12.5$, $\text{c.f.} = 4$, $f = 10$ and $h = 20$.

Substituting the values

$$Q_1 = 20 + \frac{12.5 - 4}{10} \times 20 = 37$$

Similarly, obtain Q_3 , which is the $\frac{3}{4} \times 50 = 37.5$ -th value.

The class in which Q_3 lies is 40-70, as 14 values are less than 40, and 40 values are less than 70. Therefore, using the formula given before

$$Q_3 = 40 + \frac{37.5 - 14}{26} \times 30 = 67.11$$

Hence, the **quartile deviation is**

$$\frac{67.11 - 37}{2} = \frac{30.11}{2} = 15.055 = 15.1$$

This has the same units of measurement as of x .

4. Measures of Dispersion in Terms of Deviations from a Central Value

We have noted before that the range and quartile deviation measure dispersion in

a general way. They do not refer to any particular value of the distribution. However, the deviations from a central value, or, any other value, give a better picture of dispersion about that value. For example, consider the monthly incomes, in rupees, of five households given below.

Monthly Incomes of Households

Households	A	B	C	D	E
Income (in Rs)	4500	6000	5500	3750	4700

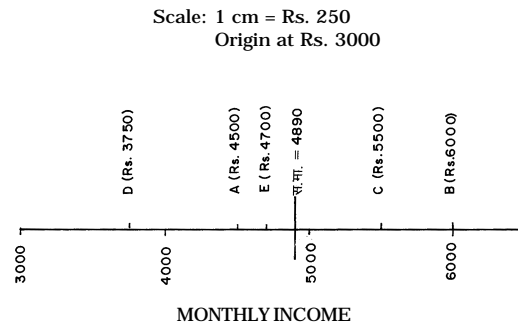


Fig. 7.1 Deviations from the arithmetic mean

The arithmetic mean of incomes is Rs 4,890; and deviations of incomes from Rs 4,890 are:

-390	1110	610	-1140	-190
------	------	-----	-------	------

As shown, in Fig. 7.1, the household D is farthest away to the left of the arithmetic mean and B is to the extreme right. Other households have incomes closer to the arithmetic mean, on either side. Similarly, we may choose some other value and examine the deviations from that value.

In general, some deviations are positive and others negative. Some are large and others small.

In order to get an over all idea about the amount of dispersion, we may consider an average of these deviations. But, while adding up deviations as they are we find that positive and negative deviations cancel out with each other, and their sum will be negligible. In fact, the sum of deviations from the arithmetic mean is exactly zero. Therefore, straight forward adding up deviations does not help. Alternatively, we may consider either the 'absolute deviations' (ignoring their signs), or, 'squared deviations'.

In the following sections, we define measures of dispersion in terms of

- (a) **absolute deviations, and**
- (b) **squares of deviations**

from the arithmetic mean of the values of the variable.

5. Mean Deviation

Suppose x_1, x_2, \dots, x_n are n values of the variable x and $\bar{x} = \frac{1}{n} \sum x_i$ is their arithmetic mean; the deviations from \bar{x} are given by

$$x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}$$

Some of these deviations will be positive and others negative. Ignoring the sign of deviations,

$$|x_1 - \bar{x}|, |x_2 - \bar{x}|, \dots, |x_n - \bar{x}|$$

are the absolute values of the deviations, where the two parallel bars $|\cdot|$ indicate that the absolute value is taken. (This is also called the 'modulus value'.)

The arithmetic mean of the absolute

deviations is called the **mean deviation**, or, **mean absolute deviation**. Thus,

$$\frac{1}{n} \sum |x_i - \bar{x}|$$

is the mean deviation of x about the arithmetic mean.

Example 2

The following table gives estimates of the population of 11 major districts of Rajasthan which have more than 20 lakh population.

Calculate the range, quartile deviation and mean deviation of the population.

Since the largest population is 49,21,000 of Jaipur and smallest is 20,13,000 of Ajmer, the range is 29,08,000. The units of measurement are 'number of persons'.

TABLE 7.2
Estimates of the Population (2001) of Major Districts of Rajasthan*

District	Population ('000)
Ajmer	2013
Alwar	2691
Bharatpur	2932
Ganganagar	2937
Jaipur	4921
Jodhpur	2275
Kota	2234
Nagaur	2606
S. Madhopur	2315
Sikar	2019
Udaipur	3389

* Major districts mean districts with population more than 20 lakh.

In order to calculate the quartiles, we arrange populations in the ascending order of magnitude:

2013, 2019, 2234, 2275, 2315, 2606, 2691, 2932, 2937, 3389, 4921

The quartile values are $Q_1 = 2234$, $Q_2 = 2606$ and $Q_3 = 2937$, i.e. $\frac{1}{4}(n+1)$ -th, $\frac{2}{4}(n+1)$ -th and $\frac{3}{4}(n+1)$ -th values, where $n=11$.

Therefore, the quartile deviation is equal to

$$\frac{Q_3 - Q_1}{2} = 351.5$$

The units of measurement are 'number of persons'.

The arithmetic mean of populations is equal to $\frac{30332}{11} \approx 2757.45$.

Table 7.3 shows the deviations from the arithmetic mean.

TABLE 7.3
Deviations from Arithmetic Mean

District	Population ('000)	Deviations from A.M. = 2757.45	Deviations from median = 2606
Ajmer	2013	-744.45	-593
Alwar	2691	-66.45	85
Bharatpur	2932	174.55	326
Ganganagar	2937	179.55	331
Jaipur	4921	2163.55	2315
Jodhpur	2275	-482.45	-331
Kota	2234	-523.45	-372
Nagaur	2606	-151.45	0
S. Madhopur	2315	-442.45	-291
Sikar	2019	-738.45	-587
Udaipur	3389	631.55	783

It should be verified that the sum of deviations from the arithmetic mean is zero. However, ignoring the sign of deviations,

$$\sum |x_i - \bar{x}| = 6298.35,$$

and, therefore, the mean deviation about the mean is 572.58 (approx.). The units of measurement are 'number of persons'.

We have calculated the median as 2606.

Consider the deviations from the median; then the mean deviation about the median is equal to

$$\frac{1}{n} \sum |x_i - \text{median}| = \frac{6014}{11} = 546.73.$$

The units of measurement are 'number of persons'.

We should note that mean deviation about the median is smaller than the mean deviation about the arithmetic mean.

In fact,

Mean deviation is the least when taken about the median.

The student may try several other values about which the mean deviation is taken and satisfy himself with the above result.

6. Calculation of Mean Deviation from Grouped Data

Example 3

The distribution of heights of 150 young ladies in a beauty contest was found to be as follows.

Height (in inches)	Frequency
62.0-63.5	12
63.5-65.0	20
65.0-66.5	28
66.5-68.0	18
68.0-69.5	19
69.5-71.0	20
71.0-72.5	30
72.5-74.0	3
Total	150

Calculate the range, quartile deviation and mean deviation about the arithmetic mean and median.

Since the largest height is 74" and smallest is 62"; the range is 12".

In order to calculate the quartiles we obtain the cumulative frequencies in the following table.

TABLE 7.4
Cumulative Frequencies

Height	Frequency	Cumulative frequencies
62.0-63.5	12	12
63.5-65.0	20	32
65.0-66.5	28	60
66.5-68.0	18	78
68.0-69.5	19	97
69.5-71.0	20	117
71.0-72.5	30	147
72.5-74.0	3	150
Total	150	

Q_1 is the $\left(\frac{1}{4}N\right)$ -th value, i.e. the 37.5 -th value. This lies in the class 65.0-66.5. Apply the formula to get Q_1 :

$$Q_1 = 65 + \frac{37.5 - 32}{28} \times 1.5 \approx 65.29.$$

$Q_2 = \text{Median}$ is the $\left(\frac{N}{2}\right)$ -th value, i.e. the 75-th value. This lies in the class 66.5 - 68.0. Apply the formula to get Q_2 :

$$\begin{aligned} \text{Median} &= Q_2 \\ &= 66.5 + \frac{75 - 60}{18} \times 1.5 = 67.75. \end{aligned}$$

Q_3 is the $\left(\frac{3}{4}N\right)$ -th value, i.e. the 112.5-th value. This lies in the class 69.5 - 71.0. Therefore, using the formula:

$$Q_3 = 69.5 + \frac{112.5 - 97}{20} \times 1.5 \approx 70.66.$$

Hence, the quartile deviation is

$$\frac{Q_3 - Q_1}{2} = 2.685.$$

The units of measurement are 'inches'.

The calculation of the arithmetic mean of heights are shown in Table 7.5.

TABLE 7.5
Calculation of Arithmetic Mean
(Assumed mean = 65.75 and common factor = 1.5)

Height (in inches)	Mid-value x	$d = x - 65.75$	$d' = d/1.5$	Frequency f	$f \times d'$ $f d'$
62.0-63.5	62.75	-3	-2	12	-24
63.5-65.0	64.25	-1.5	-1	20	-20
65.0-66.5	65.75	0	0	28	0
66.5-68.0	67.25	1.5	1	18	18
68.0-69.5	68.75	3	2	19	38
69.5-71.0	70.25	4.5	3	20	60
71.0-72.5	71.75	6	4	30	120
72.5-74.0	73.25	7.5	5	3	15
Total				150	207

Therefore,

$$\bar{d}' = \frac{\sum f_i d'_i}{N} = \frac{207}{150} = 1.38$$

$$\bar{x} = 65.75 + 1.5 \times 1.38 = 67.82.$$

We can calculate the mean deviation by **direct method** as

$$\frac{1}{N} \sum f_i |x_i - \bar{x}|;$$

or, alternatively by step-deviation method as

$$h \times \frac{1}{N} \sum f_i |d'_i - \bar{d}'|.$$

TABLE 7.6
Calculation of Mean Deviation
(By Direct Method)

Mid-value x	$ x - \bar{x} $	f	$f x - \bar{x} $
62.75	5.07	12	60.84
64.25	3.57	20	71.40
65.75	2.07	28	57.96
67.25	0.57	18	10.26
68.75	0.93	19	17.67
70.25	2.43	20	48.60
71.75	3.93	30	117.90
73.25	5.43	3	16.29
Total		150	400.92

$$\begin{aligned} \text{Mean deviation} &= \frac{1}{N} \sum f_i |x_i - \bar{x}| \\ &= \frac{400.92}{150} = 2.6728. \end{aligned}$$

TABLE 7.7
Calculation of Mean Deviation
(By Step Deviation Method)

(Assumed mean = 65.75 and common factor = 1.5)

$d' = \frac{x - 65.75}{1.5}$	$d' - \bar{d}'$	$ d' - \bar{d}' $	f	$f d' - \bar{d}' $
-2	-3.38	3.38	12	40.56
-1	-2.38	2.38	20	47.60
0	-1.38	1.38	28	38.64
1	-0.38	0.38	18	6.84
2	0.62	0.62	19	11.78
3	1.62	1.62	20	32.40
4	2.62	2.62	30	78.60
5	3.62	3.62	3	10.86
Total			150	267.28

Mean Deviation about the arithmetic mean

$$\begin{aligned} &= \frac{1}{N} \sum f_i |x_i - \bar{x}| = h \times \frac{1}{N} \sum f_i |d'_i - \bar{d}'| \\ &= 1.5 \times \frac{267.28}{150} = 2.6728. \end{aligned}$$

We should note the following characteristics of the mean deviation.

- It is rigidly defined.
- It depends on all values of the variable.
- It is based on absolute deviations from a central value.
- It is easy to understand.
- It involves harder calculations than the range and quartile deviation.
- It is amenable to algebraic treatment.
- The units of measurement of the mean deviation are the same as those of the variable (x).

7. Standard Deviation

Instead of taking absolute deviations from the arithmetic mean, we may square each deviation and obtain the arithmetic mean of squared deviations. This gives us the **variance** of the values.

The **positive** square root of the variance is called the **standard deviation** of the given values.

Suppose x_1, x_2, \dots, x_n are n values of x , their arithmetic mean is

$$\bar{x} = \frac{1}{N} \sum x_i;$$

and

$$x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}$$

are the deviations of the values of x from \bar{x} . Then,

$$\sigma^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$$

is the variance of x .

It can be shown that

$$\sigma^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 = \frac{1}{n} \sum x_i^2 - \bar{x}^2$$

which is more convenient for calculations.

It is conventional to represent the variance by the symbol σ^2 , where σ is a Greek alphabet 'sigma'. In fact, σ is small sigma and Σ is capital sigma. σ^2 is the square of σ .

Square root of the variance is the standard deviation, i.e.

$$\sigma = +\sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2}$$

We always take (+) sign before the square root. Thus, σ is always positive.

The units of measurement of the variance are expressed as (units of x)² and those of σ as 'units of x '

Example 4

Table 7.2 gives the estimates of the population (in thousands) of eleven major districts (with population more than 20 lakhs) of Rajasthan. Calculate the variance of the population.

In order to simplify calculations we consider estimates of the population nearest to ten lakhs.

Direct Method

Use the formula

$$\sigma_x^2 = \frac{1}{n} \sum x_i^2 - \bar{x}^2$$

where the suffix x to σ^2 indicates the variance of values of x (i.e. population).

The population (in lakhs) of districts are:

20, 27, 29, 29, 49, 23, 22, 26, 23, 20 and 34.

We obtain

$$\bar{x} = \frac{302}{11} \approx 27.45 \text{ and } \frac{\sum x_i^2}{n} \approx 816.91;$$

therefore,

$$\sigma_x^2 \approx 63.4075 \text{ and } \sigma_x \approx 7.96.$$

Assumed Mean Method

The variance is unaffected by the choice of the assumed mean.

Let the assumed mean be 20. Then the deviations, $d_i = x_i - 20$, are:

0, 7, 9, 9, 29, 3, 2, 6, 3, 0 and 14.

We obtain

$$\bar{d} = \frac{\sum d_i}{n} = \frac{82}{11} \approx 7.45 \text{ and}$$

$$\frac{\sum d_i^2}{n} = \frac{1306}{11} \approx 118.73;$$

and, therefore, the variance of d_i 's is

$$\sigma_d^2 = \frac{\sum d_i^2}{n} - \bar{d}^2 \approx 63.23.$$

This is the same as the variance of x (σ_x^2) obtained above (apart from rounding off errors). The standard deviation of x is the same as of d :

$$\sigma_x = +\sqrt{63.23} \approx 7.95.$$

Step Deviation Method

Firstly, choose a convenient common factor h , and divide each deviation from the assumed mean by h and obtain

$d'_i = \frac{d_i}{h}$ for $i = 1, \dots, n$. Then

$$\bar{d}' = \frac{\sum d'_i}{n}$$

and the variance of d'_1, \dots, d'_n is

$$\sigma_{d'}^2 = \frac{\sum d_i'^2}{n} - (\bar{d}')^2$$

The variance of x is

$$\sigma_x^2 = h^2 \sigma_d^2$$

so that

$$\sigma_x = h \sigma_d$$

In the present example, using the step deviation method does not help because deviations from the assumed mean are small enough.

