UNIT 1 BASIC CONCEPTS

Structure

1.0 OBJECTIVES

After going through this Unit you will be able to:

- appreciate the importance of statistics in our life;
- explain some basic concepts used in the study of statistics;
- define statistics in both singular as well as the plural sense; and
- identify the uses and misuses of statistics.

1.1 INTRODUCTION

Now-a-days the word 'statistics' has become a household word, although different people comprehend it in different senses. The modern educated person has to be a person of statistics, broadly understanding its meaning and applying it to his/her life in different ways. For example, everyday we come across different types of quantitative information in both print as well as electronic media on topics like population, exchange rate fluctuations, inflation rate, day and night temperatures (being below or above normal; lowest or highest in the century or in the last thirty years or so), etc. In order to improve our understanding of the world around us, it is necessary to:

a) measure what is being said,

b) express it numerically, i.e., in numbers/quantities like weights in so many kilograms, eggs in so many dozens, etc., and

c) utilize quantitative information or expression to draw conclusions and suggest policy measures.

Needless to say that if we cannot measure and express, in terms of numbers what is being said, then our knowledge will remain insufficient and far from being satisfactory. Statistics thus involves some sort of numerical information called "numerical data" or simply "data". For example, one may give a statement that he/she has studied statistics (that is quantitative information) on absenteeism among the educated and the uneducated workers in Indian industries and found that incidence of absenteeism is more among the latter. He/she is referring to the numerical figures or numerical information technically called data.

Other examples of data are:

a) India is suffering from population explosion, annual growth of population being around 2%.

b) Students of XIIA have shown a better result than those of XIIB because the average marks of the former are 25% more than the average marks of the latter.

c) Foreign exchange reserve of the country has been the highest so far since independence and stood at $ 110 billion.

d) As per the 2001 census population of India was 1027 million.

Many more such examples can be found and the students are expected to go through this exercise on their own.

History of Statistics

The word Statistics is the modern form of the word *statistik* which in turn has been derived from the Italian word "statista" meaning "statesman". Professor Gott Fried Achenwall used it in the 18th century. It was Dr. E.A.W. Zimmerman who introduced the word statistics into England.

Early government records show statistical information on some aspects of population, land records, military strength of different wings, mortality during epidemics and so on. Perhaps it was because of this that Statistics was called the science of kings. But as the humanity developed, the usage as well as understanding of Statistics increased and now it is difficult to imagine a field of knowledge which can do without statistics. In fact, it has become an important tool of analysis.

1.2 MEANING OF STATISTICS

Let us look into the meaning of the word 'Statistics'. It conveys different meaning to different people. A common man may simply interpret it as a mass of figures, graphs or diagrams relating to an economic, business or some other scientific activity. However, for an expert, it may also imply a *statistical method of investigation* in addition to a mere mass of figures. Let us discuss each of these.

1.2.1 Statistics in Plural Sense

Statistics in plural sense means the mass of quantitative information called 'data'. For example, we talk of information on population or demographic features of India available from the Population Census conducted every ten years by the Government of India. Similarly, we can have statistics (quantitative data or simply data) on

enrollment of students in a particular university, say, over the last ten years. Further, data are collected by almost all ministries of the Government of India relating to their activities.

Also referred to as Statistical Data, Horace Secrist describes statistics in plural sense as follows:

"*By Statistics we mean aggregates of facts affected to a marked extent by multiplicity of causes numerically expressed, enumerated or estimated according to reasonable standard of accuracy, collected in a systematic manner for a pre-determined purpose and placed in relation to each other.*"

This definition of statistics in plural sense highlights the following features:

a) *Statistics are numerical facts*: In order that information obtained from an investigation can be called as statistics or data, it must be capable of being represented by numbers. The collected data may be obtained either by the measurement of characteristics (like data on heights, weights, etc.) or by counting when the characteristics (like honesty, smoking habit, beauty, etc.) is not measurable.

b) *Statistics are aggregates of facts:* Single and unrelated figures even though expressed as quantities are not statistics. For example, in a university examination Mr. Sharma secures 65% marks does not make statistics or data. However, if we find that out of 3 lakh university students whose average marks were 55%, Mr. Sharma secured 65% marks, then these figures are statistics. So no single figure in any sphere of statistical inquiry, say production, employment, wage and income constitutes statistics.

c) *Statistics are affected to a marked extent by multiplicity of causes*: In physical sciences it is possible to isolate the effect of various forces on a particular event. But in 'Statistics' facts and figures, that is, the collected information, are greatly influenced by a number of factors and forces working together. For example, the output of wheat in a year is affected by various factors like the availability of irrigation, quality of soils, method of cultivation, type of seed, amount of fertilizer used, etc. In addition to this there may be certain factors which are even difficult to identify.

d) *Statistics are numerically expressed:* Statistics are statements of facts expressed numerically or in numbers. Qualitative statements like "the students of a school ABC are more intelligent than those of school XYZ" cannot be regarded statistics. Contrary to this the statement that "the average marks in school ABC are 90% compared with 60% in school XYZ, and that the former had 80% first division compared with only 50% in the latter", is a statistical statement.

e) *Statistics are enumerated or estimated with a reasonable degree of accuracy*: While enumerating or estimating statistics, a reasonable degree of accuracy must be achieved. The degree of accuracy needed, in an investigation, depends upon the nature and objective of investigation on one hand and upon the time and resources on the other. Thus it is necessary to have a reasonable degree of accuracy of data, keeping in mind the nature and objective of investigation and availability of time and resources. The degree of accuracy once decided must be uniformly maintained throughout the investigation.

f) *Statistics are collected in a systematic manner*: Before the collection of statistics, it is necessary to define the objective of investigation. The objective

of investigation must be specific and well defined. The data are then collected in systematic manner by proper planning which involves finding of answers to questions such as: Whether to use sample or census investigation, how to collect, arrange, present and analyse data, etc. This will be discussed in Unit 2 in greater detail.

g) *Statistics should be placed in relation to one another:* Only comparable data make some sense. Unrelated and incomparable data are no data. They are just figures. For example, heights and weights of students of a class do not have any relation with the income and qualification of their parents. For comparability, the data should be homogeneous; that is, it should belong to the same subject or class or phenomenon. For example, pocket money of the students of a class is certainly related to the income of their parents. Prices of onions and potatoes in Delhi can certainly be related to their prices in other cities of India.

Thus, it will not be wrong to say that "*all statistics are numerical statements of facts but all numerical statements of facts are not statistics*".

1.2.2 Statistics in Singular Sense

In the singular sense, Statistics refers to what is called *statistical methods* which means the ever-growing body of techniques for collection, condensation, presentation, analysis and interpretation of statistical data/quantitative information. In simple language, it means the subject of Statistics like any other subject such as Mathematics or Economics.

We can now take up definitions given by some famous statisticians.

A. L. Bowley gave a few definitions but none of them was complete and satisfactory. However, his two definitions make some sense even though incomplete. For example, he says, "*Statistics may be called the science of counting*". Here he is emphasizing on enumeration aspect of statistics, which no doubt is important. At another place he describes statistics as "*the science of measurement of the social organism...*". He is also of the view that "*Statistics may rightly be called the science of average*". Although measurement, enumeration and averages (Arithmetic, Geometric and Harmonic means; Mode and Median which we will discuss in the next Block) are important, yet they are not the only concern of Statistics, as we shall study in the subsequent units.

Croxton and Cowden have put forward a very simple and precise definition of Statistics as "*Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data*".

This definition lays emphasis on five important aspects, which in fact, constitute the very scope of the subject called **Statistics** or **Statistical Methods**. These are:

A) *Collection of data:* In any statistical inquiry, the collection of data is the first basic step. They form the foundation of statistical analysis, and therefore utmost care should be taken in collecting data. Faulty data will certainly lead to misleading results and can do more harm than good. The data can be drawn from two sources:

   a) *Primary source* where data are generated by the investigator himself through various methods discussed in detail in the next Unit, Section 2.4.

b) *Secondary source* where data are extracted from the existing published or unpublished source, that is, from the data already collected by others. It saves a lot of time, effort and money of the investigator; but then he has to be conscious and judicious in their use. A detailed discussion is available in our Unit 2, Section 2.5.

B) *Arrangement of Data:* Data from the secondary source are already arranged or organised like population data from Census of India. A minor rearrangement to suit our needs can be undertaken. However, primary data are in a haphazard form and need some arrangement so that it makes some sense. The steps involved in this process are: –

a) *Editing:* This involves the removal of omissions and inconsistencies involved in the collected information.

b) *Classification of data*: It follows editing. It involves arranging data according to some common characteristic/s. Normally the raw information received from the respondents is put on the master sheets. For example, we may conduct a survey on, say, metal based engineering industries of Orissa, from where information are collected on capital structure, output of different types of products, employment of unskilled, semi-skilled and skilled workers, cost and price structure, technology aspects, etc. All this information can be put on master sheets. For more details refer to Unit 3.