8. Calculation of Standard Deviation from Grouped Data

We use the frequency distribution given in Example 3 to illustrate the calculation of standard deviation from grouped data.

Example 5

The frequency distribution of heights of 150 young ladies in a beauty contest is given below.

Height (in inches)	Frequency
62.0-63.5	12
63.5-65.0	20
65.0-66.5	28
66.5-68.0	18
68.0-69.5	19
69.5-71.0	20
71.0-72.5	30
72.5-74.0	3
Total	150

Calculate the standard deviation of heights.

Direct Method

Calculations for standard deviation are as shown in Table 7.8.

TABLE 7.8
Calculation of Standard Deviation
(By Direct Method)

Height (in inches) (1)	Mid-value x (2)	Frequency f (3)	$f \times x$ (4)	$f \times x^2$ (5)
62.0-63.5	62.75	12	753.00	47250.75
63.5-65.0	64.25	20	1285.00	82561.25
65.0-66.5	65.75	28	1841.00	121045.75
66.5-68.0	67.25	18	1210.50	81406.125
68.0-69.5	68.75	19	1306.25	89804.6875
69.5-71.0	70.25	20	1405.00	98701.25
71.0-72.5	71.75	30	2152.50	154441.875
72.5-74.0	73.25	3	219.75	16096.6875
Total		150	10173.00	691308.375

Thus,

$$\bar{x} = \frac{\sum f_i x_i}{N} = \frac{10173}{150} = 67.82$$

and

$$\frac{\sum f_i x_i^2}{N} = 4608.7225$$

where $N = \sum f_i = 150$. Therefore, the variance of x (heights) is

$$\sigma_x^2 = \frac{\sum f_i x_i^2}{N} - \bar{x}^2 = 9.1701 \text{ (inches)}^2$$

and standard deviation is 3.03 (inches) approx.

The steps of calculation of the variance are as follows:

Step 1. Obtain the mid-values (x_i) of class intervals (Col. 2)

Step 2. Multiply frequencies (f_i) in Col. 3, with corresponding values of x_i , in Col. (2), to obtain $f_i x_i$ (in Col. (4))

Step 3. Obtain the total of Col. (4) to get $\sum f_i x_i$ and

$$\bar{x} = \frac{\sum f_i x_i}{N}, N = \sum f_i = 150$$

Step 4. Multiply $f_i x_i$, in Col. (4) with corresponding x_i in Col. (2), to obtain $f_i x_i^2$ [in Col. (5)].

Step 5. Obtain the total of Col. (5) to get

$$\Sigma f_i x_i^2 \text{ and } \frac{\Sigma f_i x_i^2}{N}.$$

Step 6. Obtain

$$\sigma_x^2 = \frac{\Sigma f_i x_i^2}{N} - \bar{x}^2.$$

The units of measurement of variance of heights (x_i 's) are expressed as (units of x)², and the units of measurement of the standard deviation of x are the same as those of x .

Assumed Mean Method

Calculations for the standard deviation by assumed mean method are shown in the following table.

TABLE 7.9
Calculation of Standard Deviation
(By Assumed Mean Method)
(Assumed Mean = 67.25)

Mid-value x (1)	$d = x - 67.25$ (2)	f (3)	$f \times d$ (4)	$f \times d^2$ (5)
62.75	-4.5	12	-54	243
64.25	-3.0	20	-60	180
65.75	-1.5	28	-42	63
67.25	0	18	0	0
68.75	1.5	19	28.5	42.75
70.25	3.0	20	60	180
71.75	4.5	30	135	607.5
73.25	6.0	3	18	108
Total		150	85.5	1424.25

Therefore,

$$\bar{d} = \frac{\Sigma f_i d_i}{N} = \frac{85.5}{150} = 0.57$$

$$\frac{\Sigma f_i d_i^2}{N} = \frac{1424.25}{150} = 9.495$$

$$\sigma_d^2 = \frac{\Sigma f_i d_i^2}{N} - \bar{d}^2 = 9.1701 \text{ (inches)}^2.$$

It should be noted that

$$\sigma_x^2 = \sigma_d^2 = 9.1701;$$

i.e. **the variance of x is unaffected by the choice of the origin.**

The standard deviation of x is the positive square root of the variance, i.e.

$$\sigma_x \approx 3.03 \text{ (inches)}.$$

The steps of calculation of the variance of x should be carefully noted.

Step 1. The assumed mean is arbitrarily chosen as 67.25.

Step 2. Obtain $d_i = x_i - 67.25$ in Col. (2), where x_i 's are mid-values of different classes.

Step 3. Obtain $f_i d_i$ in Col. (4) and $\Sigma f_i d_i$

$$\text{and } \bar{d} = \frac{\Sigma f_i d_i}{N}.$$

Step 4. Obtain $f_i d_i^2$, in Col. (5) by multiplying the corresponding values in Col. (2), and Col. (4); and obtain $\Sigma f_i d_i^2$

Step 5. Obtain $\sigma_d^2 = \frac{\Sigma f_i d_i^2}{N} - \bar{d}^2$ as the variance of d .

Step 6. The variance of d is equal to variance of x .

Step 7. The standard deviation of x is equal to the positive square root of the variance of x .

Step Deviation Method

Calculations for standard deviation by step deviation method are shown below.

TABLE 7.10
Calculation of Standard Deviation
(By Step Deviation Method)
(Assumed mean = 67.25 and common factor = 1.5)

Mid-value x	$d = x - 67.25$	$d' = d \div 1.5$	f	$f \times d'$	$f \times d'^2$
(1)	(2)	(3)	(4)	(5)	(6)
62.75	-4.5	-3	12	-36	108
64.25	-3.0	-2	20	-40	80
65.75	-1.5	-1	28	-28	28
67.25	0	0	18	0	0
68.75	1.5	1	19	19	19
70.25	3.0	2	20	40	80
71.75	4.5	3	30	90	270
73.25	6.0	4	3	12	48
Total			150	57	633

We obtain

$$\bar{d}' = \frac{\sum f_i d'_i}{N} = \frac{57}{150} = 0.38$$

$$\frac{\sum f_i d_i'^2}{N} = \frac{633}{150} = 4.22;$$

therefore,

$$\sigma_{d'}^2 = \frac{\sum f_i d_i'^2}{N} - \bar{d}'^2 = 4.0756$$

and

$$\sigma_x^2 = h^2 \sigma_{d'}^2 = 1.5^2 \times 4.0756 = 9.1701 \text{ (inches)}^2.$$

The standard deviation of x is the positive square root of the variance of x ; thus,

$$\sigma_x \approx 3.03 \text{ (inches)}.$$

The steps of calculation of the standard deviation are as follows.

Step 1. We choose the assumed mean as 67.25 and common factor 1.5.

Step 2. In Col. (2) we obtain $d_i = x_i - 67.25$

and in Col.(3) $d'_i = \frac{d_i}{1.5}$ where

x_i 's are mid-value of classes, shown in Col. (1)

Step 3. Obtain $f_i d'_i$ in Col. (5) by multiplying frequencies (f_i 's) with the corresponding values of d'_i 's

Step 4. Obtain $\bar{d}' = \frac{\sum f_i d'_i}{N}$

Step 5. Obtain $f_i d_i'^2$ in Col. (6) by multiplying corresponding values in Col.(3) and Col.(5).

Step 6. Obtain $\frac{\sum f_i d_i'^2}{N}$ and the variance of d i.e., $\sigma_{d'}^2 = \frac{\sum f_i d_i'^2}{N} - \bar{d}'^2$

Step 7. We note that

$$\sigma_x^2 = h^2 \sigma_{d'}^2$$

is the variance of x and the positive square root of the variance is the standard deviation σ_x .

It should be observed that the calculations are simplified a great deal by using the step deviation method.

9. Comparison of Alternative Measures of Dispersion

Let us compare the characteristics of the four measures of dispersion discussed above.

(i) *Rigidly Defined*

All four measures, the range, quartile deviation, mean deviation and standard deviation are rigidly defined. There is no vagueness in their definition.

(ii) *Ease of Calculation*

The range is the easiest to calculate.

Quartile deviation requires calculation of the upper and lower quartiles, but that is also easy enough. However, the mean deviation and standard deviation require a little more systematic calculations. They, too, are easy.

(iii) *Simple Interpretation*

All measures of dispersion are easy to interpret. While the range and quartile deviation measure dispersion in a general way, the mean deviation and standard deviation measure dispersion in terms of deviations from a central value. Thus, the mean deviation and standard deviation give a better idea about the dispersion of values within the range.

(iv) *Based on All Values*

The range and quartile deviation do not depend on all values; whereas, the mean deviation and standard deviation use all values of the variable. The range is affected the most by extreme values.

(v) *Amenable to Algebraic Treatment*

The standard deviation is perhaps the easiest for analytical work. Other measures also can be dealt with analytically but derivations are harder.

10. Relative Measure of Dispersion – The Coefficient of Variation

The measures of dispersion discussed above are **absolute** measures. All of them are measured in the same units as those of the variable considered. Thus, if we are considering variation of incomes (in rupees), the range, quartile deviation, mean deviation and standard deviation of incomes are all in rupees. For heights measured in cm. they are all in terms of cm., etc. This feature of measures of dispersion may create difficulty, if we

want to compare dispersion in two sets of values, which have

- (i) different central values, and/or
- (ii) different units of measurement.

For example, suppose we measure income in paise instead of rupees. The dispersion will increase 100 times. This may lead us to think that dispersion of incomes has increased when it might have actually remained unchanged in every respect.

Also, comparison of dispersion in two sets of values is difficult, if the units of measurement are not the same. For example, it is difficult to compare the dispersion of incomes in India with those in USA. The latter are in terms of dollars, while the former are in rupees.

In order to overcome this difficulty, it is desirable to eliminate the units of measurement. This can be done if we use a **relative measure of dispersion**, which is a **pure number** (and does not depend on units of measurement). The relative measure of dispersion is called **coefficient of variation**. We might express this simply as a ratio, or, express it in percentage terms.

The most commonly used coefficient of variation is the ratio of standard deviation to arithmetic mean. Symbolically, $\frac{\sigma}{\bar{x}}$ is the coefficient of variation, where σ is the standard deviation and \bar{x} is the arithmetic mean of the variable. The ratio is a pure number. We may also express it in percentage terms as $\frac{\sigma}{\bar{x}} \times 100$.

We may also compute the 'coefficient of variation' as

$$\frac{x_{\max} - x_{\min}}{x_{\max} + x_{\min}} = \frac{\text{Range}}{x_{\max} + x_{\min}}.$$

if we are using the range as a measure of dispersion; where x_{\max} is the largest and x_{\min} is the smallest value of the variable; or, as

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

if we are using quartile deviation as measure of dispersion.

Similarly, using mean deviation as a measure of dispersion,

$$\frac{\text{mean deviation}}{\text{arithmetic mean}}$$

or

$$\frac{\text{mean deviation}}{\text{median}}$$

are also coefficients of variation.

EXERCISES

1. Illustrate the meaning of the term 'dispersion' with examples.
2. Why should we measure dispersion about some particular value? Do the range and quartile deviation measure dispersion about some value?
3. What are the properties of a good measure of dispersion?
4. What are the four alternative measures of 'absolute' dispersion? Discuss their properties.
5. 'The coefficient of variation is a **relative measure** of dispersion'. We may calculate coefficient of variation using any of the measures of dispersion, such as, range, quartile deviation, mean deviation and standard deviation. Illustrate the use of coefficient of variation in these cases.
6. Calculate the arithmetic mean and standard deviation of the following values
 - (i) without grouping
 - (ii) grouping the values in classes 140-145, 145-150, ...
 - (iii) grouping them in classes 140-150, 150-160, ...

140	143	143	146	146
146	154	156	159	162
164	164	166	166	167
167	168	168	169	169
169	171	175	175	176
176	178	180	182	182
182	182	182	183	184
186	188	190	190	191
191	192	195	202	227

7. In a certain examination 15 students of a class secured the following percentage of marks in mathematics and economics.

<i>Mathematics</i>				
31	25	78	65	80
56	58	42	39	54
60	63	58	45	72

<i>Economics</i>				
42	39	45	49	58
56	49	44	60	62
44	50	51	38	40

Using R (the range), Mean Deviation (about the arithmetic mean) and standard deviation (s.d.) of marks, examine if the performance of the students has larger variation in mathematics than in economics.

Would your conclusion change, if you used a relative measure of dispersion?

8. The distribution of the cost of production (in rupees) of a quintal of wheat in 50 farms is as follows:

<i>Cost (in rupees)</i>	<i>Number of farms</i>
40-50	3
50-60	6
60-70	12
70-80	18
80-90	9
90-100	2
Total	50

- (a) Calculate the variance
- by direct method
 - by step deviation method
- and compare your results with the mean deviation about the arithmetic mean.
- (b) Calculate the coefficient of variation by using
- the standard deviation of costs, and
 - the mean deviation of costs about the arithmetic mean
- and compare the two. What is your conclusion about variation of costs?
9. A variable x takes the following values:
7, 9, 18, 11, 10, 8, 17, 13, 11, 16
- (a) Calculate
- the arithmetic mean (\bar{x}) of x
 - the standard deviation of the values of x
 - the mean deviation of the values of x about \bar{x}
- (b) Also, calculate
- $\Sigma(x_i - 10)^2$
 - $\Sigma |x_i - \text{median}|$

- (c) Examine, if
- $\Sigma(x_i - 10)^2 > \Sigma(x_i - \bar{x})^2$
 - $\Sigma|x_i - \bar{x}| > \Sigma|x_i - \text{median}|$
- 10.** In a survey it was found that the average per capita consumption of milk is 0.5 litre per day and the coefficient of variation (expressed in percentage terms) is 20. What is the **variance** of per capita milk consumption?
- 11.** A study of certain examination results of 1000 students in the year 2000 gave average marks secured as 50 per cent with a standard deviation of 3 per cent. A similar study in 2001 revealed average marks secured and standard deviation as 55 per cent and 5 per cent, respectively. Have the results improved over the year?
- 12.** Suppose the variable x takes the integral values from 1 to 10. Calculate.
- the arithmetic mean
 - the standard deviation
 - the mean deviation about the arithmetic mean
 - the mean deviation about the median of the values of x .
- Also, calculate the coefficient of variation of the values of x , using
- standard deviation
 - mean deviation
- 13.** 'The standard deviation of heights measured in inches will be larger than the standard deviation of heights measured in feet for the same group of individuals.' Comment on the validity, or, otherwise of this statement with appropriate explanation.
- 14.** The following is the frequency distribution of the population of 55 villages in India.

<i>Population</i>	<i>Number of villages</i>
Less than 200	5
200 – 400	14
400 – 1000	21
1000 – 2000	9
2000 – 5000	3
More than 5000	3
Total	55

Calculate a suitable measure of dispersion and justify your choice.

CHAPTER 8

Coefficient of Correlation

In the earlier chapters, we studied **univariate** distributions, where observations were given on a **single** variable. We obtained a measure of central tendency (like the arithmetic mean, median, etc.) and a measure of dispersion (like the standard deviation, mean deviation, etc.) of the set of values of the variable.

Now, suppose we have observations on two variables, X and Y for several individuals. We have a bivariate distribution of X and Y.

We can still calculate the central value (arithmetic mean, median, etc.) and dispersion (standard deviation, mean deviation, etc.) of each variable, X and Y, separately. However, we would also like to know, if there is any **association** between the values of the two variables. We would like to know, for example, how does the value of one of the variables change, if the value of the other increases, or, decreases by a certain amount. Is the change in the same direction and in the same proportion; or, the change is more

or less than proportionate and in the reverse direction?

A numerical measure of association between two variables is given by Karl Pearson's **coefficient of correlation**.

In the first instance, however, we may study the **form of association** between two variables with the help of a **scatter diagram**.

1. Scatter Diagram

In a scatter diagram, we plot the values of the two variables, as a set of points, on a graph paper. The cluster of points, so obtained, is called the **scatter diagram**. Let us illustrate this with the following example.

Example 1

Suppose we have observations on
(i) the monthly income (X), and
(ii) the total monthly expenditure on food (Y),
in rupees, of five rural households, as shown in Table 8.1.

TABLE 8.1
Monthly Income and Expenditure on Food of Five Rural Households

Variable	Households				
	1	2	3	4	5
Income (X) in rupees	550	600	800	700	650
Expenditure on food (Y) (in rupees)	400	450	550	550	400

We note that the income of the first household is Rs 550 per month and its expenditure on food per month is Rs 400. We may plot this as a point (X,Y) on the graph paper, where $X=550$ and $Y=400$. We measure 550 along the X-axis and 400 along the Y-axis. Similarly, for the second household $X=600$ and $Y=450$; so that the coordinates of the second point are (600,450). We measure 600 along the X-axis and 450 along the Y-axis, and so on.

The cluster of points is shown in Fig. 8.1. This is called the scatter diagram.

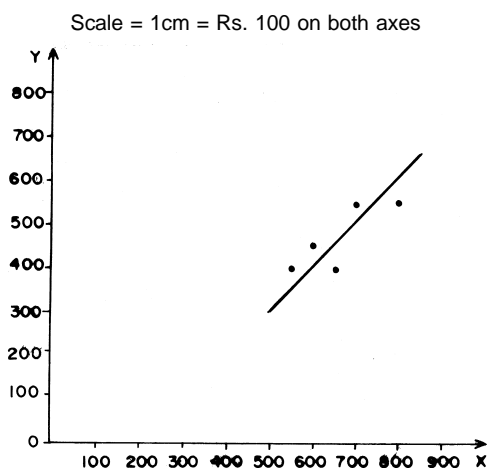


Fig.8.1 Scatter diagram (Data in Table 8.1)

It is clear from the scatter diagram that the points tend to cluster about an upward sloping straight line. In other words, the expenditure on food (Y) increases as the income (X) of the household increases.

In general, if the straight line is sloping upward, the values of X and Y change in the same direction (i.e. an

increase in the value of X is accompanied with an increase in the value of Y). Otherwise, if the straight line is sloping downwards, an increase in the value of X is accompanied with a decrease in the value of Y.

It is the **slope** of the straight line (which depends on the angle that the straight line makes with the X-axis and is equal to $\frac{dy}{dx}$) that determines the **rate of change in the value of Y for a unit change in the value of X**.

Let us illustrate this with the following example.

Example 2

In the following table, we are given the price (p) and quantity demanded (q) of potatoes in a certain wholesale market in different seasons of a year.

TABLE 8.2
Demand for Potatoes

Price (p) (Rs/kg)	5	6	7	8	9
Quantity (q) (Quintals)	10	9	7	5	4

Draw a scatter diagram, and determine the form of association between q and p.

The scatter diagram is shown in Fig. 8.2. We measure price (p) along the X-axis and quantity demanded (q) along the Y-axis.

There is an inverse relationship between q and p, because the points in the scatter diagram tend to cluster about a downward sloping straight line.

The straight line makes an angle of approximately 50° , as shown in

Fig. 8.2, with the X-axis. Therefore, the decrease in quantity demanded is **almost** in the same proportion as the increase in price.

In general, if the straight line makes an angle of 45° with the X-axis, the change in the value of Y is **exactly** in the same proportion as the change in the value of X.

The change in the value of Y is more than proportionate to the change in the value of X, if the angle that the straight line makes with the X-axis is greater than 45° . See the hypothetical Figures 8.3, 8.4, 8.5 and 8.6. Figure 8.7 shows that the value of Y does not change, at all, if the value of X increases. (An example of this last case would be the relationship between the quantity demanded of a commodity like salt with its price). Figure 8.8 shows a non-linear relationship between X and Y; and Fig. 8.9 shows no relation.

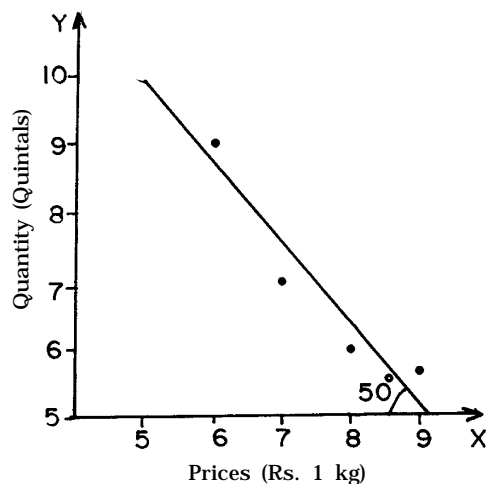


Fig.8.2 Showing almost equal change

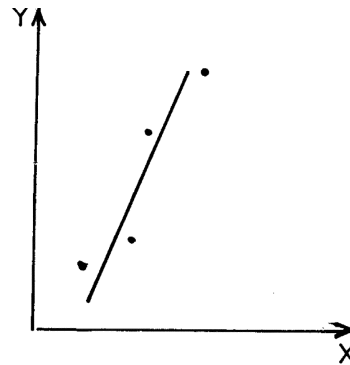


Fig.8.3 Showing more than proportionate change

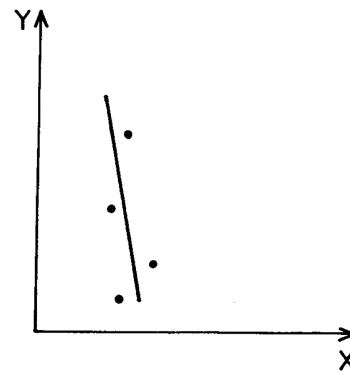


Fig.8.4 Showing more than proportionate change

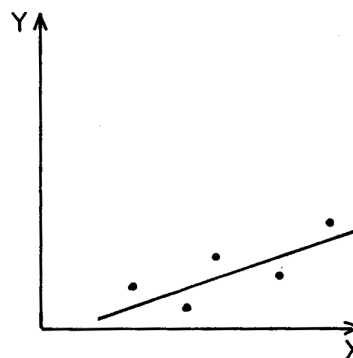


Fig.8.5 Showing less than proportionate change

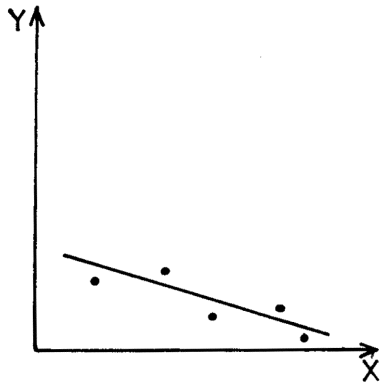


Fig.8.6 Showing less than proportionate change

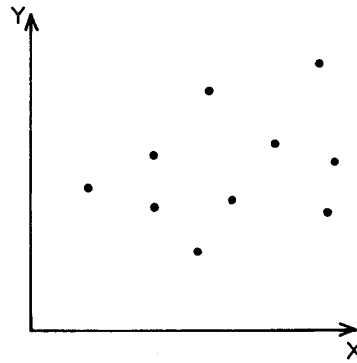


Fig.8.9 No relation

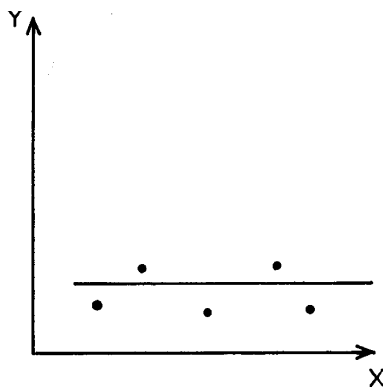


Fig.8.7 Showing no change

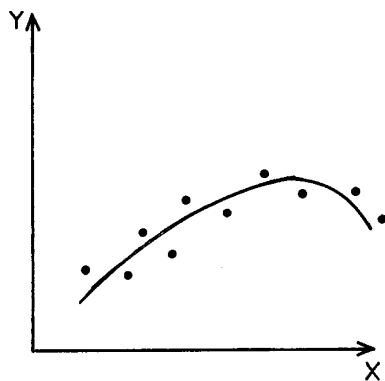


Fig.8.8 Non-linear relationship

2. Merits and Demerits of a Scatter Diagram

Merits

- It is easy to draw a scatter diagram.
- It is an easy first step in determining the form of relationship (linear or non-linear) between two variables.
- In case of linear relationship between Y and X, it gives a clear visual picture of the proportionate change in the value of Y for a change in the value of X.

Demerits

- The strength or the degree of association between two variables cannot be determined in numerical terms from a scatter diagram.
- Scatter diagram does not indicate the direction of causation. It does not tell, if Y causes X or X causes Y.
- It is not possible to draw a scatter diagram on a graph paper, if there are more than two variables.

3. Karl Pearson's Coefficient of Correlation — Product Moment Correlation

If the relationship between two variables X and Y is **linear**, we can determine the **degree of association** between them, **numerically**, with the help of Karl Pearson's coefficient of correlation. This is also called the **product moment correlation**.

Suppose, X_1, X_2, \dots, X_n are values of X, and Y_1, Y_2, \dots, Y_n are the corresponding values of Y. The arithmetic means of X and Y are

$$\bar{X} = \frac{1}{n} \sum X \text{ and } \bar{Y} = \frac{1}{n} \sum Y$$

and their variances are

$$\sigma_x^2 = \frac{1}{n} \sum (X - \bar{X})^2 = \frac{1}{n} \sum X^2 - X^2, \text{ and}$$

$$\sigma_y^2 = \frac{1}{n} \sum (Y - \bar{Y})^2 = \frac{1}{n} \sum Y^2 - Y^2.$$

The standard deviations, σ_x and σ_y , of X and Y, respectively, are the positive square roots of their variances.

Let us write

$$x = X - \bar{X} \text{ and } y = Y - \bar{Y}$$

for deviations of X and Y from their respective means.

The **product moment correlation between X and Y** is defined as

$$r = \frac{\sum xy}{n\sigma_x\sigma_y}$$

which can also be expressed as

$$r = \frac{\sum XY - \frac{1}{n} (\sum X)(\sum Y)}{\sqrt{\sum X^2 - \frac{1}{n} (\sum X)^2} \sqrt{\sum Y^2 - \frac{1}{n} (\sum Y)^2}}$$

or,

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{n\sum X^2 - (\sum X)^2} \sqrt{n\sum Y^2 - (\sum Y)^2}}$$

Let us illustrate calculation of the correlation coefficient by using these formulae.

Example 3

Calculate the product moment correlation between 'quantity demanded' and 'price' of potatoes for data given in Example 2.

Calculations are shown in Table 8.3 by direct method and by using deviations from actual means in Table 8.4.

TABLE 8.3
Calculation of Correlation Coefficient
(Direct Method)

Price (p)	Quantity (q)	p^2	q^2	pq
5	10	25	100	50
6	9	36	81	54
7	7	49	49	49
8	5	64	25	40
9	4	81	16	36
$\Sigma p=35$	$\Sigma q=35$	$\Sigma p^2=255$	$\Sigma q^2=271$	$\Sigma pq=229$

We use the formula.

$$r = \frac{\Sigma pq - \frac{1}{n} (\Sigma p)(\Sigma q)}{\sqrt{\Sigma p^2 - \frac{1}{n} (\Sigma p)^2} \sqrt{\Sigma q^2 - \frac{1}{n} (\Sigma q)^2}}$$

This is called the **direct method**.

Substituting the values, we get

$$r = \frac{229 - \frac{1}{5} \times 35 \times 35}{\sqrt{255 - \frac{1}{5} (35)^2} \sqrt{271 - \frac{1}{5} (35)^2}} = -0.99$$

Example 4

From the data in Example 2, calculate the product moment correlation coefficient between p and q by calculating deviations from their respective means.

The calculations are shown in Table 8.4 below.

$$\text{We have } \bar{p} = \frac{1}{5}\Sigma p = 7 \text{ and } \bar{q} = \frac{1}{5}\Sigma q = 7.$$

Using the formula

$$r = \frac{\Sigma(p - \bar{p})(q - \bar{q})}{\sqrt{\Sigma(p - \bar{p})^2} \sqrt{\Sigma(q - \bar{q})^2}} = \frac{-16}{\sqrt{10} \sqrt{26}} = -0.99$$

This is called **actual mean method**.

Example 5

The data on

X = monthly income (in rupees),
and

Y = monthly expenditure (in rupees)

of five rural households are given in Table 8.1.

Calculate the product moment correlation between X and Y by **direct method**.

The calculations are shown in Table 8.5 below.

TABLE 8.4
Calculation of Correlation
(Deviations from actual means)

P	q	$p - \bar{p}$	$q - \bar{q}$	$(p - \bar{p})^2$	$(q - \bar{q})^2$	$(p - \bar{p})(q - \bar{q})$
5	10	-2	3	4	9	-6
6	9	-1	2	1	4	-2
7	7	0	0	0	0	0
8	5	1	-2	1	4	-2
9	4	2	-3	4	9	-6
$\Sigma p = 35$	$\Sigma q = 35$			$\Sigma(p - \bar{p})^2 = 10$	$\Sigma(q - \bar{q})^2 = 26$	$\Sigma(p - \bar{p})(q - \bar{q}) = -16$

TABLE 8.5
Calculation of Correlation
(Direct method)

X	Y	X^2	Y^2	XY
550	400	302500	160000	220000
600	450	360000	202500	270000
800	550	640000	302500	440000
700	550	490000	302500	385000
650	400	422500	160000	260000
$\Sigma X = 3300$	$\Sigma Y = 2350$	$\Sigma X^2 = 2215000$	$\Sigma Y^2 = 1127500$	$\Sigma XY = 1575000$

Using the formula

$$\begin{aligned}
 r &= \frac{\sum XY - \frac{1}{n} (\sum X)(\sum Y)}{\sqrt{\sum X^2 - \frac{1}{n} (\sum X)^2} \sqrt{\sum Y^2 - \frac{1}{n} (\sum Y)^2}} \\
 &= \frac{1575000 - \frac{1}{5} (3300)(2350)}{\sqrt{2215000 - \frac{1}{5} (3300)^2} \sqrt{1127500 - \frac{1}{5} (2350)^2}} \\
 &= \frac{24000}{\sqrt{37000 \times 23000}} = \frac{24000}{29171.9} \\
 &= 0.82
 \end{aligned}$$

Example 6

The data on X (monthly income in rupees) and Y (monthly expenditure on food in rupees) of five households are given in the Table 8.1.