C) *Tabulation:* It is the last step in the arrangement process. From the master sheets (or coded sheets) information is tabulated in the form of frequency distributions or tables, where information is arranged in columns and rows. For more details refer to Unit 3.

D) *Presentation of Data*: After the data have been arranged and tabulated, they can now be presented in the form of diagrams and graphs to facilitate the understanding of various trends as well as the process of comparison of various situations. Two different types of presentation of data are normally used, detailed study of which will be made in Unit 3.

These are:

a) Statistical tables

b) Graphs including line graphs.

E) *Analysis of Data*: It is the most important step in any statistical inquiry. A major portion of this course in Statistics is devoted to the methods used for analysing the collected data to derive some policy conclusions. The tools of analysis will be discussed in details in later units. For the time being we can summarize them as follows:

TOOLS OF STATISTICAL ANALYSIS

I) Theoretical Statistics

a) *Uni-variate analysis*.

i) Measures of Central tendency: This includes mathematical averages such as arithmetic mean $(\bar{X})$geometric mean (G) and harmonic mean (H) and positional averages such as mode (M_o) and median (M_d), and other partition values which include quartiles (Q), octiles (O), deciles (D) and percentiles (P).

ii) Measures of dispersion: These include crude measures such as range (R), quartile deviation (QD), Mean Deviation (d), standard deviation (s), etc.

iii) Measures of Skewness (S_k) — Karl Pearsons, Robert Bowley's, and Moment based (b_1 coefficient) measures.

iv) Measures of Kurtosis (b_2 coefficient) based on moments.

v) Probability and probability distributions such as Binomial, Poisson and Normal.

b) *Bi-variate analysis*

It includes analysis using two variables like amount of fertilizers (x) and the amount of yield (y) where it is known that yield (y) is affected by the amount of fertilizers (x) used. In this context we will discuss linear correlation (r_{xy}) and regression analysis in Units 7 and 8.

II) **Applied Statistics**

Here we use the tools developed in I to analyse some very useful aspects of our daily life. These include:

a) Time series

b) Index numbers

c) Vital Statisctics

d) Inferential statistics, e.g., testing of hypothesis, etc.

F) *Interpretation of Data*: It is the last but very crucial stage of a statistical inquiry or investigation. It is a job in itself which requires high degree of aptitude, skill and experience. In case of faulty interpretation, the very purpose of the investigation is lost. Our policies and actions later on depend very much on how soundly and correctly we interpreted our data. On this basis Wallis and Robert (Statistics - A New Approach) have rightly remarked that statistics may be regarded as "*a body of methods for making wise decisions in the face of uncertainty*".

1.2.3 Meaning of the Word 'Statistic'

You must have been buying a few kilograms of wheat every month for your family consumption. How do you judge the quality of wheat contained in a bag of 100 kgs? Theoretically two methods are open to you:

a) *Census method* where each and every grain of wheat is examined. You will study in Unit 2 how this method is costly, time consuming, boring and at the same time unnecessary because almost same results can be obtained from a sample inquiry.

b) *Sample method* where one or more samples, each containing few grains, are selected and examined. If you are satisfied with the sample/s, you buy the grains, assuming that all grains in the bag are of similar quality.

The statistical values of the characteristics of a population (such as mean height of students in a university for) are known as *parameters*. On the other hand, the mean, standard deviation, etc. of the sample taken from the population, are known as *statistic* and are the estimators of the parameter values.

Check Your Progress 1

1) Are the following statements correct? Give reasons in two or three lines:-

   a) Statistics has no use for a modern man.

   b) Statistics and statistic imply the same thing.

c) Statistics in singular sense implies statistical methods.

d) Statistics may rightly be called the science of averages.

e) Statistics need not be numerically expressed.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

2) Give five examples of the use of Statistics in daily life.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

3) Use the words Statistics, statistics and statistic in three separate sentences to bring out the difference in them.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

4) How would you judge the quality of potatoes from a bag of 100 kg. from which you want to buy 5 kg. Use the words population, sample, statistic and parameters in your explanation. (Restrict your answer to six lines).

.....

.....

1.3 IMPORTANCE OF STATISTICS

We have seen in Section1.1, that Statistics has a very wide application in our daily life. It is required in every field of inquiry; its knowledge has become necessary to study and understand our day to day problems. A Statistician H. G. Wells had rightly pointed out that "*statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write*". Further, Statistics has acquired universal application. Even A. L. Bowley remarked that, "*Statistics cannot be confined to any one science*". Thus, statistics is/are applied and interpreted in different ways in different fields of knowledge.

1.3.1 Statistics and Economics

The relationship between Statistics and Economics is very old. In 1690 Sir William Petty wrote a book titled "Political Arithmetic" using Statistics in Economics. In the late 19th century Alfred Marshall had observed that "*Statistics are the straw out of which I, like every other economist, have to make bricks*".

In the 20th century, economists largely based their theories on statistical inquiry — on empirical evidence of human behavior rather than on deductive methods of analysis. J.M. Keynes, V. Pareto and others used statistics very extensively. Recently Statistics and Economics have been so intermixed that a new branch known as *Econometrics* has developed. Mean, Standard deviation, Regression analysis, Normal distribution, Sampling theory, etc. are being used extensively in economic analysis. In addition to this, following are the other important uses of statistics. The list is only illustrative and not exhaustive.

1) Estimation and analysis of national income.
2) Input-output analysis.
3) Empirical analysis of production function.
4) Financial statistics as contained in Reserve Bank of India Bulletins.
5) Statistical studies of population or demographic features like death rate, birth rate, life expectancy, etc.
6) Statistical studies of market structures like oligopolies and monopolies, etc.
7) Macro-economic variables like price level, employment, money supply, etc.
8) It would be impossible to understand and steer the growth process in underdeveloped economies without the availability of sufficient and reliable statistical information. Economic planning is just not possible without the availability of sufficient and reliable data.

1.3.2 Statistics and Business

Statistics helps in business too. For a progressive business concern, analysis of costs, revenue, profits, labour and capital, marketing, etc. are essential. Business planning involves business forecasting based on market surveys on demand, availability of substitute brands, opinions of consumers regarding different brands, consumers preferences, etc. Using time series analysis, one may isolate the effects of secular trend, seasonal variations, cyclical factors and irregular factors, on a business activity (see Unit 10).

Statistical methods are useful to business in formulating its business policies and activities in the field of production, finance, personnel, accounting and quality control. Modern business firms make extensive use of graphs, charts, and diagrams in their sale promotion efforts and display of their production achievements.

1.3.3 Statistics and Physical Sciences

Statistics has proved to be useful in physical sciences like Physics, Geology, Astronomy, Biology, Medicine, etc. A modern doctor relies heavily on the information on various parameters of a patient in diagnosing his disease. These include his body temperature behaviour, blood pressure and blood sugar level, ECG, etc. Doctor needs this information all the more when performing surgery.

Further, before introducing a new drug, data are collected and analysed for its effects on rats, monkeys, rabbits, etc. If found statistically satisfactory, the experiments are then conducted on human beings. The efficacy of the medicine is studied statistically. For example, researchers may be interested in finding whether quinine is still effective in the control of malaria with a new strain of mosquito. They may conduct the experiment on, say, 1000 patients selected at random. If the percentage of success is quite high, researchers may declare that quinine is still effective in the control of malaria.

Similarly, statistical studies are conducted in other physical sciences. Perhaps, it will not be an exaggeration to say that there is hardly any scientific study where use of statistical methods is not undertaken. The Gaussian "*Normal Law of Errors*" was used to study the movements of stars and planets. Thus, as Bowley pointed out, statistics can "*prove useful at any time under any circumstance*".

1.3.4 Statistics and Mathematics

The relation between Statistics and Mathematics is known to exist since the 17th century. The theory of probability has bearing on various statistical methods. In the last 100 years or so Statistics and Mathematics have come very close to each other to evolve a new subject called *Mathematical Statistics*.

1.3.5 Statistics and other Social Sciences

Similarly, scholars are increasingly using Statistics in Education, Political Science, Geography, Psychology, Anthropology, etc. All public opinion polls are based on Statistics. Other fields where Statistics is useful are all types of insurance, war/defense preparedness, index numbers and dearness allowances formulae, etc.

1.4 MISUSES OF STATISTICS

Although Statistics is indispensable in almost all fields of learning as pointed out above, yet it is likely to be misused and misinterpreted by the vested interests. These interests, like a ruling party, can always manipulate figures to arrive at the predetermined favourable results. Because of the various misuses, Statistics is sometimes called an *unscrupulous science*. Various facts can be twisted, distorted and presented with an evil design. This becomes easy when the state or the other vested interests have the monopoly of collecting and presenting statistics.

All this, no doubt therefore, has produced various misgivings about Statistics such as:

a) "Statistics can prove anything"

b) "Statistics are the lies of the first order".

c) "There are three kinds of lies, namely, lies, damned lies and statistics"

d) "Statistics is the rainbow of lies."

Statistical conclusions may be misinterpreted and hence can be disastrous. A story goes that a mathematician finding average height of his family members higher than average depth of a stream, decided to cross it safely. But on the other bank of the stream he found that except himself, all other members were drowned because his abnormal height had pulled up the average.

1.5 LIMITATIONS OF STATISTICS

As mentioned earlier in Section 1.3 of this Unit, and according to H.G Wells, "Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write". However, statistics are not like Alladin's lamp which can perform all tricks. The following list of limitations is worth mentioning:

Firstly, statistical analysis depends upon the type of variable under consideration. For qualitative data such as beauty, health, goodwill and honesty, attempts have been made indirectly in the form of Analysis of Association of Attributes. Here we do not measure things like honesty but count their number.

Secondly, statistics deal only with aggregates. That is no significance is attached to individual items which make this aggregate. For example, one state of India may be richer than other states, but some people may be much poorer in the rich state than some people of the poorer states. Averages sometimes may be misleading.

Thirdly, statistical conclusions are not mathematically exact. It is possible that with wrong samples, taken knowingly or unknowingly, the results may be favourable by fluke.

Fourthly, as mentioned in Section 1.4, statistical measures like averages may be misinterpreted and hence can prove disastrous.

Check Your Progress 2

1) Comment on the following statements in 3-4 sentences

   a) Statistics are confined only to Economics and Business.

c) If the per capita income of a country is Rs. 4050, it means that everybody is getting that income.

d) Econometrics is an intermix of Economics and Mathematics.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

2) Give five examples of the use of Statistics in Economics and Business.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

3) Mention four fields where Statistics is being used prominently.

.....

.....

.....

.....

.....

.....

.....

.....

.....

4) Explain the following terms:

a) Mathematical Statistics

b) Statistics as an unscrupulous science

c) Statistics as Alladin's lamp

.....

.....

1.6 LET US SUM UP

A modern man must possess knowledge of Statistics like that of reading and writing. The word Statistics in singular sense implies statistical methods aimed at collecting, arranging, presenting, analysing and interpreting data. In the plural sense, it mean mass of quantitative information like population data.

The word statistic (as against statistics) means an estimator obtained from a sample with a purpose to infer about the population value called parameter.

Statistics has utility in almost all branches of knowledge. In Economics and Business it has a special utility. Combination of Economics, Statistics and Mathematics has led to a new subject called *Econometrics*.

In spite of immense utility, some unscrupulous persons have misused statistics driving it to the level that is worse than damned lies. Because of this, sometimes, it has been termed as unscrupulous science.

1.7 KEY WORDS

Statistics: In plural sense, it means a set of numerical figures commonly known as statistical data.

Statistics: In singular sense, it means scientific methods for collection, presentation, analysis and interpretation of data.

Statistic: It is measure, like arithmetic mean, median, geometric mean, standard deviation, etc., calculated from sample. It is also termed as estimator in the theory of estimation.

Parameter: It is a measure like arithmetic mean, median, geometric mean, standard deviation, etc., calculated by using all values of population.

Quantitative data: These are information on measurable characteristics. Such data are available in the form of numerical figures.

Qualitative data: These are information on a non-measurable characteristics like honesty, beauty, color, caste, etc.

Population: The totality of all the units falling under the scope of an investigation.

Sample: A sample is a fraction of the population used to study its one or more characteristics.

Census: A method of investigation in which information is collected from all units of the population.

Sampling: A method of investigation in which information is collected from sampled units only.

1.8 SOME USEFUL BOOKS

Elhance, D.N. and V. Elhance, 1988, *Fundamental of Statistics,* Kitab Mahal, Allahabad.

Nagar, A.L. and R.K. Dass, 1983, *Basic Statistics,* Oxford University Press, Delhi

Mansfield, E., 1991, *Statistics for Business and Economics: Method and Applications,* W.W. Norton and Co.

Yule, G.U. and M.G. Kendall, 1991, *An Introduction to the Theory of Statistics,* Univeristy Books, Delhi.

1.9 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) (a) false, (b) false, (c) true, (d) true, (e) false.
2) Refer Section 1.1
3) Refer Sub-Sections 1.2.1 to 1.2.3
4) Refer Sub-Section 1.2.3

Check Your Progress 2

1) a) Refer Section 1.3
   b) and (c) refer Section 1.4
   d) Refer Sub-Section 1.3.1
2) Refer Sub-Sections 1.3.1 and 1.3.2
3) Refer Section 1.3
4) a) Refer Sub-Section 1.3.4
   b) Refer Section 1.4
   c) Refer Section 1.5

UNIT 2 DATA COLLECTION METHODS

Structure

2.0 OBJECTIVES

On going through this Unit you will be able to:

- explain the concept and types of data;
- identify the importance of data in a statistical inquiry;
- explain various survey techniques; and
- identify the uses and limitations of secondary data.

2.1 INTRODUCTION

We face problems in various fields of our life, which force us to think and discover their solutions. When we are genuinely serious about the solution of a problem faced, a thinking process starts. Statistical Thinking or Statistical Inquiry is one kind of thinking process which requires evidence in the form of some information, preferably quantitative, which is known as data/statistical information.

In a statistical inquiry, the first step is to procure or collect data. Every time the investigator may not start from the very beginning. He must try to use what others have already discovered. This will save us in cost, efforts and time.

As discussed in Unit 1 (Section 1.2.1) data imply related quantitative information. They are collections of any number of related observations with a predetermined goal. We can collect information on the number of T.V. sets sold by a particular salesman or a group of salesmen, on weekdays in different parts of Delhi to study the pattern of sales, lean days, effect of competitive products, income behaviour and other related matters. The information thus collected is called a data set and a single observation a data point.

All types of information collected without proper aim or objective is of no use. For example, John's height is 5'6" or monthly wage of Mr. X on 1st January 2004 were Rs.15000/- are not data. Not all quantitative information are statistical. Isolated measurements are not statistical data. Statistics (that is in singular sense) is concerned with collection of data relevant to the solution of a particular problem. According to Simpson and Kafka (Basic Statistics),"*Data have no standing in themselves; they have a basis for existence only where there is a problem*".

2.2 PURPOSE OF DATA COLLECTION

By now you have known that data could be classified in the following three ways:

a) Quantitative and Qualitative Data.

b) Sample and Census Data.

c) Primary and Secondary data.

a) Quantitative and Qualitative data: Quantitative data are those set of information which are quantifiable and can be expressed in some standard units like rupees, kilograms, litres, etc. For example, pocket money of students of a class and income of their parents can be expressed in so many rupees; production or import of wheat can be expressed in so many kilograms or lakh quintals; consumption of petrol and diesel in India as so many lakh litres in one year and so on.

Qualitative data, on the other hand, are not quantifiable, that is, cannot be expressed in standard units of measurement like rupees, kilograms, litres, etc. This is because they are 'features', 'qualities' or 'characteristics' like eye-colours, skin complexion, honesty, good or bad, etc. These are also referred to as attributes. In this case, however, it is possible to count the number of individuals (or items) possessing a particular attribute.

b) Sample and Census Data: It was discussed in Section 1.2.3 of Unit 1 that data can be collected either by census method or sample method. Information collected through sample inquiry is called sample data and the one collected through census inquiry is called census data. Population census data are collected every ten years in India.

c) Primary and Secondary Data: As discussed very briefly in Section 1.2.2, primary data are collected by the investigator through field survey. Such data are in raw form and must be refined before use. On the other hand, secondary data are extracted from the existing published or unpublished sources, that is, from the data already collected by others.

Collection of data is the first basic step towards the statistical analysis of any problem. The collected data are suitably transformed and analysed to draw conclusions about the population. These conclusions may be either or both of the following:

i) To estimate one or more parameters of a population or the nature of the population itself. This forms the subject matter of the theory of estimation (discussed in Block 7).

ii) To test a hypothesis. A hypothesis is a statement regarding the parameters or the nature of population (discussed in Block 7).