Calculate the product moment correlation between X and Y by taking **deviations from actual means**.

The calculations are shown in Table 8.6 below.

$$\bar{X} = \frac{1}{n} \sum X = 660, \bar{Y} = \frac{1}{n} \sum Y = 470$$

Using the formula

$$\begin{aligned}
 r &= \frac{\sum xy}{\sqrt{\sum x^2} \sqrt{\sum y^2}} = \frac{24000}{\sqrt{37000 \times 23000}} \\
 &= \frac{24000}{29171.9} = 0.82
 \end{aligned}$$

4. Use of Step Deviation Method to Calculate Correlation Coefficient

The calculation of correlation coefficient can be simplified a great deal, if

- we calculate deviations of X and Y from **assumed means**, and
- divide the deviations by a convenient value, or, the common factor.

We will illustrate this by using the examples given above.

This is also called **change of origin** and **change of scale** of the values of X and Y.

The following is an important result.

The value of the correlation coefficient is unaffected by the change of origin and change of scale of X and Y.

TABLE 8.6
Calculation of Correlation
(Deviations from actual means)

X	Y	$X - \bar{X} = x$	$Y - \bar{Y} = y$	x^2	y^2	xy
550	400	-110	-70	12100	4900	7700
600	450	-60	-20	3600	400	1200
800	550	140	80	19600	6400	11200
700	550	40	80	1600	6400	3200
650	400	-10	-70	100	4900	700
$\sum X = 3300$	$\sum Y = 2350$			37000	23000	24000

Example 7

The data on

X = monthly income (in rupees), and
Y = monthly expenditure (in rupees)
of five rural households are given in
Table 8.1.

Calculate the product moment correlation between X and Y.

The calculations by step deviation method are shown in Table 8.7.

We choose '**assumed mean**' for X as 600 and for Y as 550, and divide deviations by common factor 50; thus

$$d'_x = \frac{X - 600}{50} \quad \text{and} \quad d'_y = \frac{Y - 550}{50}$$

Using the formula

$$r = \frac{n \sum d'_x d'_y - \sum d'_x \sum d'_y}{\sqrt{n \sum d_x'^2 - (\sum d'_x)^2} \sqrt{n \sum d_y'^2 - (\sum d'_y)^2}}$$

$$= \frac{5 \times 0 - 6 \times (-8)^2}{\sqrt{5 \times 22 - 6^2} \sqrt{5 \times 22 - (-8)^2}}$$

$$= 0.82$$

Example 8

Calculate the correlation coefficient between

X = years of schooling, and
Y = annual yield per acre in rupees
for the data given below in Table 8.8

TABLE 8.8

X	0	2	4	6	8	10	12	14	14	16
Y	4	4	6	10	10	8	12	10	8	6

For comparison, we calculate the coefficient of correlation by

- direct method (in Table 8.9),
- by taking deviations from actual means (in Table 8.10),
- by step deviation method (in Table 8.11).

TABLE 8.9

Calculation of the Coefficient of Correlation (Direct method)

X	Y	X ²	Y ²	XY
0	4	0	16	0
2	4	4	16	8
4	6	16	36	24
6	10	36	100	60
8	10	64	100	80
10	8	100	64	80
12	12	144	144	144
14	10	196	100	140
14	8	196	64	112
16	6	256	36	96
$\Sigma X = 86$		$\Sigma Y = 78$	$\Sigma X^2 = 1012$	$\Sigma Y^2 = 676$
		$\Sigma XY = 744$		

TABLE 8.7

Calculation of Correlation Coefficient (Step deviation method)

X	Y	$d'_x = \frac{X - 600}{50}$	$d'_y = \frac{Y - 550}{50}$	$d_x'^2$	$d_y'^2$	$d'_x d'_y$
550	400	-1	-3	1	9	3
600	450	0	-2	0	4	0
800	550	4	0	16	0	0
700	550	2	0	4	0	0
650	400	1	-3	1	9	-3
		$\Sigma d'_x = 6$	$\Sigma d'_y = -8$	$\Sigma d_x'^2 = 22$	$\Sigma d_y'^2 = 22$	$\Sigma d'_x d'_y = 0$

Using the formula

$$r = \frac{\sum XY - n\bar{X}\bar{Y}}{\sqrt{\sum X^2 - n\bar{X}^2} \sqrt{\sum Y^2 - n\bar{Y}^2}}$$

$$= \frac{744 - 10 \times 8.6 \times 7.8}{\sqrt{1012 - 10 \times (8.6)^2} \sqrt{676 - 10 \times (7.8)^2}}$$

$$= \frac{73.2}{\sqrt{272.4} \sqrt{67.6}} = \frac{73.2}{135.7} = 0.54$$

TABLE 8.10
Calculation of the Coefficient of Correlation
(Deviations from actual means)

$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	y^2	xy
-8.6	-3.8	73.96	14.44	32.68
-6.6	-3.8	43.56	14.44	25.08
-4.6	-1.8	21.16	3.24	8.28
-2.6	2.2	6.76	4.84	-5.72
-0.6	2.2	0.36	4.84	-1.32
1.4	0.2	1.96	0.04	0.28
3.4	4.2	11.56	17.64	14.28
5.4	2.2	29.16	4.84	11.88
5.4	0.2	29.16	0.04	1.08
7.4	-1.8	54.76	3.24	-13.32
$\sum x = 0$	$\sum y = 0$	$\sum x^2 = 272.4$	$\sum y^2 = 67.60$	$\sum xy = 73.20$

Using the formula

$$r = \frac{\sum xy}{\sqrt{\sum x^2} \sqrt{\sum y^2}} = \frac{73.20}{\sqrt{272.4} \sqrt{67.6}} = 0.54$$

TABLE 8.11
Calculation of Coefficient of Correlation
(Step deviation method)
Assumed mean of X is 10 and of Y is 8; Common Factor = 2

X	$d_x' = \frac{X - 10}{2}$	$d_x'^2$	Y	$d_y' = \frac{Y - 8}{2}$	$d_y'^2$	$d_x' d_y'$
0	-5	25	4	-2	4	10
2	-4	16	4	-2	4	8
4	-3	9	6	-1	1	3
6	-2	4	10	1	1	-2
8	-1	1	10	1	1	-1
10	0	0	8	0	0	0
12	1	1	12	2	4	2
14	2	4	10	1	1	2
14	2	4	8	0	0	0
16	3	9	6	-1	1	-3
$\sum d_x' = -7$		$\sum d_x'^2 = 73$	$\sum d_y' = -1$		$\sum d_y'^2 = 17$	$\sum d_x' d_y' = 19$

$$r_{xy} = r = \frac{\sum d_x' d_y'}{\sqrt{\sum d_x'^2} \sqrt{\sum d_y'^2}} = \frac{19}{\sqrt{73 \times 17}} = 0.54$$

5. Certain Properties of the Correlation Coefficient

The following properties of the correlation coefficient should be noted.

- (i) **The correlation coefficient is never less than -1 and is never greater than +1:**

$$-1 \leq r \leq +1$$

- (ii) **A negative value of r indicates an inverse relationship between the variables X and Y.** That is, an increase in the value of X is associated with a decrease in the value of Y; and vice versa. For example, the quantity demanded and price are inversely related.

- (iii) **If an increase in the value of X is associated with an increase in the value of Y, r must be positive.** For example, income and quantity demanded may be positively related. Likewise, if a decrease in the value of X is associated with a decrease in the value of Y, then r must be positive.

- (iv) **If $r=0$, the two variables X and Y are uncorrelated.** In this case, a change in the value of one variable is not **linearly** related with the value of the other.
- (v) **If $r = +1$, or, $r = -1$, the two variables, X and Y are perfectly linearly related.** The relationship between X and Y is **exact**.
- (vi) **A high value of r (close to $+1$, or, -1) indicates strong linear relation between X and Y.**
- (vii) **A value of r close to zero (positive or negative) indicates weak linear relation between X and Y.**
- (viii) **The value of r is unaffected by change of origin and change of scale.**

Suppose,

$$d'_x = \frac{X - A}{B} \quad \text{and} \quad d'_y = \frac{Y - C}{D}$$

where A and C are **assumed means** and B and D are any values (or, common factors), then

$$r_{d'_x d'_y} = r_{XY}$$

- (ix) **A limitation of the correlation coefficient is that it does not indicate the direction of causation, i.e. it says nothing about, if Y causes X, or, X causes Y.**

6. Spearman's Rank Correlation

Sometimes, it is not possible to measure variables in numerical terms. For example, intelligence of people and their physical appearance, or, aptitudes of people to art and music, etc. cannot be numerically measured

in the same way as their heights and weights. Such variables are called **attributes**. We may rank individuals, objects, etc. according to their quality of attributes. The ranks assigned to them are then used for purposes of analysis.

Suppose, n individuals have been ranked according to their intelligence and physical appearance, and their ranks are as follows.

	Individuals		
	1st	2nd	nth
Intelligence (R)	R_1	R_2	R_n
Physical Appearance (R')	R'_1	R'_2	R'_n

Spearman's coefficient of rank correlation provides a measure of linear association between ranks assigned to individuals or objects, according to the quality of their attributes.

The steps followed in the calculation of the rank correlation are:

Step 1. Obtain differences in ranks $D = R - R'$ for each individual.

Step 2. Obtain the sum of squares of differences in ranks, ΣD^2 , for all individuals.

Step 3. Spearman's rank correlation is obtained by the formula

$$r_k = 1 - \frac{6\Sigma D^2}{n^3 - n}, \text{ where } n \text{ is the number of individuals.}$$

We will consider the following cases to calculate the rank correlation.

Case 1. When actual ranks are given.

Case 2. When ranks are not given, but can be derived from actual values.

Case 3. When ranks are sometimes repetitive, i.e. when there are tie-ups.

We will illustrate these cases with examples.

Case 1. When actual ranks are given.

Example 9

Five students have been ranked according to their ability in mathematics and economics. Their ranks are given in the following table.

	Students				
	A	B	C	D	E
Rank in Maths	1	2	3	4	5
Rank in Eco	4	2	1	3	5

Calculate Spearman's rank correlation.

The calculations are shown in the Table 8.12 given below.

TABLE 8.12
Calculation of the Rank Correlation

Student	Rank in maths (R)	Rank in economics (R')	R-R'=D	D ²
A	1	4	-3	9
B	2	2	0	0
C	3	1	2	4
D	4	3	1	1
E	5	5	0	0
				$\Sigma D^2=14$

The rank correlation is

$$r_k = 1 - \frac{6\Sigma D^2}{n^3 - n}$$

and, therefore, substituting

$$n = 5 \text{ and } \Sigma D^2 = 14,$$

we get

$$r_k = 1 - \frac{6 \times 14}{5^3 - 5} = 1 - \frac{84}{120} = 0.3.$$

The ranks are positively but weakly associated.

Example 10

Seven students were ranked on the basis of their performance in 'oral' and 'written' examinations as follows.

Rank in the written examination	Rank in the oral examination
1	2
2	1
3	5
4	3
5	6
6	7
7	4

Given that the two examinations were conducted by independent examination bodies, do you think that the two sets of rankings are fairly close to each other?

The steps of calculations of Spearman's rank correlation are shown in Table 8.13.

TABLE 8.13
Calculation of Rank Correlation

Rank in written exam (R)	Rank in oral exam (R')	R- R' = D	D ²
1	2	-1	1
2	1	1	1
3	5	-2	4
4	3	1	1
5	6	-1	1
6	7	-1	1
7	4	3	9
			$\Sigma D^2=18$

Therefore, substituting

$$n = 7 \text{ and } \Sigma D^2 = 18$$

in the formula

$$r_k = 1 - \frac{6\Sigma D^2}{n^3 - n}$$

we get

$$r_k = 1 - \frac{6 \times 18}{7^3 - 7} = 0.68$$

Since, $r_k = 0.68$ is moderately high correlation, our conclusion would be that the rankings obtained by students in 'written' and 'oral' examinations are reasonably close to each other.

Case 2. When ranks can be derived from actual values.

Let us illustrate this with the following example.

Example 11

We are given the percentage of marks, obtained by six students in mathematics and economics, as shown in Table 8.14.

TABLE 8.14
Percentage of Marks in Mathematics and Economics

Student	Mathematics X	Economics Y
A	85	60
B	60	48
C	55	49
D	65	50
E	75	55
F	90	62

Calculate the coefficient of rank correlation between X and Y.

The steps of calculations are shown in Table 8.15.

Rank 1 is given to the largest value; Rank 2 to the entry next to the largest value, and so on. Rank six is given to the smallest value. We have no tie-ups of ranks.

Therefore,

$$r_k = 1 - \frac{6\Sigma D^2}{n^3 - n} = 1 - \frac{6 \times 2}{210} = 0.94$$

The ranks are highly correlated.

Example 12

Calculate rank correlation between X and Y for the data given below.

TABLE 8.15
Calculation for Rank Correlation

Marks in mathematics X	Rank in mathematics R	Marks in economics Y	Rank in economics R'	D=R-R'	D ²
85	2	60	2	0	0
60	5	48	6	-1	1
55	6	49	5	1	1
65	4	50	4	0	0
75	3	55	3	0	0
90	1	62	1	0	0
Total					2

X: 64 63 39 40 97 31 07 84 46 82
Y: 26 44 04 48 65 43 40 51 11 58

The calculations are shown in Table 8.16.

As before we start with the largest value and give Rank 1 to it; Rank 2 is given to the next value, and so on. The smallest value has Rank 10. There are no tie-ups of ranks.

Therefore,

$$\begin{aligned} r_k &= 1 - \frac{6\Sigma D^2}{n^3 - n} \\ &= 1 - \frac{6 \times 58}{10^3 - 10} \\ &= 1 - \frac{348}{990} = 0.65 \end{aligned}$$

It should be noted that the rank correlation r_k is equal to the product moment correlation between the ranks.

For example, compute the product moment correlation between ranks in Table 8.16.

Calculations are shown in Table 8.17.

TABLE 8.17
Product Moment Correlation Between Ranks
(in Table 8.16)

R	R'	R^2	R'^2	RR'
4	8	16	64	32
5	5	25	25	25
8	10	64	100	80
7	4	49	16	28
1	1	1	1	1
9	6	81	36	54
10	7	100	49	70
2	3	4	9	6
6	9	36	81	54
3	2	9	4	6
$\Sigma R=55$		$\Sigma R'=55$	$\Sigma R^2=385$	$\Sigma R'^2=385$
$\Sigma RR'=356$				

$$\begin{aligned} r_{RR'} &= \frac{n\Sigma RR' - (\Sigma R)(\Sigma R')}{\sqrt{n\Sigma R^2 - (\Sigma R)^2} \sqrt{n\Sigma R'^2 - (\Sigma R')^2}} \\ &= \frac{10 \times 356 - (55)(55)}{\sqrt{10 \times 385 - 55^2} \sqrt{10 \times 385 - 55^2}} \\ &= 0.65 \end{aligned}$$

which is the same as before.

In Example 9 we calculated

$$r_k = 1 - \frac{6\Sigma D^2}{n^3 - n} = 0.3$$

calculating the product moment correlation between the ranks, we get

$$r_{RR'} = 0.3.$$

TABLE 8.16
Calculation of Rank Correlation

X	Rank (R) according to X	Y	Rank (R') according to Y	$D=R-R'$	D^2
64	4	26	8	-4	16
63	5	44	5	0	0
39	8	4	10	-2	4
40	7	48	4	3	9
97	1	65	1	0	0
31	9	43	6	3	9
07	10	40	7	3	9
84	2	51	3	-1	1
46	6	11	9	-3	9
82	3	58	2	1	1
Total					58

Case 3. When ranks are repeated.

In a certain examination, two or more students may secure equal marks and thus, have a tie-up in their ranks among all students. Similarly, comparing different districts with respect to their level of literacy we may come across districts with equal literacy rates and thus, have the same rank among all districts, and so on.

Let us illustrate the method of computing rank correlation in such cases with the following example.

Example 13

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

Calculate the rank correlation between X and Y.

Arrange the values of X and Y in descending order of their magnitude.

Give Rank 1 to the largest value, Rank 2 to the second largest value and, so on. The lowest Rank 8 is given to the smallest value.

We find that X = 35 occurs both at the 4-th and 5-th place.

Give an **average rank**

$$\left(\frac{4+5}{2}\right)\text{-th} = 4.5\text{th}$$

to both.

Then the ranks are shown in Table 8.18.

The calculation of r_k is then done on the same lines as before,

$$r_k = 1 - \frac{6\sum D^2}{n^3 - n}$$

TABLE 8.18
Calculation of Rank Correlation

X	Y	R	R'	R-R'=D	D ²
25	55	6	2	4	16
45	60	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					$\sum D^2=65.5$

$$r_k = 1 - \frac{6 \times 65.5}{8^3 - 8} = 1 - \frac{393}{512-8} = 1 - \frac{393}{504} = 0.22.$$

EXERCISES

1. What is a scatter diagram? How does it help in determining the form of relationship between two variables X and Y?
2. What kind of relationship between X and Y is indicated, if the points of the scatter diagram tend to cluster about (a) a straight line parallel to the X-axis, (b) a straight line parallel to the Y-axis, (c) a straight line sloping upward, (d) a straight line sloping downward?
3. If the points in a scatter diagram tend to cluster about a straight line which makes an angle of 30° with the X-axis, what would you say about the strength of association between X and Y?

4. (a) How is Karl Pearson's coefficient of correlation defined?
 (b) What are the limits of the correlation coefficient r ?
 (c) If $r = +1$ or $r = -1$, what kind of relationship exists between X and Y ?
5. (a) Define Spearman's rank correlation r_k .
 (b) What are the limits of r_k ?
 (c) If the values of X and Y have been ranked and we compute product moment correlation between ranks of X and Y , will this correlation be equal to the value of r_k ?

6. Covariance of X and Y is defined as

$$\text{cov}(X, Y) = \frac{1}{n} \sum (X - \bar{X})(Y - \bar{Y})$$

in the same way as the variance of X and variance of Y .

What is the advantage in using the correlation coefficient defined as

$$r = \frac{1}{n} \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sigma_X \sigma_Y}$$

as a measure of association instead of the $\text{cov}(X, Y)$?

7. An epidemic spread in a certain state of the country and number of deaths were reported from different villages. The government took immediate steps and sent teams of doctors to the affected villages. Later, a statistician collected data on
 X = number of the deaths in different villages, and
 Y = number of doctors sent to different villages.

It was found that $r_{X,Y} = 0.8$

Is this high value of r sensible? Explain.

8. For the data on X and Y given as

X	-3	-2	-1	0	1	2	3
$Y(=X^2)$	9	4	1	0	1	4	9

Calculate the correlation between X and Y , and interpret your results.

9. The percentage of marks, obtained by ten students in a 'written' and 'oral' examination are as follows:

Written	81	62	74	78	93	69	72	83	90	84
Oral	76	71	69	76	87	62	80	75	92	79

- (a) Draw a scatter diagram.
 (b) What kind of association between two sets of marks do you infer from the scatter diagram?
 (c) Would you like to calculate the correlation coefficient to determine the strength of association between X and Y ?
10. The values of X and Y are given as follows:

X	0	1	2	3	4	5
Y	10	10	10	10	10	10

- (a) Draw a scatter diagram and determine the form of association.
 (b) Calculate the value of r .

11. You are given data on X and Y as follows:

X	64	63	39	40	97	31	07	84	46	82
Y	26	44	4	48	65	43	40	51	11	58

- Draw a scatter diagram and determine the form of association between X and Y.
- Calculate the correlation coefficient between X and Y.
- Rank the values of X and Y; and obtain rank correlation between X and Y.
- Calculate the product moment correlation between the ranks of X and Y obtained in (c).

12. The following table gives absolute values (nearest to thousand crores of rupees) of exports and imports of India at current prices.

Year	Exports	Imports
1996-97	119	139
1997-98	130	154
1998-99	142	176
1999-2000	119	149

Calculate the correlation between exports and imports.

13. The revenue receipts (X) and total expenditure (Y) of the Central Government are given in the table below.

(Rupees in thousand crores)

Year	Revenue receipts*	Total expenditure**
1990-91	55	98
1995-96	110	168
1996-97	126	190
1997-98	134	216
1998-99	151	255
1999-2000	183	284

* Revenue receipts are equal to the sum of tax and non-tax revenue.

** Total expenditure includes the Plan and Non-plan expenditure.

Calculate the correlation between X and Y.

14. Given that $n=8$, $\Sigma X=360$, $\Sigma X^2 = 20400$, $\Sigma XY = 13440$, $\Sigma Y=272$, $\Sigma Y^2 = 9920$ show that the product moment correlation between X and Y is 0.7.

15. (a) If the covariance between X and Y

$$\text{cov}(X,Y) = \frac{1}{n} \Sigma (X-\bar{X})(Y-\bar{Y})$$

is negative, will you get a negative or a positive correlation between X and Y?

- What are the units of measurement of cov (X,Y)?
- Can we get a value greater than +1 or less than -1 for the covariance between X and Y?

- 16.** The following table gives the private consumption expenditure and gross national product (GNP) for a ten year period.

<i>Private consumption expenditure (rupees in crores)</i>	<i>Gross National Product (rupees in crores)</i>
11761	14950
12427	15883
12929	16992
14567	19542
17355	22896
17826	23947
20349	27380
24561	32138
24945	33291
27158	36942

- Construct a scatter diagram from the data given above, plotting the gross national product on the X-axis and private consumption expenditure on the Y-axis. What relationship between the two variables is suggested by the scatter diagrams?
- Calculate the coefficient of product moment correlation r between the two variables. Does this value of r confirm the nature of relationship suggested above?

- 17.** Data on the price index (P) and money supply (M) for certain years are as follows:

<i>Price Index (P)</i>	<i>Money Supply (M) (rupees in crores)</i>
94.5	1862
78.3	2049
100.0	2725
135.5	4237
179.7	6729

Calculate the product moment correlation between P and M. What does this indicate?

CHAPTER 9

Introduction to Index Numbers

1. Meaning

An index number measures changes in the values of such variables as

- (a) prices of commodities
- (b) industrial production
- (c) agricultural production
- (d) cost of living, etc.

As a very simple example, suppose the price of onion in October, 2001 was Rs 2.50 per kilogram and it rose to Rs 3.00 per kilogram in December, 2001. The change in price of onion over the period of two months may be measured in two ways.

(i) *Actual Difference*

The actual difference in the prices of onion over the two months is equal to Rs 3.00 – Rs 2.50 = 0.50. That is, the price of onion in December was 50 paise per kilogram more than in October.

(ii) *Relative Change*

The relative change in prices is expressed as

$$\frac{\text{Rs } 3.00 - \text{Rs } 2.50}{\text{Rs } 2.50} = 0.20$$

i.e. **the actual difference in prices, relative to the original price.**

We may also express this as

$$\frac{\text{Rs } 3.00}{\text{Rs } 2.50} - 1 = 0.20;$$

or, in percentage terms

$$0.20 \times 100 = 20.$$

That is, the price of onion in December was 20 per cent more than in October.

The ratio of prices in the two months

$$\frac{\text{Rs } 3.00}{\text{Rs } 2.50} = 1.20$$

is called **the price relative**.

The price relative is a **pure number**.

The price relative minus '1' measures the relative change.

The price relative for a single commodity may be called an index number for that commodity.

Let us take another example. Suppose, we are considering two commodities, onion and cloth. Their prices in October, 2001 and December, 2001 were as follows.

	October 2001	December 2001
Onion	Rs 2.50 per kg	Rs 3.00 per kg
Cloth	Rs 15.00 per meter	Rs 15.50 per meter

In both cases, the price in December is 50 paise more than in October.

However, it should be noted that an increase of 50 paise per kilogram in the price of onion, **over Rs 2.50 per kilogram**, may be of much greater concern than an increase of 50 paise per meter in the price of cloth, **over Rs 15.00 per meter**.

If we measure the relative change in prices, we find that the onion price in December was

$$\frac{3.00 - 2.50}{2.50} \times 100 = 20$$

per cent more; where as the cloth price in December was

$$\frac{15.50 - 15.00}{15.00} \times 100 = 3.33$$

per cent higher.

The actual numerical difference in prices is not important. What is important is the change in price **in relation to its original level**.

In the example given above we had two commodities. Can we obtain a **composite index** of the changes in prices of onion and cloth together?

The absolute differences in prices cannot be combined, as they have different units of measurement. But the price relatives

$$\frac{\text{Rs } 3.00}{\text{Rs } 2.50} \text{ and } \frac{\text{Rs } 15.50}{\text{Rs } 15.00}$$

are pure numbers and we may combine them meaningfully.

Let us obtain the **arithmetic mean of price relatives**, to obtain.

$$\frac{1}{2} \left(\frac{\text{Rs } 3.00}{\text{Rs } 2.50} + \frac{\text{Rs } 15.50}{\text{Rs } 15.00} \right) = 1.12$$

as the **composite index** of prices of onion and cloth together.

The difference $1.12 - 1.00 = 0.12$ indicates the relative change in prices of the two commodities together. In percentage terms the **composite price** of the two commodities increased by $0.12 \times 100 = 12$ per cent.

If there are more than two commodities, we may again obtain the arithmetic mean of the price relatives to obtain a composite price index.

Conventionally, the index numbers are expressed in percentage terms.

2. Purpose

An index number of prices measures the relative change in the prices of a group of commodities. Similarly, we may construct an index number to measure the relative change in the 'quantities' of different commodities consumed, or, produced to obtain a 'quantity index'.

In the construction of an index number, some questions are of basic importance. For example,

- (i) which commodities should be included, and
- (ii) what kind of prices (retail, or, wholesale) should be used, etc.?

The answer to these questions would depend on the purpose for

which the index number is to be constructed. Suppose, we were to measure the change in the general price level, we would require to use wholesale prices of industrial, agricultural and other products. If we were to measure the change in cost of living, we would require retail prices of consumer goods and services. Similarly, for measuring change in industrial and agricultural production, we would require different kinds of information.

3. Choice of the Base Period

In our examples given above, we measured the prices in December, 2001 relative to those in October, 2001. Thus, in this example, October is called the **base period** and December is called the **current period**.

The choice of the base period (which is, in fact, the reference period) depends on the following considerations.

(i) **The base period should be a normal period.**

We should avoid using abnormal periods like period of war, drought, or, floods, etc. as base periods.

(ii) **The base period should be neither too short nor too long.**

As a working rule, it should be neither less than a month, nor more than a year.

(iii) **It should be the period for which actual data are available.**

(iv) **The base period should not be too far back in the past.**

The base period is the reference period with which the current period

is compared. If it is too far back in the past the comparison may become meaningless. The tastes change over time and many commodities may go out of use. In such a situation it is useful to shift the base period to a more recent period.

4. Selection of Commodities

The exact number and type of commodities to be selected for the construction of an index number depends on the purpose for which that index is to be used.

It is not possible to include all the commodities in any particular case. Therefore, first of all, we should identify the group for which the index is to be constructed.

It is necessary to ensure that the selected commodities are **representative of the group**.

For example, in the construction of a general price index, we should include wholesale prices of some major industrial and agricultural commodities and other goods and services. Similarly, for the construction of the consumer price index, we should include the retail prices of those commodities which are important in the family budget.

5. Construction of Index Numbers

Notation

The suffix '0' is used to refer to the base period and '1' refers to the current period. Thus, p_{i1} is the price of the i -th commodity in the current period and p_{i0} is the price in the base period.

P_{01} expresses the index number of prices for the current period '1' compared with those of the base period '0'.

Similarly, Q_{01} may be used to express a quantity index for the quantities in the current period '1' compared with the quantities in the base period '0'.

We may define a value index V_{01} in the same way.

Simple Aggregative Index

A simple aggregative price index is defined as

$$P_{01} = \frac{\sum p_{i1}}{\sum p_{i0}} \times 100;$$

which is the sum of prices of commodities in the current period '1' expressed as percentage of the sum of prices in the base period '0'.

However, since the units of measurement of prices of various commodities are, generally, not the same (these may be Rs per kilogram, Rs per meter, Rs per litre, etc.) a straight forward summation of prices is ruled out. Therefore, a simple aggregative index of prices or quantities has limited applicability.

Unweighted Index Number

As discussed earlier, the index number can be constructed for one or more commodities. In case we have only one commodity, the index number is simply the ratio of the price of the commodity in the current period to that in the base period. This is

usually expressed in percentage terms.

In case of several commodities, the index number is defined as the arithmetic mean of price relatives, expressed in percentage terms.

Suppose, the prices of n commodities in the current period '1' are $p_{11}, p_{21}, \dots, p_{n1}$, and those in the base period '0' are $p_{10}, p_{20}, \dots, p_{n0}$.

Then, the price relatives for the n commodities are $\frac{p_{11}}{p_{10}}, \frac{p_{21}}{p_{20}}, \dots, \frac{p_{n1}}{p_{n0}}$.