2.3 COLLECTION OF DATA

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

2.3.1 Statistical Inquiry — Planning and Conduct

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

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

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

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

Any statistical survey consists of two stages:

i) Planning Stage

ii) Executing Stage

2.3.2 Planning Stage — Requisites of a Statistical Inquiry

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

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

Unless the investigator answers this question most satisfactorily, (s)he cannot proceed in the right direction and can go astray. Both money and efforts will be lost if data, not relevant to inquiry, are collected. Not only this, one must also be clear about how much data are required and hence ensure that only the necessary data get collected. For example, if we want to collect data on pattern of wheat production in a particular state, we need to collect data on the type of land, agricultural inputs, educational levels of farmers involved, presence or absence of defects of land tenure system, availability and cost of agricultural finance, nature of marketing, etc.

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

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

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

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

1) *Census* or *Sample* inquiry. In census method (s)he examines each and every item / individual of the population whereas in sample method (s)he examines only the item / individual included in the sample. For example, in census method (s)he examines each and every persons in a village, but in sample method, (s)he examines only a limited number of persons.

2) *Direct* or *Indirect* inquiry. In a direct inquiry the observations can be directly obtained in quantitative terms as for example, sales of T.V. sets and the advertisement cost in rupees. On the other hand, in an indirect inquiry, like intelligence of a group of students, marks secured by them are used to judge their intelligence.

3) *Original* or *Repetitive* inquiry. An inquiry conducted for the first time is original but if it is undertaken over and over again, it is repetitive. For example, population census in India is conducted every 10 years. All these inquiries must be related.

4) *Open* or *Confidential* inquiry. In open inquiry the results are made public, as for example, the population and national income data. On the other hand, the results of many government inquiries are kept confidential for reasons of national security, as for example, data on defence, atomic energy, space research and development, etc.

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

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

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

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

In various economic and business studies, absolute accuracy is neither necessary nor possible. In population data, accuracy till the last person is not required. For example, population of India is 98,89,70,510 or 98,89,00,000 does not matter much. However, the degree of accuracy required will determine the choice between different methods of collecting data. Further, the degree of accuracy, once decided, must be maintained throughout the survey.

2.3.3 Execution Stage

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

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

2) *Selection* and *Training* of field staff called interviewers or investigators or research staff or enumerators. They will approach the respondents in different ways as explained in Section 2.4. Investigators should be properly trained, should be honest and hard working. Any error at this stage will jeopardise the whole process of investigation giving misleading results. To obtain the best possible results from a survey, it is desirable to have the field staff who is familiar with the language of the respondents and have patience and tact of dealing with them.

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

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

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

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

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

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

2.3.4 Primary and Secondary Data

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

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

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

A) **Published Sources**

1) Official publications of the government at all levels — Central, State, Union Territories and Councils

2) Official publications of foreign countries.
3) Official publications of international bodies like IMF, UNESCO, WHO, etc.
4) Newspapers and Journals of repute, both local and international.
5) Official publications of RBI, and other Banks, LIC, Trade Unions, Stock Exchange, Chambers of Commerce, etc.
6) Reports submitted by reputed economists, research scholars, universities, commissions of inquiry, if made public.

Some main sources of published data in India are:

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

ii) **National Sample Survey Organisation** (N.S.S.O.): Under Ministry of Statistics and Programme Implementation, this organisation provides us data on all aspects of national economy, such as agriculture, industry, labour and consumption expenditure.

iii) **Reserve Bank of India Publications** (R.B.I.): It publishes financial statistics. Its publications are Report on Currency and Finance, Reserve Bank of India Bulletin, Statistical Tables Relating to Banks in India, etc.

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

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

B) **Un-published Sources**

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

Check Your Progress 1

1) Explain the following terms :

(Answers should not exceed three sentences each.)

a) Questionnaire b) Non-response
c) Hypothesis d) Statistical unit
e) Statistical inquiry f) Question Schedule

.....

.....

.....

.....

.....

.....

.....

.....

2) Distinguish between the following terms :

(Answers should not exceed four sentences each.)

a) Data, Statistical Data and Statistics

b) Data Set and Data Point

c) Primary and Secondary Data

d) Quantitative and Qualitative Data

e) Sample and Census Data

f) Sample and Census Inquiry

g) Planning and Execution of Statistical Inquiry

h) Survey and Experiment

i) Direct Investigation and Investigation through Secondary Source.

.....

.....

.....

.....

.....

.....

3) What are different sources of information?

.....

.....

.....

.....

.....

.....

2.4 COLLECTION OF PRIMARY DATA — SURVEY TECHNIQUES

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

a) Direct Personal Investigation

b) Indirect Oral Investigation

c) Use of Local Reports

d) Questionnaire Method

a) Direct Personal Investigation

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

1) She/he should be polite, unbiased and tactful.

2) She/he should know the local conditions, customs and traditions so that she/

3) She/he should be intelligent possessing good observation power.
4) She/he should use simple, easy and meaningful questions to extract information.

This method is suitable only for intensive investigations. It is a costly method in terms of money, effort and time. Further, the personal bias of the investigator cannot be ruled out and it can do a lot of harm to the investigation. The method is a complete flop if the investigator does not possess the above mentioned qualities.

b) Indirect Oral Investigation Method

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

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

c) Use of Local Reports

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

d) Questionnaire Method

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

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

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

i) The task of soliciting information from people in desired form and with sufficient accuracy is the most difficult problem. By their nature people are not willing to reveal any information because of certain fears. Many a times they provide incomplete and faulty information. Therefore, it is necessary that the respondents be taken into confidence. They should be assured that their individual information will be kept confidential and no part of it will be revealed to tax and other government investigative agencies. This is very essential indeed.

ii) Where providing information is not legally binding, the informant has to be induced through appeals or by using clever arguments. They must be explained

and convinced that the results of the survey will help the authorities to frame policies which will ultimately benefit them. It is obvious that some element of good salesmanship is also required in the investigation.

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

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

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

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

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

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

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

   Are you married? Yes/No

   Tick (✓) the right answer.

viii) Questions should be short and clear. That is, they should not be ambiguous and confusing. As far as possible, attempt should be made to suggest the possible answers to a question and the respondent may be asked to simply tick the answer/s he/she thinks is/are correct.

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

   Following is an example of a question:

   Why do people not exercise their right to vote?

   Tick (✓) the right answer:

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

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

   c) The polling booths are far from their residence.

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

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

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

   g) Any other reason, please state.

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

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

A sample questionnaire on family planning is reproduced below.

Survey on Family Planning

1. Name .....
2. Father's / Husband's Name .....
3. Residential address .....
4. Place of Work .....
5. Age ..... 6. Male/Female .....
7. Religion ..... 8. Telephone No. .....
9. Profession:
   a) Self ..... b) Spouse .....
10. Annual Income of the family from all sources .....
11. Educational Qualifications: (Tick (✓) the right answer)
    a) Illiterate b) Primary standard
    c) Middle standard d) Secondary
    e) Sr. Secondary f) Graduate g) Post graduate
12. Educational Qualifications of spouse: (Tick (✓) the right answer)
    a) Illiterate b) Primary standard
    c) Middle standard d) Secondary
    e) Sr. Secondary f) Graduate
    g) Post graduate
13. Number of years of married life .....
14. Number of children born: Girls ..... Boys .....
15. Number of surviving children: Girls ..... Boys .....
16. State the gap in years between the children
    a) Between marriage and first child : .....
    b) Between first and second child : .....
    c) Between second and third child : .....
    d) Between third and fourth child : .....
    e) Between fourth and fifth child : .....
    f) Between fifth and sixth child : .....
17. Do you favour family planning? (Yes/No)
18. If no, what are the reasons?
    a) Children are natural gift: (Yes/No)
    b) Family planning is against my religion: (Yes/No)
    c) Family planning means murdering an unborn child: (Yes/No)
    d) Number of children is the part of my fate: (Yes/No)
    e) Any other reason, please state: .....

19. If you favour family planning, state the reasons
    a) Small family is a happy family: (Yes/No)
    b) Two children can be controlled easily: (Yes/No)
    c) Two children can be properly educated and fed: (Yes/No)
    d) There are fewer complications in life: (Yes/No)
    e) The health of the mother is not adversely affected: (Yes/No)
    f) Any other reason, please state: .....
20. State Age, Educational Level and Health Condition of your children.

Sl.No	Name	Age	Educational Level	Health condition*
1.	.....	.....	.....	.....
2.	.....	.....	.....	.....
3.	.....	.....	.....	.....
4.	.....	.....	.....	.....