The **unweighted index number** is defined as the arithmetic mean

$$\frac{1}{n} \left(\frac{p_{11}}{p_{10}} + \frac{p_{21}}{p_{20}} + \dots + \frac{p_{n1}}{p_{n0}} \right)$$

or, briefly

$$\frac{1}{n} \sum \frac{p_{i1}}{p_{i0}}.$$

A limitation of the unweighted index number is that it gives equal weight (importance) to all the commodities, even though some may be more important than others.

Weighted Index Number

In order to allow adequate importance to different commodities in a composite index, we assign suitable weights to them. The weighted index number is simply the weighted arithmetic mean of price relatives defined as

$$P_{01} = \sum w_i \frac{p_{i1}}{p_{i0}}$$

where, the weights w_1, w_2, \dots, w_n are such that

$$\sum w_i = 1$$

A straight forward method of determining weights for various commodities is to have them equal to the proportion (or percentage) of expenditure on them in the total expenditure.

Example

Calculate the 'unweighted' and 'weighted' index of prices for the data given below.

Commodity	Price		Weight	Price relative P_{11}/P_{10}
	Base period	Current period		
Wheat	Rs 10 per kg	Rs 15 per kg	30 per cent	1.50
Rice	Rs 15 per kg	Rs 25 per kg	40 per cent	1.67
Salt	Rs 2 per kg	Rs 2.50 per kg	2 per cent	1.25
Ghee	Rs 40 per kg	Rs 60 per kg	5 per cent	1.50
Milk	Rs 12 per litre	Rs 15 per litre	20 per cent	1.25
Cloth	Rs 40 per meter	Rs 60 per meter	3 per cent	1.5
			100	8.67

The **unweighted index of prices** is

$$\frac{1}{6} \sum \frac{P_{11}}{P_{10}} = 1.45$$

and the **weighted index of prices** is

$$\frac{\sum w_i \left(\frac{P_{11}}{P_{10}} \right)}{\sum w_i} = \frac{(30 \times 1.50 + 40 \times 1.67 + 2 \times 1.25 + 5 \times 1.5 + 20 \times 1.25 + 3 \times 1.5)}{100} = 1.51$$

Expressed in percentage terms the weighted index is 151 and the unweighted index is 145.

Selection of Weights

The weights reflect the relative importance of different commodities in the aggregate. They may be defined as 'value shares' of different commodities in the total expenditure.

Suppose, $p_{10}, p_{20}, \dots, p_{n0}$ are the base year prices and $q_{10}, q_{20}, \dots, q_{n0}$ are the quantities of corresponding

commodities consumed in the base period. Then,

$$p_{10} q_{10}, p_{20} q_{20}, \dots, p_{n0} q_{n0}$$

are 'values' of different commodities consumed in the base period. The total expenditure in the base period is

$$\sum p_{i0} q_{i0}$$

and, therefore, the 'value shares' of different commodities in the base period are

$$w_1 = \frac{p_{10} q_{10}}{\sum p_{i0} q_{i0}}, w_2 = \frac{p_{20} q_{20}}{\sum p_{i0} q_{i0}}, \dots, w_n = \frac{p_{n0} q_{n0}}{\sum p_{i0} q_{i0}}.$$

Using these value shares as weights, a weighted average of price relatives would be

$$\begin{aligned}
 P_{01} &= \sum w_i \left(\frac{P_{i1}}{P_{i0}} \right) \\
 &= \sum \frac{P_{i0} q_{i0}}{\sum P_{i0} q_{i0}} \left(\frac{P_{i1}}{P_{i0}} \right) \\
 &= \frac{\sum P_{i1} q_{i0}}{\sum P_{i0} q_{i0}}
 \end{aligned}$$

This is called Laspeyres' index number.

Alternatively, we could choose the 'value shares' of different commodities consumed in the current period at base period prices. In that case the weights would be

$$w_1' = \frac{P_{10} q_{11}}{\sum P_{i0} q_{i1}}, w_2' = \frac{P_{20} q_{21}}{\sum P_{i0} q_{i1}}, \dots, w_n' = \frac{P_{n0} q_{n1}}{\sum P_{i0} q_{i1}}$$

Therefore, the index number would be

$$\begin{aligned}
 P_{01} &= \sum w_i' \left(\frac{P_{i1}}{P_{i0}} \right) \\
 &= \sum \frac{P_{i0} q_{i1}}{\sum P_{i0} q_{i1}} \left(\frac{P_{i1}}{P_{i0}} \right) \\
 &= \frac{\sum P_{i1} q_{i1}}{\sum P_{i0} q_{i1}}
 \end{aligned}$$

This is called Paasche's index number.

6. Some Important Index Numbers in Use

The following index numbers have been regularly published by the government and are used for policy purposes.

(a) Consumer Price Index (CPI)

The consumer price index is also called the **cost of living index**. It measures the average change in 'retail prices' at which the consumers buy goods and services in the market at a given point of time.

Since, the number of commodities is large, it is necessary to identify a particular group of consumers and select those commodities which are generally consumed by them.

The major groups of consumers for whom the consumer price index numbers have been constructed in India are

1. the industrial workers,
2. the urban non-manual workers, and
3. the agricultural labourers.

The consumer price index for industrial workers is by far the most popular index. This has been constructed annually for quite a long period of time; the most recent series is being constructed with 1993-94 as the base year.

A sample survey of households, in the particular group, is conducted, to obtain information on shares of expenditures on various items in the total household expenditure, in a fixed period.

A **representative** family budget is then prepared for the whole group. This is used to select the goods and services and their relative weights (their relative shares in the total household expenditure) for the construction of the consumer price index for the whole group.

Uses of Consumer Price Index

An increase in retail prices raises the consumer price index. The purchasing power of money is reduced and real wages decline.

- (i) The consumer price index is used to determine the purchasing power of money and real wages.

Suppose, the consumer price index was 250 in 1998-99; whereas it was 100 in the base period 1993-94. Then, a rupee in 1998-99 would be equal to

$$\text{Rs } \frac{100}{250} = 0.40,$$

i.e. 40 paise in 1993-94. This shows a decline of 60 paise in the purchasing power of rupee in 1998-99.

If the money wage of the consumer was Rs 800 p.m. in 1998-99, his real wage (according to 1993-94 prices) would be

$$\text{Rs } 800 \times \frac{100}{250} = \text{Rs } 320 \text{ p.m.}$$

The consumer price index is called the **price deflator of income**.

(ii) When the consumer price index increases beyond a certain level, the government decides to compensate workers by paying them additional money (i.e. the dearness allowance). The quantum of relief is determined by the increase in the consumer price index.

(iii) If the prices of certain essential commodities (like wheat, rice, sugar, cloth, etc.) increase, due to shortages, the government may decide to provide them through fair price shops or rationing.

(b) The Wholesale Price Index (WPI)

The wholesale price index is an indicator of change in the 'general price level'. This has been regularly published on **weekly basis** by the Office of the Economic Adviser, Ministry of Industry, Government of India.

The new series of wholesale price index numbers, published by the government, has the base period as 1993-94; whereas for the old series it was 1981-82. It covers a large number of commodities, which are broadly classified in the following three major groups;

- (i) primary goods such as food and non-food items,
- (ii) fuel, power, light and lubricants, and
- (iii) manufactured items.

Appropriate weights are attached to each group.

Uses of the Wholesale Price Index Numbers

- (i) The time series of wholesale price index numbers can be used to forecast future prices.
- (ii) Since the prices affect both the demand and supply, one can use an appropriate model to estimate the future demand and supply situations.

- (iii) The wholesale price index is used to measure the **rate of inflation**.

Suppose, the wholesale price index for the t-th and (t-1)-th weeks were X_t and X_{t-1} , respectively then the **weekly inflation rate** would be calculated as

$$\frac{X_t - X_{t-1}}{X_{t-1}} \times 100.$$

The annual inflation rate is calculated by using the **annual averages** of wholesale price index numbers. For example, the annual averages of the wholesale price index numbers for 1997-98 and 1998-99 were 134.4 and 142.4, respectively. Therefore, the annual inflation rate over this period was

$$\frac{142.4 - 134.4}{134.4} \times 100 = 5.95 \text{ per cent}$$

One can also calculate inflation rates for different commodities or, commodity groups as required for policy purposes.

(iv) Wholesale price index can be used to eliminate the effect of changes in prices on aggregates such as national income, capital formation, etc. The national income is defined as the value of goods and services produced in a certain year. If we calculate the value of goods and services according to prices prevailing in the same year, we get the **national income at current prices**.

However, an increase in the national income at current prices may be due to

- (i) an increase in the general price level, or
- (ii) an increase in the real output.

In order to determine the real output, we must eliminate the effect

of changes in prices.

Let us illustrate this with the following example.

Example

Suppose the total output in 1997-98 at current prices was Rs 1000 crores and that in the previous year 1996-97 it was Rs 885 crores. The WPI in 1997-98 was 134.4 and in 1996-97 it was 128. Calculate the output in 1997-98 at 1996-97 prices as

$$\frac{128}{134.4} \times 1000 = 952$$

Thus, the increase in real output was only $952 - 885 = 67$ and not $1000 - 885 = 115$.

The real output calculated at the base year prices would be

$$\frac{100}{134.4} \times 1000 = 744$$

(c) *Index of Industrial Production*

Index of industrial production is a quantity index. It measures changes in industrial production.

This index has been constructed in India annually with base period 1993-94 for the current series.

The three major groups for which the index has been constructed are

- (i) mining,
- (ii) manufacturing, and
- (iii) electricity.

Appropriate weights have been assigned to each.

EXERCISES

1. What is the purpose of constructing an index number of (a) prices, and (b) quantities?
2. Discuss the general method of constructing an index number and uses of an index number.
3. What are the considerations underlying the choice of base period in the construction of an index number?
4. Discuss the Simple Aggregative Price Index. What are its limitations?
5. Discuss the 'weighted' and unweighted' index of prices.
6. What are the considerations underlying the selection of
(a) weights, and
(b) commodities
in the construction of a weighted index of prices?
7. The following table provides prices of some major food items in 1991 and 2001 together with the data on the average expenditure per household per month in 1991.

Items	Units	Prices in rupees		Average expenditure/ month(Rs) in 1991
		1991	2001	
Rice	Rs per kg	15.00	35.00	804
Wheat	Rs per kg	9.00	15.00	310
Pulses	Rs per kg	20.00	25.00	245
Milk	Rs per litre	7.00	15.00	115
Oil	Rs per litre	30.00	65.00	110
Fish	Rs per kg	30.00	75.00	260
Tea	Rs per cup	0.75	3.00	130

Calculate

- (a) the Simple Aggregative Price Index
 - (b) the Unweighted Index of Price Relatives
 - (c) suitable Weighted Index Number of Price Relatives and compare the results.
8. The index numbers of agricultural production for foodgrains and non-foodgrains are given below:

Years	Base 1981-82 = 100	
	Index Nos. of agricultural production	
	Foodgrain Weight = 62.92	Non-foodgrain Weight = 37.08
1993-94	150.2	169.4
1994-95	155.9	180.9
1995-96	146.1	185.4
1996-97	160.9	200.9
1997-98	155.7	180.6
1998-99	164.8	198.1

Obtain the index numbers of total agricultural production.

9. Index numbers of area under principal crops with base period 1981-82 (= 100) are given below.

<i>Base: 1981-82 = 100</i>		
<i>Years</i>	<i>Foodgrains</i>	<i>Non-foodgrains</i>
1990-91	100.7	120.0
1993-94	96.7	127.3
1994-95	97.6	126.2
1995-96	95.3	131.7
1996-97	97.4	134.6
1997-98	97.6	133.5
1998-99	98.8	135.4

Foodgrains include (a) cereals (rice, wheat, coarse cereals), and (b) pulses.

Non-foodgrains include (a) oilseeds (ground nuts, rapeseed and mustard), (b) fibres (cotton, jute, etc.), (c) plantation crops (tea, coffee and rubber), and (d) others (sugarcane, tobacco, potato).

Interpret the data. What conclusions would you draw from these data? Write a short note on your findings.

CHAPTER 10

Project on Application of Statistical Methods in Economics

Any project that we undertake for studying a socio-economic problem requires collection and analysis of data on certain variables. The data may be 'secondary' (obtained from published sources), or, they may be 'primary' (obtained by conducting a field survey by the investigator himself). We have discussed the methods of collection of primary data (census and sampling methods) and the related problems in Chapter 2. At present let us note the following points.

1. The Purpose and Aim of the Study Must be Clearly Specified

For example, with a view to make an assessment of the prevalence of malpractices among sellers, we may plan a study on 'consumer awareness among households', or, to improve efficiency in the production processes, we may plan a study on 'productivity awareness among enterprises'.

2. Population to be Studied

If we are planning a field survey, to assess consumer awareness among

households, we should specify the target group of households, which we want to consider. There are very rich and very poor households. There are also middle class households. They differ significantly with regard to awareness about quality of goods they buy. Rural and urban households, and literate and illiterate ones react differently. Similarly, if we want to study productivity awareness among enterprises, we have large and small firms, which would differ significantly with regard to productivity awareness, and so on.

3. Data to be Used

If we are using secondary data from published sources, we must pay attention to the definitions of the variables on which data are to be obtained. We should try to find out how the data were collected. If they are index numbers, for example, what was the method of construction (weights, coverage, etc.); if the data are on output of firms, is it in value terms at current, or, constant prices, etc.

We must use published data with great care. Otherwise, we may arrive at wrong conclusions.

If we are planning a field survey to study consumer awareness among households we note that the population is not homogeneous (as noted above). Therefore, we may either restrict to a homogeneous group and use the method of simple random sampling or, adopt stratified random sampling method. The stratified random sampling method has been discussed in Chapter 2.

4. Questionnaire

We should prepare the questionnaire with great care, as suggested in Chapter 2. The questionnaire may be prepared in parts, as illustrated in Appendix B.

As discussed in Chapter 2, we may follow either the **interview method**,

or, send questionnaires by mail to the respondents. Both the methods have their advantages and disadvantages (see Chapter 2).

5. Analysis of Data

We may obtain the proportions of households

- (i) who examined goods when they made any purchases
- (ii) who reported to the supplier about defective/adulterated items
- (iii) proportion of households whose complaints were attended by the suppliers to their satisfaction
- (iv) proportion of households who took action by reporting to consumer courts.

Do these proportions differ significantly from stratum to stratum?

Do you conclude that educated households have higher degree of consumer awareness?