(*State whether Poor, Below Normal or Excellent)

How to approach the Respondent with a Questionnaire?

There are three methods available to us:

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

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

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

2.5 COLLECTION OF SECONDARY DATA

As pointed out in Section 2.3.4 that direct investigation, though desirable, is costly in terms of money, time and efforts. Alternatively, information can also be obtained through a secondary source. It means drawing or collecting data from the already collected data of some other agency. Technically, the data so collected are called secondary data.

Limitations of Secondary Data

Although the secondary source is cheap in terms of money, time and effort, utmost care should be taken in their use. It is desirable that such data should be vast and reliable; and the terms and definitions must match the terms and definitions of the

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

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

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

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

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

Secondary data should not only be reliable and suitable, but also *adequate* for the present inquiry. It is always desirable that the available data be much more than required by the inquiry. For example, data on, say, consumption pattern of a state cannot be derived from the data on its major cities and towns.

Check Your Progress 2

1) State, with reasons, whether the following statements are correct or incorrect?
   a) Secondary Data are better than Primary Data.
   b) Data obtained from population census of India 2001 are primary source of data.
   c) Secondary data should not be accepted without scrutiny.
   d) A questionnaire with a very long list of questions is justified.
   e) Of all the survey techniques, the questionnaire method is the best.

.....

.....

.....

2) State two most important characteristics of an investigator when direct personal investigation method is being used for collection of primary data.

3) State and explain various survey techniques.

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

2.6 LET US SUM UP

Data / Statistics are quantitative information and can be distinguished as sample or census data; primary or secondary data.

For conducting an inquiry, we need data which can be collected afresh or from a secondary source. Both require statistical survey which has a planning stage and an executing stage. In the planning stage, the investigator should decide whether to use primary or secondary source, census or sample inquiry, nature of the statistical units and the units of measurement, degree of accuracy desired and so on.

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

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

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

2.7 KEY WORDS

Data Point: It is an observation from an individual or item.

Data Set: It is the collection of all data points.

Census Data: The data obtained by observing all the items of population.

Sample Data: The data obtained by observing only those items which are included in the sample.

Primary Data: Data obtained by observing the items or individuals under the ambit of a problem under consideration.

Secondary Data: Data obtained from the already collected data of some agency.

Questionnaire or Schedule: It is a list of questions that are relevant to the inquiry at hand.

Statistical Inquiry: An inquiry that requires information in the form of figures for its investigation.

Statistical Survey: It is a method for the collection of data by observing all or a sample of items under the ambit of a given problem.

Statistical Unit: It is a characteristic or a set of characteristics of an item that are observed to collect data.

Respondent: The person who supplies the information.

Investigator: The person responsible for the collection of information from respondents.

Hypothesis: It is an assertion or statement about a population.

Test of Hypothesis: Testing the validity of a hypothesis on the basis of collected data.

2.8 SOME USEFUL BOOKS

Elhance, D. N. and V. Elhance, 1988, *Fundamentals of Statistics,* Kitab Mahal, Allahabad.

Nagar, A. L. and R. K. Dass, 1983, *Basic Statistics,* Oxford University Press, Delhi.

Mansfiedl, E., 1991, *Statistics for Business and Economics: Methods and Applications,* W.W. Norton and Co.

Yule, G. U. and M. G. Kendall, 1991, *An Introduction to the Theory of Statistics,* Universal Books, Delhi.

2.9 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) a), b), f) see Sub-section 2.3.3
   c) see Section 2.2
   d) see Sub-section 2.3.2
   e) see Sub-section 2.3.1
2) a) see Section 2.1
   b) see Section 2.2
   c), d), e), h) see Section 2.2
   f) see Sub-section 2.3.2
   g), (i) see Sub-section 2.3.1
   j) see Sub-section 2.3.4
3) See Sub-section 2.3.4

Check Your Progress 2

1) a), b), d) incorrect
   c), e) correct
2) See Section 2.4
3) See Sub-section 2.4
4) See Sub-section 2.5.1

UNIT 3 TABULATION AND GRAPHICAL REPRESENTATION OF DATA

Structure

3.0 OBJECTIVES

On going through this Unit, you will be able to explain:

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

3.1 INTRODUCTION

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

Data can be summarized numerically also. Here we use summary measures like measures of central tendency (such as Arithmetic, Geometric and Harmonic Means, Mode and Median); measures of dispersion (such as Range, Quartile Deviation, Mean Deviation, and Standard Deviation); measures of association in bivariate analysis (such as Correlation and Regression), Index Numbers, etc. In this Unit we plan to discuss how data can be summarized using tables and graphs. Numerical summarization will be discussed in subsequent Blocks (2, 3 and 4). It must be kept in mind that a good summarization and presentation of data is not undertaken for its own sake. It is not an end in itself. In fact it sets the stage for useful analysis and interpretation of data. Again, a good presentation helps us to highlight significant facts and their comparisons. Figures can be made to speak out thereby making possible their intelligent use.

3.2 STAGES OF STATISTICAL INQUIRY

As studied in Section 2.3 of Unit 2, a statistical survey or inquiry is undertaken in two stages, namely, the planning stage and the executing stage. In this Unit we plan to concentrate on some aspects of the executing stage. This involves organising and condensing data in the form of simple array (ascending and descending order), frequency array and continuous frequency distributions, etc.; and presentation of statistical data in the form of tables and graphs.

3.3 ARRANGEMENT OF DATA

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

The first step in the analysis and interpretation of data is its classification and tabulation. The process of arranging data into groups according to their common characteristics is known as its classification. On the other hand tabulation implies a systematic presentation of data in rows and columns according to some salient

In Unit 2, a questionnaire was prepared on Family Planning. Suppose this questionnaire was used to collect information from 50 families of C-III Block of XYZ Colony, New Delhi. Let us assume that it produced that following types of information as given in Tables 3.1 and 3.2. *Can we make any head or tail out of it?*

Table 3.1
Number of Children per family in C-III block, XYZ Colony, New Delhi

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

Table 3.2
Monthly Income of 50 families of C-III block, XYZ Colony, New Delhi

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

As pointed out earlier, to make any head or tail out of the mass of raw data, such as presented above, we have to classify and arrange it. This can be done either by forming a simple array or a frequency array (discrete frequency distribution) or a continuous frequency distribution Sub-sections 3.3.1, 3.3.2 and 3.3.3 attempt to explain this aspect.

3.3.1 Simple Array

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

We can prepare both types of simple arrays from Table 3.1. In the following table, the figures have been arranged in ascending order. From the arrangement, it is clear that the lowest value is 0 and the highest one is 5.

Table 3.3
Number of Children per Family in C-III Block of XYZ Colony, New Delhi
Simple Array — Ascending Order

0	0	0	0	0	1	1	1	1	1
1	1	1	1	1	1	1	2	2	2
2	2	2	2	2	2	2	2	2	2
2	3	3	3	3	3	3	3	3	3

After the arrangement of data in ascending order as in Table 3.3 the raw data make some sense.

The possible conclusions that can be drawn from this arrangement of data (see Table 3.3) are that five families are issueless, twelve families have one child each, fourteen have two children each, ten families posses three children each, six families have four children each and three families have five children each.

3.3.2 Frequency Array or Discrete Frequency Distribution

Here different observations are not repeatedly written as in simple array like 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, etc. We count the number of times (i.e., frequency) an observation repeats itself. For example, in Table 3.3 the observation 4 is repeated 6 times. Thus the frequency of 4 above is 6. The frequency array, for the simple array given in Table 3.3, will look like as given below in Table 3.3.

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

Table 3.4

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

When the number of observations is large enough, the counting process is often undertaken by the use of *tally bars*. In this method, all possible values of the variable are written in a column. For every observation, a tally bar denoted by (|) is noted against its corresponding values. Every fifth repetition is marked by crossing the previous four bars as (𝍸). In this way, we get blocks of five which simplify counting at the end. Thus a number or an observation repeated fourteen times will be marked as (𝍸 𝍸 IIII). Note that after representing each observation by a bar on the *tally sheet*, the same will be ticked (3) or crossed (5) so that it is not duplicated.

The data of Table 3.1 is rewritten in the form of frequency distribution as shown in Table 3.5 below:

Table 3.5
Frequency Distribution of Number of Children per Family

No. of Children	*Tally Sheet*	*Frequency*
0	𝍸	5
1	𝍸 𝍸 II	12
2	𝍸 𝍸 IIII	14
3	𝍸 𝍸	10
4	𝍸 I	6
5	III	3
Total		**50**

3.3.3 Continuous or Grouped Frequency Distribution

Numbers like 1, 2, 3, 4, 5, 20, 40, etc. are discrete numbers and are used where no value between the two consecutive numbers is possible. As in the case of the number of children, it will be impossible as well as funny to say that a particular family has 2.083 or 2.1 or 2.75 number of children. The family can have either 2 or 3 children and not a fraction in between. Out of the two examples of raw data mentioned in Section 3.3, the number of children (Table 3.1) is an example of discrete data while monthly income (Table 3.2) is an example of continuous variable giving continuous data.

In this Sub-section we propose to illustrate the construction of continuous or grouped frequency distribution from the raw data of Table 3.2 on monthly income of the 50 families.

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

Table 3.6
Frequency Distribution of Monthly Income of Families

Monthly Income (Rs.)	*Tally Sheet*	*No. of Families (Frequency)*
500 - 550	𝍸	5
550 - 600	𝍸 I	6
600 - 650	𝍸 𝍸	10
650 - 700	𝍸 𝍸 II	12
700 - 750	𝍸 IIII	9
750 - 800	𝍸	5
800 - 850	III	3
Total		**50**

In Table 3.6 we have completed an exercise where the variable "*income of the family*" has been grouped in order to reduce it to a manageable form called grouped data or *Continuous Frequency Distribution*. However, prior to the construction of any grouped frequency distribution, it is very important to find answers to the following questions:

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

1) What should be the number of class intervals?

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

that the minimum number of classes should not be less than 5 or 6 and in any case, there should not be more than 20 classes.

Usually the formula to determine the number of classes is given by

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

where N is the total number of observations.

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

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

2) **What should be the width of each class interval?**

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

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

$$\textit{Width of a Class} = \frac{\textit{Largest Observation} - \textit{Smallest Observation}}{\textit{Number of Class Intervals}}$$

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

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

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

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

3) **How will the class limits be designated?**

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

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

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

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

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

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

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

The major benefit of exclusive class intervals is that it ensures continuity of data because the upper limit of one class is the lower limit of the next class. In our example on monthly income (Table 3.6), there are 5 families whose income lies between Rs. 500 to Rs. 550, i.e., Rs. 500 to 549 and 6 families whose income lies between Rs. 550 to Rs. 600, i.e., Rs. 550 to 599, and so on. Based on this presumption we can rewrite this frequency distribution in the form of Table 3.7 also.

Table 3.7
Exclusive Class Intervals

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

ii) **Inclusive Method :** In this method, all the observations with magnitude greater than or equal to the lower limit but less than or equal to the upper limit of a class is included in it. Now observe Table 3.8. Income of Rs. 549 is included in the class 500 to 549 so that an income of Rs. 550 automatically goes to the next class of 550 to 599. Since the upper limit of one class is not equal to the lower limit of the following class, this saves us from the confusion whether Rs. 550 goes to (500 to 549) or (550 to 599) class.

Table 3.8
Inclusive Class Intervals

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

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

Table 3.9
Class Boundaries of Inclusive Class Intervals

Monthly Income (Rs.)	*Number of Families* (Frequency)
499.5 - 549.5	5
549.5 - 599.5	6
599.5 - 649.5	10
649.5 - 699.5	12
699.5 - 749.5	9
749.5 - 799.5	5
799.5 - 849.5	3
Total	**50**

Mid-Value of a Class

In exclusive type of class intervals, the mid-value or class mark of a class is defined as the arithmetic mean of its lower and upper limits. However, in case of inclusive class intervals, there is a gap between the upper limit of a class and the lower limit of the following class. This gap is eliminated by adding half of the gap to the upper limit and subtracting half of the gap from the lower limit. The new class limits, thus obtained, are known as *class boundaries*. The class boundaries of the inclusive class intervals in Table 3.8 are given in Table 3.9.

3.3.4 Various Forms of Frequency Distributions

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

a) Open End Frequency Distribution

b) Frequency Distribution with Unequal Class Width

c) Cumulative Frequency Distribution

d) Relative Frequency Distribution

a) Open End Frequency Distribution

Open-end frequency distribution is one which has at least one of its ends open. Either the lower limit of the first class or upper limit of the last class or both are not specified. The words "below" or "less than" and "above" or "more than" are used. In the former the value extends to $-\infty$ and in the latter to $+\infty$. Example of such a frequency distribution is given in Table 3.10.

Table 3.10
Open-end Class Frequency

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

Table 3.11
Unequal Class Frequency

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

b) A Frequency Distribution with Unequal Class Width

The classes of a frequency distribution may or may not be of equal width. A frequency distribution with unequal class width is reproduced in Table 3.11. Here, the width of 1st, 2nd and 5th classes is 5, while that of 3rd is 10 and that of 4th is 15. As we will see in Unit 4, *mode* is not a representative value in such types of series and hence not defined.

c) Cumulative Frequency Distribution

Suppose that, with reference to data given in Table 3.6, we ask the following questions:

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

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

The answers to the above questions can be easily obtained by forming an appropriate cumulative frequency distribution. To answer the first question, we need to form a "less than type" cumulative frequency distribution while a "greater than type" cumulative frequency distribution is required for answering the second question. These distributions are given in Tables 3.12 and 3.13 respectively.

Table 3.12
"Less-than type" Cumulative Frequency Distribution

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

Table 3.13
"More-than type" Cumulative Frequency Distribution

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

d) Relative Frequency Distribution

So far we have expressed the frequency of a value or that of a class as the number of times an observation is repeated. We can also express these frequencies as a *fraction* or a *percentage* of the total number of observations. Such frequencies are known as *the relative frequencies*. Table 3.14 demonstrates the construction of relative frequency distribution.

Table 3.14
Relative Frequency Distribution of Monthly Income of 50 Families

Class	Frequency	Relative Frequency	
		As a fraction	As a percentage
500 - 549	5	5 ÷ 50 = 0.10	0.10 ÷ 100 = 10
550 - 599	6	6 ÷ 50 = 0.12	0.12 ÷ 100 = 12
600 - 649	10	10 ÷ 50 = 0.20	0.20 ÷ 100 = 20
650 - 699	12	12 ÷ 50 = 0.24	0.24 ÷ 100 = 24
700 - 749	9	9 ÷ 50 = 0.18	0.18 ÷ 100 = 18
750 - 799	5	5 ÷ 50 = 0.10	0.10 ÷ 100 = 10
800 - 849	3	3 ÷ 50 = 0.06	0.06 ÷ 100 = 6
Total	**50**	**1**	**100**

From the above table it is clear that sum of the relative frequencies should be either 1 (in case of fraction) or 100 (in case of percentage).

Check Your Progress 1

1) Distinguish between the following, giving at least two points of distinction.
   a) Discrete and continuous frequency distributions
   b) Simple and cumulative frequency distributions
   c) Exclusive and inclusive class intervals
   d) Simple and frequency array

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

2) Explain the following terms giving examples:

a) Ungrouped data

b) Class mark

c) Open end classes

d) Class limits

e) Class boundaries

f) Class frequencies

g) Tally bar

h) Relative frequencies

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

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

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

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

   a) The number of classes, and

   b) Width of the class interval?

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

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

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

.....

.....

.....

.....

.....

.....

.....

.....

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

.....

.....

.....

.....

.....

.....

.....

3.4 TABULATION OF DATA

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

We will discuss presentation of statistical data under three heads.

i) Formal tables

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

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

In this Section we concentrate on tabular forms of presentation.

3.4.1 Meaning and Types of Tables

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

Table 3.15 is based on hypothetical figures of exports and imports of country X with country B for three years 1995, 1996 and 1997.

Table 3.15
Imports and Exports of X with Country B during 1995 - 1997

(In Crore of Rupees)

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

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

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

Types of Tables

Basically, we have two types of tables:

1) Reference tables or general purpose tables

2) Text tables or special purpose tables.

1) *Reference tables* are a general purpose tables and are a store of information with the aim of presenting detailed statistical information. From these tables, we can derive our information (i.e., secondary source). Tables presented by different government departments, ministries, Reserve Bank of India, Economic Surveys, etc. are reference tables and are a routine work of these departments.

Another important example is the Population Census tables prepared by the Registrar General of India giving detailed information on the demographic features of India. Students are advised to consult the latest issue of "Economic Survey" which is issued every year along with the union budget of India. Prepare from it a table on exports and imports of India to USA, UK, Russia, Canada and Germany for three or four years.

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

   Tables can be simple and one way, like the tables given in Section 3.3, where we deal with only one variable, say, income. Alternatively, it is called a univariate frequency distribution. In addition to this, we can have two-way or multi-way tables where we deal with two or more related characteristics (for example, Table 3.15).