APPENDIX A

TABLE OF TWO-DIGIT RANDOM NUMBERS

03 47 43 73 86	36 96 47 36 61	46 98 63 71 62	33 26 16 80 45	60 11 14 10 95
97 74 24 67 62	42 81 14 57 20	42 53 32 37 32	27 07 36 07 51	24 51 79 89 73
16 76 62 27 66	56 50 26 71 07	32 90 79 78 53	13 55 38 58 59	88 97 54 14 10
12 56 85 99 26	96 96 68 27 31	05 03 72 93 15	57 12 10 14 21	88 26 49 81 76
55 59 56 35 64	38 54 82 46 22	31 62 43 09 90	06 18 44 32 53	23 83 01 30 30
16 22 77 94 39	49 54 43 54 82	17 37 93 23 78	87 35 20 96 43	84 26 34 91 64
84 42 17 53 31	57 24 55 06 88	77 04 74 47 67	21 76 33 50 25	83 92 12 06 76
63 01 63 78 59	16 95 55 67 19	98 10 50 71 75	12 86 73 58 07	44 39 52 38 79
33 21 12 34 29	78 64 56 07 82	52 42 07 44 38	15 51 00 13 42	99 66 02 79 54
57 60 86 32 44	09 47 27 96 54	49 17 46 09 62	90 52 84 77 27	08 02 73 43 28
18 18 07 92 46	44 17 16 58 09	79 83 86 19 62	06 76 50 03 10	55 23 64 05 05
26 62 38 97 75	84 16 07 44 99	83 11 46 32 24	20 14 85 88 45	10 93 72 88 71
23 42 40 64 74	82 97 77 77 81	07 45 32 14 08	32 98 94 07 72	93 85 79 10 75
52 36 28 19 95	50 92 26 11 97	00 56 76 31 38	80 22 02 53 53	86 60 42 04 53
37 85 94 35 12	83 39 50 08 30	42 34 07 96 88	54 42 06 87 98	35 85 29 48 39
70 29 17 12 13	40 33 20 38 26	13 89 51 03 74	17 76 37 13 04	07 74 21 19 30
56 62 18 37 35	96 83 50 87 75	97 12 25 93 47	70 33 24 03 54	97 77 46 44 80
99 49 57 22 77	88 42 95 45 72	16 64 36 16 00	04 43 18 66 79	94 77 24 21 90
16 08 15 04 72	33 27 14 34 09	45 59 34 68 49	12 72 07 34 45	99 27 72 95 14
31 16 93 32 43	50 27 89 87 19	20 15 37 00 49	52 85 66 60 44	38 68 88 11 80
68 34 30 13 70	55 74 30 77 40	44 22 78 84 26	04 33 46 09 52	68 07 97 06 57
74 57 25 65 76	59 29 97 68 60	71 91 38 67 54	13 58 18 24 76	15 54 55 95 52
27 42 37 86 53	48 55 90 65 72	96 57 69 36 10	96 46 92 42 45	97 60 49 04 91
00 39 68 29 61	66 37 32 20 30	77 84 57 03 29	10 45 65 04 26	11 04 96 67 24
29 94 98 94 24	68 49 69 10 82	53 75 91 93 30	34 25 20 57 27	40 48 73 51 92
16 90 82 66 59	83 62 64 11 12	67 19 00 71 74	60 47 21 29 68	02 02 37 03 31
11 27 94 75 06	06 09 19 74 66	02 94 37 34 02	76 70 90 30 86	38 45 94 30 38
35 24 10 16 20	33 32 51 26 38	79 78 45 04 91	16 92 53 56 16	02 75 50 95 98
38 23 16 86 38	42 38 97 01 50	87 75 66 81 41	40 01 74 91 62	48 51 84 08 32
31 96 25 91 47	96 44 33 49 13	34 86 82 53 91	00 52 43 48 85	27 55 26 89 62
66 67 40 67 14	64 05 71 95 86	11 05 65 09 68	76 83 20 37 90	57 16 00 11 66
14 90 84 45 11	75 73 88 05 90	52 27 41 14 86	22 98 12 22 08	07 52 74 95 80
68 05 51 18 00	33 96 02 75 19	07 60 62 93 55	59 33 82 43 90	49 37 38 44 59
20 46 78 73 90	97 51 40 14 02	04 02 33 31 08	39 54 16 49 36	47 95 93 13 30
64 19 58 97 79	15 06 15 93 20	01 90 10 75 06	40 78 78 89 62	02 67 74 17 33
05 26 93 70 60	22 35 85 15 13	92 03 51 59 77	59 56 78 06 83	52 91 05 70 74
07 97 10 88 23	09 98 42 99 64	61 71 62 99 15	06 51 29 16 93	58 05 77 09 51
68 71 86 85 85	54 87 66 47 54	73 32 08 11 12	44 95 92 63 16	29 56 24 29 48
26 99 61 65 53	58 37 78 80 70	42 10 50 67 42	32 17 55 85 74	94 44 67 16 94
14 65 52 68 75	87 59 36 22 41	26 78 63 06 55	13 08 27 01 50	15 29 39 39 43

APPENDIX A (Cont.)

17 53 77 58 71	71 41 61 50 72	12 41 94 96 26	44 95 27 36 99	02 96 74 30 83
90 26 59 21 19	23 52 23 33 12	96 93 02 18 39	07 02 18 36 07	25 99 32 70 23
41 23 52 55 99	31 04 49 69 96	10 47 48 45 88	13 41 43 89 20	97 17 14 49 17
60 20 50 81 69	31 99 73 68 68	35 81 33 03 76	24 30 12 48 60	18 99 10 72 34
91 25 38 05 90	94 58 28 41 36	45 37 59 03 09	90 35 57 29 12	82 62 54 65 60
34 50 57 74 37	98 80 33 00 91	09 77 93 19 82	74 94 80 04 04	45 07 31 66 49
85 22 04 39 43	73 81 53 94 79	33 62 46 86 28	08 31 54 46 31	53 94 13 38 47
09 79 13 77 48	73 82 97 22 21	05 03 27 24 83	72 89 44 05 60	35 80 39 94 88
88 75 80 18 14	22 95 75 42 49	39 32 82 22 49	02 48 07 70 37	16 04 61 67 87
90 96 23 70 00	39 00 03 06 90	55 85 78 38 36	94 37 30 69 32	90 89 00 76 33
53 74 23 99 67	61 32 28 69 84	94 62 67 86 24	98 33 41 19 95	47 53 53 38 09
63 38 06 86 54	99 00 65 26 94	02 82 90 23 07	79 62 67 80 60	75 91 12 81 19
35 30 58 21 46	06 72 17 10 94	25 21 31 75 96	49 28 24 00 49	55 65 79 78 07
63 43 36 82 69	65 51 18 37 88	61 38 44 12 45	32 92 85 88 65	54 34 81 85 35
98 25 37 55 26	01 91 82 81 46	74 71 12 94 97	24 02 71 37 07	03 92 18 66 75
02 63 21 17 69	71 50 80 89 56	38 15 70 11 48	43 40 45 86 98	00 83 26 91 03
64 55 22 21 82	48 22 28 06 00	61 54 13 43 91	82 78 12 23 29	06 66 24 12 27
85 07 26 13 89	01 10 07 82 04	59 63 69 36 03	69 11 15 83 80	13 29 54 19 28
58 54 16 24 15	51 54 44 82 00	62 61 65 04 69	38 18 65 18 97	85 72 13 49 21
34 85 27 84 87	61 48 64 56 26	90 18 48 13 26	37 70 15 42 57	65 65 80 39 07
03 92 18 27 46	57 99 16 96 56	30 33 72 85 22	84 64 38 56 98	99 01 30 98 64
62 95 30 27 59	37 75 41 66 48	86 97 80 61 45	23 53 04 01 63	45 76 08 64 27
08 45 93 15 22	60 21 75 46 91	98 77 27 85 42	28 88 61 08 84	69 62 03 42 73
07 08 55 18 40	45 44 75 13 90	24 94 96 61 02	57 55 66 83 15	73 42 37 11 61
01 85 89 95 66	51 10 19 34 88	15 84 97 19 75	12 76 39 43 78	64 63 91 08 25
72 84 71 14 35	19 11 58 49 26	50 11 17 17 76	86 31 57 20 18	95 60 78 46 75
88 78 28 16 84	13 52 53 94 53	75 45 69 30 96	73 89 65 70 31	99 17 43 48 76
45 17 75 65 57	28 40 19 72 12	25 12 74 75 67	60 40 60 81 19	24 62 01 61 16
96 76 28 12 54	22 01 11 94 25	71 96 16 16 88	68 64 36 74 45	19 59 50 88 92
43 31 67 72 30	24 02 94 08 63	38 32 36 66 02	69 36 38 25 39	48 03 45 15 22
50 44 66 44 21	66 06 58 05 62	68 15 54 35 02	42 35 48 96 32	14 52 41 52 48
22 66 22 15 86	26 63 75 41 99	58 42 36 72 24	58 37 52 18 51	03 37 18 39 11
96 24 40 14 51	23 22 30 88 57	95 67 47 29 83	94 69 40 06 07	18 16 36 78 86
31 73 91 61 19	60 20 72 93 48	98 57 07 23 69	65 95 39 69 58	56 80 30 19 44
78 60 73 99 84	43 89 94 36 45	56 69 47 07 41	90 22 91 07 12	78 35 34 08 72
84 37 90 61 56	70 10 23 98 05	85 11 34 76 60	76 48 45 34 60	01 64 18 39 96
36 67 10 08 23	98 93 35 08 86	99 29 76 29 81	33 34 91 58 93	63 14 52 32 52
07 28 59 07 48	89 64 58 89 75	83 85 62 27 89	30 14 78 56 27	86 63 59 80 02
10 15 83 87 60	79 24 31 66 56	21 48 24 06 93	91 98 94 05 49	01 47 59 38 00
55 19 68 97 65	03 73 52 16 56	00 53 55 90 27	33 42 29 38 87	22 13 88 83 34
53 81 29 13 39	35 01 20 71 34	62 33 74 82 14	53 73 19 09 03	56 54 29 56 93
51 86 32 68 92	33 98 74 66 99	40 14 71 94 58	45 94 19 38 81	14 44 99 81 07
35 91 70 29 13	80 03 54 07 27	96 94 78 32 66	50 95 52 74 33	13 80 55 62 54
37 71 67 95 13	20 02 44 95 94	64 85 04 05 72	01 32 90 76 14	53 89 74 60 41
93 66 13 83 27	92 79 64 64 72	28 54 96 53 84	48 14 52 98 94	56 07 93 89 30

APPENDIX A (Cont.)

02 96 08 45 65	13 05 00 41 84	93 07 54 72 59	21 45 57 09 77	19 48 56 27 44
49 83 43 48 35	82 88 33 69 96	72 36 04 19 76	47 45 15 18 60	82 11 08 95 97
84 60 71 62 46	40 80 81 30 37	34 39 23 05 38	25 15 35 71 30	88 12 57 21 77
18 17 30 88 71	44 91 14 88 47	89 23 30 63 15	56 34 20 47 89	99 82 93 24 98
79 69 10 61 78	71 32 76 95 62	87 00 22 58 40	92 54 01 75 25	43 11 71 99 31
75 93 36 57 83	56 20 14 82 11	74 21 97 90 65	96 42 68 63 86	74 54 13 26 94
38 30 92 29 03	06 28 81 39 38	62 25 06 84 63	61 29 08 93 67	04 32 92 08 09
51 29 50 10 34	31 57 75 95 80	51 97 02 74 77	76 15 48 49 44	18 55 63 77 09
21 31 38 86 24	37 79 81 53 74	73 24 16 10 33	52 83 90 94 76	70 47 14 54 36
29 01 23 87 88	58 02 39 37 67	42 10 14 20 92	16 55 23 42 45	54 96 09 11 06
95 33 95 22 00	18 74 72 00 18	38 79 58 69 32	81 76 80 26 92	82 80 84 25 39
90 84 60 79 80	24 36 59 87 38	82 07 53 89 35	96 35 23 79 18	05 98 90 07 35
46 40 62 98 82	54 97 20 56 95	15 74 80 08 32	16 46 70 50 80	67 72 16 42 79
20 31 89 03 43	38 46 82 68 72	32 14 82 99 70	80 60 47 18 97	63 49 30 21 30
71 59 73 05 50	08 22 23 71 77	91 01 93 20 49	82 96 59 26 94	66 39 67 98 60
22 17 68 65 84	68 95 23 92 35	87 02 22 57 51	61 09 43 95 06	58 24 82 03 47
19 36 27 59 46	13 79 93 37 55	39 77 32 77 09	85 52 05 30 62	47 83 51 62 74
16 77 23 02 77	09 61 87 25 21	28 06 24 25 93	16 71 13 59 78	23 05 47 47 25
78 43 76 71 61	20 44 90 32 64	97 67 63 99 61	46 38 03 93 22	69 81 21 99 21
03 28 28 26 08	73 37 32 04 05	69 30 16 09 05	88 69 58 28 99	35 07 44 75 47
93 22 53 64 39	07 10 63 76 35	87 03 04 79 88	08 13 13 85 51	55 34 57 72 69
78 76 58 54 74	92 38 70 96 92	52 06 79 79 45	82 63 18 27 44	69 66 92 19 09
23 68 35 26 00	99 53 93 61 28	52 70 05 48 34	56 65 05 61 86	90 92 10 70 80
15 39 25 70 99	93 86 52 77 65	15 33 59 05 28	22 87 26 07 47	86 96 98 29 06
58 71 96 30 24	18 46 23 34 27	85 13 99 24 44	49 18 09 79 49	74 16 32 23 02
57 35 27 33 72	24 53 63 94 09	41 10 76 47 91	44 04 95 49 66	39 60 04 59 81
48 50 86 54 48	22 06 34 72 52	82 21 15 65 20	33 29 94 71 11	15 91 29 12 03
61 96 48 95 03	07 16 39 33 66	98 56 10 56 79	77 21 30 27 12	90 49 22 23 62
36 93 89 41 26	29 70 83 63 51	99 74 20 52 36	87 09 41 15 09	98 60 16 03 03
18 87 00 42 31	57 90 12 02 07	23 47 37 17 31	54 08 01 88 63	39 41 88 92 10
88 56 53 27 59	33 35 72 67 47	77 34 55 45 70	08 18 27 38 90	16 95 86 70 75
09 72 95 84 29	49 41 31 06 70	42 38 06 45 18	64 84 73 31 65	52 53 37 97 15
12 96 88 17 31	65 19 69 02 83	60 75 86 90 68	24 64 19 35 51	56 61 87 39 12
85 94 57 24 16	92 09 84 38 76	22 00 27 69 85	29 81 94 78 70	21 94 47 90 12
38 64 43 59 98	98 77 87 68 07	91 51 67 62 44	40 98 05 93 78	23 32 65 41 18
53 44 09 42 72	00 41 86 79 79	68 47 22 00 20	35 55 31 51 51	00 83 63 22 55
40 76 66 26 84	57 99 99 90 37	36 63 32 08 58	37 40 13 68 97	87 64 81 07 83
02 17 79 18 05	12 51 52 57 02	22 07 90 47 03	28 14 11 30 79	20 69 22 40 98
95 17 82 06 53	31 51 10 96 46	92 06 88 07 77	56 11 50 81 69	40 23 72 51 39
35 76 22 42 92	96 11 83 44 80	34 68 35 48 77	33 42 40 90 60	73 96 53 97 86
26 29 13 56 41	85 47 04 66 08	34 72 57 59 13	82 43 80 46 15	38 26 61 70 04
77 80 20 75 82	72 82 32 99 90	63 95 73 76 63	89 73 44 99 05	48 67 26 43 18
46 40 66 44 52	91 36 74 43 53	30 82 13 54 00	78 45 63 98 35	55 03 36 67 68
37 56 08 18 09	77 53 84 46 47	31 91 18 95 58	24 16 74 11 53	44 10 13 85 57
61 65 61 68 66	37 27 47 39 19	84 83 70 07 48	53 21 40 06 71	95 06 79 88 54

APPENDIX A (Cont.)

93 43 69 64 07	34 18 04 52 35	56 27 09 24 86	61 85 53 83 45	19 90 70 99 00
21 96 60 12 99	11 20 99 45 18	48 13 93 55 34	18 37 79 49 90	65 97 38 20 46
95 20 47 97 97	27 37 83 28 71	00 06 41 41 74	45 89 09 39 84	51 67 11 52 49
97 86 21 78 73	10 65 81 92 59	58 76 17 14 97	04 76 62 16 17	17 95 70 45 80
69 92 06 34 13	59 71 74 17 32	27 55 10 24 19	23 71 82 13 74	63 52 52 01 41
04 31 17 21 56	33 73 99 19 87	26 72 39 27 67	53 77 57 68 93	60 61 97 22 61
61 06 98 03 91	87 14 77 43 96	43 00 65 98 50	45 60 33 01 07	98 99 46 50 47
85 93 85 86 88	72 87 08 62 40	16 06 10 89 20	23 21 34 74 97	76 38 03 29 63
21 74 32 47 45	73 96 07 94 52	09 65 90 77 47	25 76 16 19 33	53 05 70 53 30
15 69 53 82 80	79 96 23 53 10	65 39 07 16 29	45 33 02 43 70	02 87 40 41 45
02 89 08 04 49	20 21 14 68 86	87 63 93 95 17	11 29 01 95 80	35 14 97 35 33
87 18 15 89 79	85 43 01 72 73	08 61 74 51 69	89 74 39 82 15	94 51 33 41 67
98 83 71 94 22	59 97 50 99 52	08 52 85 08 40	87 80 61 65 31	91 51 80 32 44
10 08 58 21 66	72 68 49 29 31	89 85 84 46 06	59 73 19 85 23	65 09 29 75 63
47 90 56 10 08	88 02 84 27 83	42 29 72 23 19	66 56 45 65 79	20 71 53 20 25
22 85 61 68 90	49 64 92 85 44	16 40 12 89 88	50 14 49 81 06	01 82 77 45 12
67 80 43 79 33	12 83 11 41 16	25 58 19 68 70	77 02 54 00 52	53 43 37 15 26
27 62 50 96 72	79 44 61 40 15	14 53 40 65 39	27 31 58 50 28	11 39 03 34 25
33 78 80 87 15	38 30 06 38 21	14 47 47 07 26	54 96 87 53 32	40 36 40 96 76
13 13 92 66 99	47 24 49 57 74	32 25 43 62 17	10 97 11 69 84	99 63 22 32 98

APPENDIX B

QUESTIONNAIRE**Part A : Personal Information**Schedule (Question Schedule) Number Area Code Household Number

Type of Locality :

(a) Very rich (b) Upper Middle Class (c) Middle Class (d) Very Poor (slum) (e) Rural (f) Urban

Respondent's Name :

Address

Respondent's

(a) Caste

(b) Religion

Family Type

(a) Nuclear (b) Joint **Family Profile**

<i>S.No.</i>	<i>Names of household members</i>	<i>Relation to the respondent</i>	<i>Age</i>	<i>Sex</i>	<i>Education</i>
1.	Respondent	Self			
2.					
3.					
4.					
5.					

Work Status

<i>S.No</i>	<i>Names of household members</i>	<i>Earning (E)/not earning (NE)</i>	<i>Nature of job</i>	<i>Income</i>
1.	Respondent			
2.				
3.				
4.				
5.				

Part B : Food Items

Q. Where do you regularly buy your grocery from?

(a) Supermarket

(b) Nearby grocer

(c) Different grocers

(d) Elsewhere (specify)

Q. Where do you regularly buy milk from?

(a) Mother Dairy booth

(b) DMS booth

(c) Polypacks from grocer

(d) Milkman

(e) Elsewhere (specify)

Q. Where do you regularly buy meat from?

(a) Nearby meat shop

(b) Elsewhere (specify)

Q. Where do you regularly buy fish from?

(a) Nearby fish shop

(b) Vendor

(c) Elsewhere (specify)

Q. Where do you regularly buy vegetables from?

(a) Nearby shop

(b) Vendor

(c) Elsewhere (specify)

Part C : Clothing and other Consumer Durables

Q. Do you, generally, buy readymade garments?

(a) Yes ☐ (b) No ☐ (c) Sometimes ☐

Q. Do you get your clothes made to order by tailor?

(a) Yes ☐ (b) No ☐

Q. Where do you buy your durable goods? (Furniture, Radio, TV, Refrigerator, etc.)

Specify ☐

Part D : Consumer Awareness

Q Do you examine the goods when you buy them?

(a) Yes ☐ (b) No ☐

Q. Did you ever come across adulteration in

(a) food stuff (a) Yes ☐ (b) No ☐

(b) milk (a) Yes ☐ (b) No ☐

Q. If yes, did you complain to

(a) the shopkeeper ☐

(b) main supplier ☐

(c) elsewhere (specify) ☐

Q. Was your complaint to the supplier/shopkeeper attended by him to your satisfaction?

(a) Yes ☐ (b) No ☐

Q. Do you check the prices, of goods you buy, from alternative sources?

(a) Yes ☐ (b) No ☐

Q. Are you aware of consumer courts, for redressal of grievances of consumers?

(a) Yes ☐ (b) No ☐

Q. Have you ever filed a case in the consumer court?

(a) Yes ☐ (b) No ☐

SOME IMPORTANT SOURCES OF SECONDARY DATA

1. *Statistical Abstract India 2000*, published by the Central Statistical Organisation, Ministry of Statistics and Programme Implementation, Government of India.
2. *Annual Plan*, Government of India, Planning Commission, New Delhi.
3. *Reserve Bank of India Bulletin*, Reserve Bank of India.
4. *National Accounts Statistics 2000*, Central Statistical Organisation, Ministry of Statistics and Programme Implementation, Government of India.
5. *Census of India 2001*, Provisional Population Totals, Rajasthan, Director of Census Operations, Rajasthan.
6. *Monthly Review of the Indian Economy*, January 2002, published by Centre for Monitoring Indian Economy Pvt. Ltd. Economic Intelligence Service.
7. *Economic Survey 2001-2002*, Government of India, Ministry of Finance, Economic Division.

GLOSSARY OF STATISTICAL TERMS

Population in statistics, implies the total set of values of a certain variable (or variables) on all individual units in the whole region. We have a **univariate population**, if the values are on one single variable; otherwise, if the values are on more than one variable, we have a **bivariate or multivariate population**.

Parameter is a certain unknown constant or constants, of the population. For example, the arithmetic mean or the variance of the population are parameters.

Estimator is the method of obtaining an estimate of the parameter values from sample data. For example, the sample arithmetic mean is an estimator of the population arithmetic mean; and the sample variance is an estimator of the population variance, etc.

Estimate is the numerical value of the estimator that we obtain from a given sample.

Sampling Error is the numerical difference between the estimate and the true value of the parameter.

Census Method is a method of data collection which requires that observations are taken on all individual units in a certain region.

Sample Method requires that observations are obtained on a representative set of individuals selected from the population. Random sampling guarantees equal probability of selection to all individual units in the population. If the population is not homogeneous, but homogeneous strata can be defined, then Stratified Random Sampling is used for selection.

Non-Sampling Errors arise in data collection due to (i) errors in measuring variables, (ii) recording mistakes, (iii) errors of non-response, etc.

Questionnaire is a list of questions, prepared by the investigator, on the subject of enquiry. The respondent is required to answer the questions.

Average is a measure of central tendency of the distribution of the values of a certain variable. For example, the arithmetic mean, median and mode are averages which measure the central value in different ways. The arithmetic mean is the central value in the sense that the sum of numerical deviations (of the values of the variable) from the arithmetic mean is zero; the median is the central value in the sense that the number of values greater than the median is equal to the number of values less than the median (actual magnitudes of the values are ignored); and mode is the value of the variable with highest frequency.

Dispersion is a measure of scatter or dispersion of values about the central value. For example, the standard deviation and mean deviation are measures of dispersion about the arithmetic mean. The range and the quartile deviation are other measures

of dispersion, but they do not consider deviations from any value. They measure dispersion in a general way.

Partition Values The median, quartiles, deciles and percentiles are called partition values. The median is that value of the variable, which divides the entire set of values in two equal halves; the quartiles are the values (Q_1, Q_2, Q_3) which divide them in four equal parts; deciles divide them in ten equal parts and percentiles in hundred equal parts.

Correlation Coefficient is a measure of association between two variables. However, its interpretation is difficult, if the relationship between variables is not linear, or, there are more than two variables. If the relationship between two variables is linear, the square of the correlation measures the strength of association between them. For example, suppose the correlation coefficient is equal to 0.6; equal to 0.6. Now, $(0.6)^2 = 0.36$, and this means that 36 per cent of the variation in one variable is explained by its linear relationship with the other. In other words, 64 per cent variation is unexplained by the relationship. Obviously, $r = 0.6$ cannot be called high correlation. Similarly, suppose the correlation between x and y is 0.3 and between other two variables u and v is 0.6. Should we say that the strength of association between u and v is twice as much as between x and y ? The answer is no; because $r_{x,y}^2 = 0.09$ and $r_{u,v}^2 = 0.36$. Thus, in fact, the strength of association between u and v is four times as much as between x and y . The correlation coefficient does not indicate the direction of causation. It does not tell, if x causes y , or, y causes x .

Index Number is a barometer of economic activity. It measures relative change in prices, quantities and values over time. For example, the prices in current period are compared with base year prices, etc. Index numbers are useful for various policy purposes.

ANSWERS

Chapter 2

Q.12 120

Chapter 3

- Q.9** (i) range = 45%
 (iia) using exclusive method
- | | | | | | |
|-------|-------|-------|-------|-------|--------|
| 45-55 | 55-65 | 65-75 | 75-85 | 85-95 | 95-100 |
| 5 | 20 | 7 | 9 | 8 | 1 |
- (iib) using exclusive method
- | | | | | |
|-------|-------|-------|-------|-------|
| 50-55 | 55-60 | 60-65 | 65-70 | 70-75 |
| 5 | 7 | 13 | 5 | 2 |
-
- | | | | | |
|-------|-------|-------|-------|--------|
| 75-80 | 80-85 | 85-90 | 90-95 | 95-100 |
| 8 | 1 | 5 | 3 | 1 |

Q.10 (a) range = Rs 9539

Chapter 6

- Q.6** A.M. = 29.1925, Median = 28.50, Mode = 28
Q.7 A.M. \approx 6.96, Median = 6.75, Mode \approx 6.33
Q.8 A.M. \approx 79.47, Median = 84, Mode = 84
Q.9 Median \approx 46.74
Q.10 Lower quartile $Q_1 = 32.09$ and upper quartile $Q_3 = 47.70$
Q.11 Median = 41.17
Q.12 $Q_1 \approx 202.38$, Median ≈ 261.90 , $Q_3 = 390$, Mode ≈ 252.17
 [For **moderately asymmetric distributions** we may use the **approximate** relation
 A.M. - Mode ≈ 3 (A.M. - Median)
 to derive the arithmetic mean]
 Arithmetic Mean ≈ 266.765

Chapter 7

- Q.6** (i) A.M. \approx 173.16, s.d. \approx 17.27
 (ii) A.M. \approx 173.72, s.d. \approx 17.52
 (iii) A.M. \approx 173.67, s.d. \approx 17.96

Q.7	<i>Mathematics</i>	<i>Economics</i>
Range	55%	24%
A.M.	55.07%	48.47%
s.d.	15.59%	7.45%
M.D. (about arith. mean)	12.59%	6.30%

Relative measures of dispersion:

	<i>Mathematics</i>	<i>Economics</i>
$R/x_{\max} + x_{\min}$	0.52	0.24
s.d./A.M.	0.28	0.15
M.D./A.M.	0.23	0.13

- Q.8** (a) Variance of costs = 144 (Rs)², s.d. of costs = Rs 12
Mean Deviation of costs about the arithmetic mean = Rs 9.84
- (b) (i) $\frac{s.d.}{A.M.} = 0.17$ and
(ii) $\frac{M.D.}{A.M.} = 0.14$ as the
Arithmetic Mean = 71.
- Q.9** (a) (i) $\bar{x} = 12$
(ii) s.d. = 3.66
(iii) M.D. about $\bar{x} = 3.20$
- (b) (i) $\Sigma(x - 10)^2 = 174$
(ii) Median = 11; $\Sigma|x - \text{median}| = 30$
- (c) $\Sigma(x_i - 10)^2 = 174$
 $\Sigma(x_i - \bar{x})^2 = 134$
 $\Sigma|x_i - \bar{x}| = 32$
 $\Sigma|x_i - \text{median}| = 30$
- Q.10** Variance¹ = 0.01 (litre)²
- Q.11** Coefficient of variation in 2000 is 6 and in 2001 it is 9. Thus, the relative dispersion in 2001 is more than in 2000. Hence, the results have not improved in 2001.
- Q.12** (i) Arithmetic Mean = 5.5
(ii) Standard Deviation ≈ 2.87
(iii) Mean Deviation about the Arithmetic Mean = 2.5
(iv) Mean Deviation about the Median = Mean Deviation about the Arithmetic Mean, because A.M. = Median = 5.5.
Coefficient of Variation ≈ 52 per cent using the standard deviation; and the Coefficient of Variation using the mean deviation is 45.45 per cent.

Chapter 8

- Q.10** Observe that the variance of Y is zero; therefore, we cannot apply the formula for product moment correlation. However, we see that the points in the scatter diagram lie exactly on a straight line parallel to the X-axis. The slope of the straight line is zero. Therefore, $r = 0$.
- Q.11** (b) $r \approx 0.49$
(c) $r_k \approx 0.63$
(d) $r \approx 0.63$
- Q.12** $r \approx 0.95$
- Q.13** $r \approx 0.99$
- Q.16(a)** Express both the variables nearest to thousand crores to obtain:

Private consumption expenditure (Rs 000 crores)	Gross National Product (Rs 000 crores)
12	15
12	16
13	17
15	20
17	23
18	24
20	27
25	32
25	33
27	37

This would facilitate drawing the scatter diagram

(b) $r = 0.996$

Q.17 $r \approx 0.98$

Chapter 9

Q.7 (b) 229.42 per cent

(c) 220.55 per cent

Q.8 157.3, 165.2, 160.7, 175.7, 164.9 and 177.2