3.4.2 Parts of a Table

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

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

2) **Title of the table** gives the indication of the type of information contained in the body of the table. It is said that the *title is to the table what heading is to an essay*. Next to the table number, we mention the title of the table. Its purpose is to answer the questions like:

   a) *What* is in the table?

   b) *Where* is it in the table?

   c) *When* did a particular information occur?

   d) *How* has a particular information been arranged?

In respect of a sample of a table on exports and imports, (Table 3.15), these questions will be answered as below:

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

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

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

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

Dos and Don'ts of the Title

Don't opt for long sentences. Title should be brief and to the point. Present the title in bold letters and/or in capital letters. Expressions used should not convey more than one meaning. Avoid the expressions like 'Table Presents .....' or 'A

Detailed Comparison of Data Relating to .....', etc. It should be like a telegraphic message.

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

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

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

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

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

6) **Main body of the table**, also called *field* of the table, is its most important and bulky part. It contains the relevant numerical information about which a hint is already contained in the title of the table. In our example of Table 3.15 the title amply suggests that the body of the table contains numerical information on exports and imports of country X for a period of three years.

7) **Foot Note**, is a qualifying statement put just below the table (at the bottom). Its purpose is to caution about the limitations of the data or certain omissions. For example, in Table 3.15, the foot note reads that "@ figures are quick estimates". This implies that the figures for the year 1997 where a superscript '@' is given are not final.

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

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

Table 3.16

(———— TITLE ————)

(In Crore of Rupees)

Stub Head	**———— Caption ————**			
	Column Head I		**Column Head II**	
	Sub-head	**Sub-head**	**Sub-head**	**Sub-head**
Stub Entries	MAIN	BODY OF	THE	TABLE
Total				

Footnote(s) :
Source :

3.4.3 Importance of Tables

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

Check Your Progress 2

1) Distinguish between
   a) Caption, stub-head and stub-entries
   b) One-way and two-way tables
   c) Reference tables and text tables
   d) Column entry and row entry
   e) Head note and foot note

.....

.....

.....

.....

.....

.....

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

.....

.....

.....

.....

.....

.....

3) Enumerate the various parts of a Statistical table.

.....

.....

.....

.....

.....

.....

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

Take hypothetical data.

.....

.....

.....

.....

.....

.....

3.5 GRAPHICAL PRESENTATION OF DATA

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

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

3.5.1 Line Graphs

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

Let us take a demand schedule and plot it on the graph. The resultant curve on joining different points, assuming continuity, will give us line graph expressing relation between price and quantity demanded. Such a line graph in Economics is called a *demand curve*. Note that price is measured on Y-axis and quantity demanded on X-axis. The demand curve for data given in Table 3.17 is given in Fig. 3.1.

Table 3.17
Demand Schedule

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

Table 3.18
Time Series Data

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

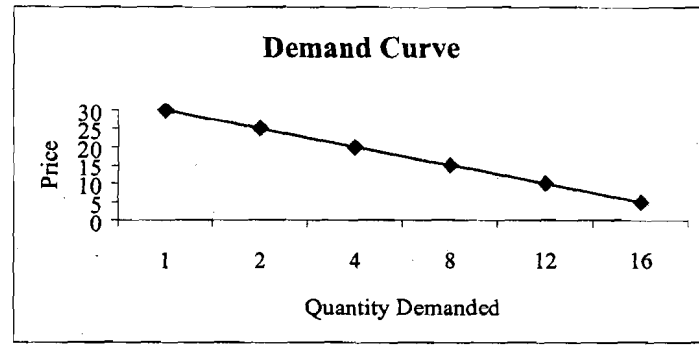


Fig. 3.1

A line graph may be used to show changes in some economic variable, say, steel production over time. In other words, if out of the two variables, one happens to be time (months, years, etc.), we get a line graph over time or simply *time series graph* or *historigram*. A time series expresses behaviour of an economic variable over time. An example of time series data is given in Table 3.18. Measuring years on X-axis and steel production on Y-axis, we can plot time series data on a graph, as shown in Fig.3.2.

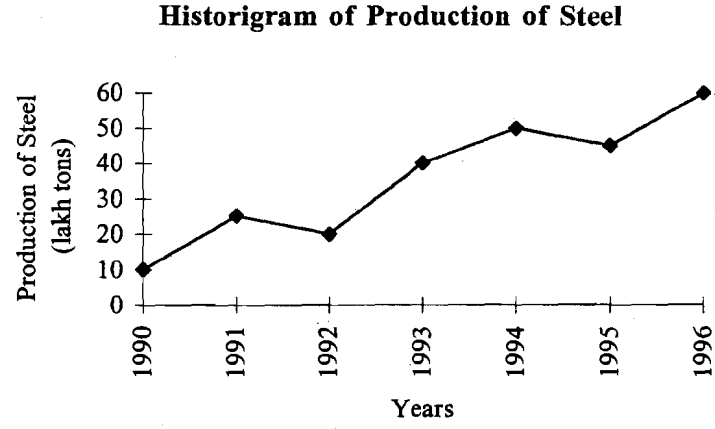


Fig. 3.2

3.5.2 Histogram, Frequency Polygon and Frequency Curve

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

a) It is a rectangular diagram.

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

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

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

Construction of Histogram

To plot a histogram of the frequency distribution given in Table 3.6 on a graph paper, we mark off class intervals like 500 - 550, 550 - 600, etc. on the horizontal axis. Similarly, we mark off frequencies on the vertical axis. Since all the classes

are of equal width, the height of each rectangle is taken to be equal to the frequency of the respective class. The histogram is shown in Fig. 3.3.

Histogram

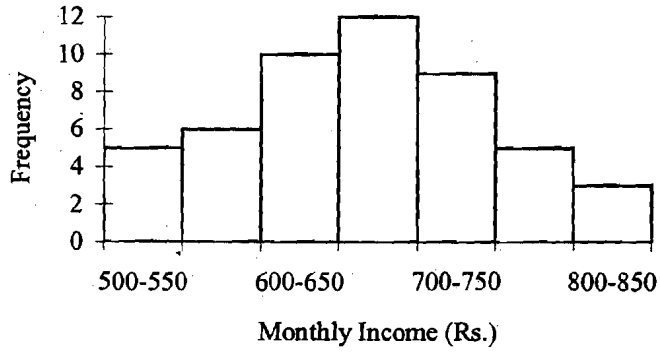


Fig. 3.3

Advantages of histogram are:

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

Frequency Polygon

Frequency Polygon has been derived from the word "polygon" which means many sides. In statistics, it means a graph of a frequency distribution. A frequency polygon is obtained from a histogram by joining the mid-points of the top of various rectangles with the help of straight lines, as shown in Fig. 3.4. In order that total area under the polygon remains equal to the area under histogram, two arbitrary classes, each with zero frequency, are added on both ends, as shown below.

Frequency Polygon

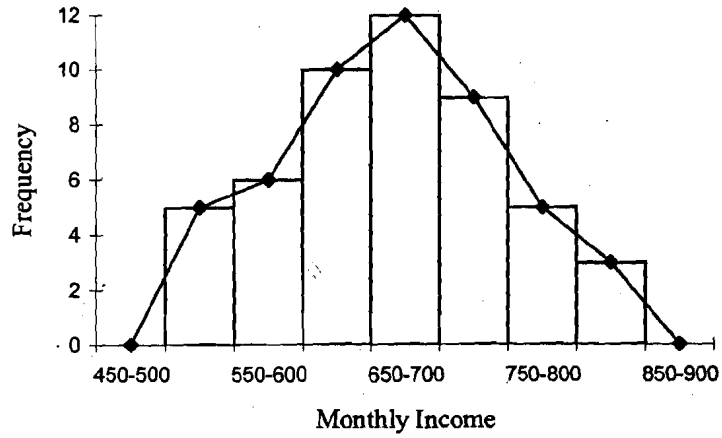


Fig. 3.4

Frequency Curve

If the points, obtained in the case of frequency polygon are joined with the help of a smooth curve, we get a frequency curve as shown in Fig. 3.5.

Frequency Curve

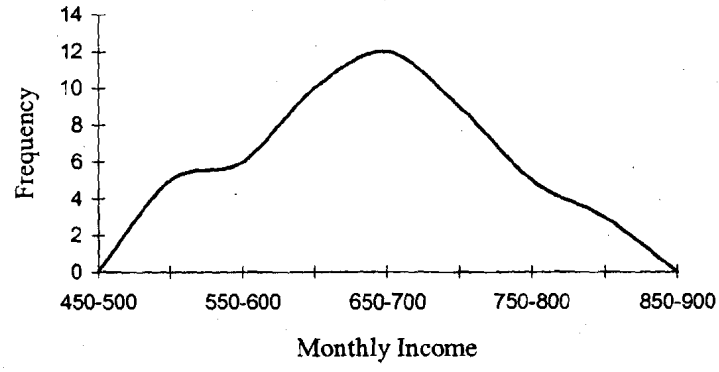


Fig. 3.5

3.5.3 Cumulative Frequency Curve — Ogives

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

Ogives can be used to locate, graphically, certain partition values. We can also determine the percentage of observations lying between given limits. The ogives for the cumulative frequency distributions given in Tables 3.12 and 3.13 are drawn in Fig. 3.6.

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

'Less than' and 'More than' type Ogives

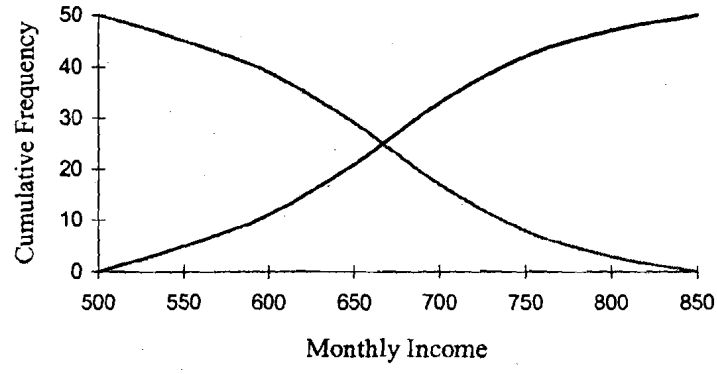


Fig 3.6

3.6 DIAGRAMMATIC PRESENTATION OF DATA

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

3.6.1 One Dimensional Diagrams

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

Table 3.19
Number of students in four zones of a country

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

The bar diagram of the above data is drawn in Fig. 3.7. To make the bar diagram beautiful we can either colour the bars or shade them in different ways. This is left to the aesthetic taste of the investigator.

Bar Diagram

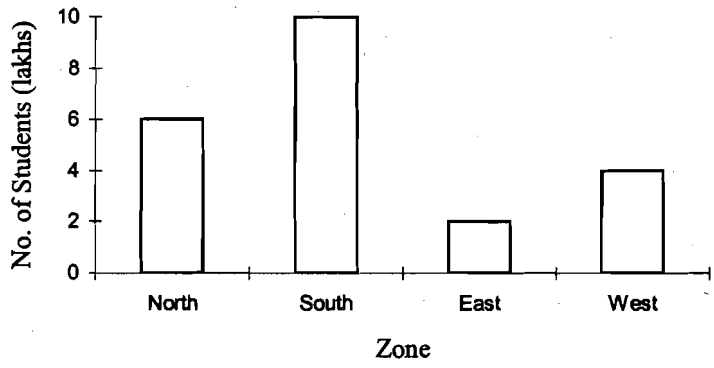


Fig. 3.7

Sub-divided or Component Bar Diagram

A sub-divided bar diagram is used when it is desired to represent the comparative values of different components of a phenomenon. In this diagram, the bars, corresponding to each phenomenon, is divided into various components. The portion of the bar occupied by each component denotes its share in the total. The sub-divisions of different bars must always be done in the same order and these should be distinguished from each other by using different colours or shades. A sub-divided bar diagram for the hypothetical data on sales of T.V. sets, given in Table 3.20 is drawn in Fig. 3.8.

Table 3.20
Zone-wise sale of T.V. sets (1995-1997)

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

Sub-divided or Component Bar Diagram

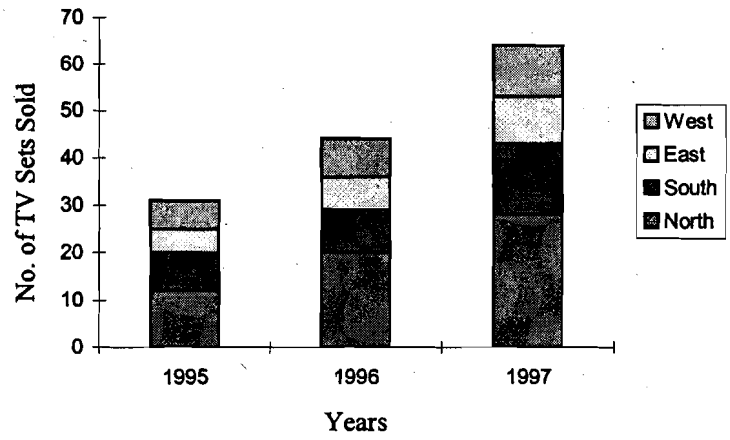


Fig. 3.8

Multiple Bar Diagram

This diagram is used when comparisons are to be shown between two or more sets of data. A set of bars for a period, place or a related phenomenon are drawn side by side without gap. Different bars are distinguished by different shades or colours. A multiple bar diagram for the hypothetical data given in Table 3.21 is drawn in Fig. 3.9.

Table 3.21
Total revenue, total cost and profit of M/S XYZ (1990-92)

(Rupees thousand)

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

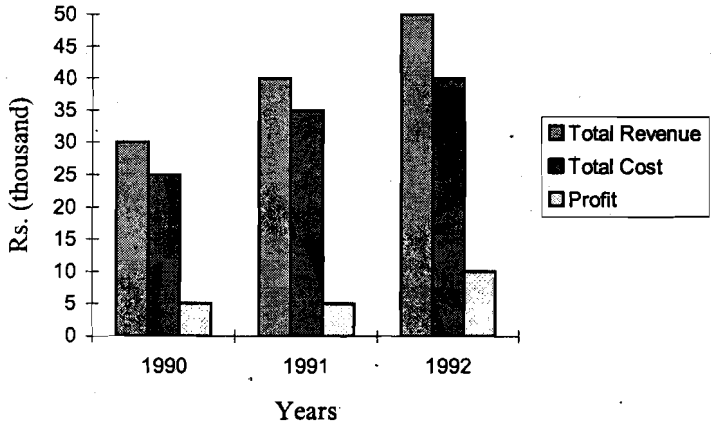


Fig. 3.9

3.6.2 Two Dimensional Diagrams or Area Diagrams

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

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

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

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

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

Table 3.22
Average Salaries of University Teachers as on 1/1/1998

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

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

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

Now take a scale of 2 cm = 100, so that the first rectangle has dimensions of 2 cm. × 5 cm, the second one has the dimensions of 2 cm × 3.2 cm and the third one has the dimensions of 2 cm × 1.8 cm. After this, we are in a position to draw the rectangles as area diagrams (Fig. 3.10).

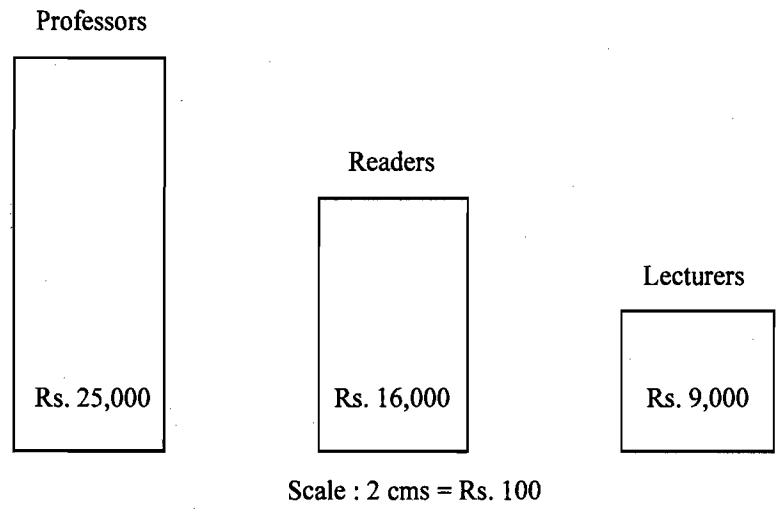


Fig. 3.10: Area Diagram (rectangles)

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

1) $\sqrt{25,000} = 158.114$

2) $\sqrt{16,000} = 126.491$

3) $\sqrt{9000} = 94.868$

Chose a scale 1 cm = 50 so that the first square has each side approximately equal to 3.2 cm. (since 158.114/50 = 3.2), second has the side of 2.53 cm. and the third has the side of 1.9 cm. The relevant squares are drawn in Fig. 3.11.

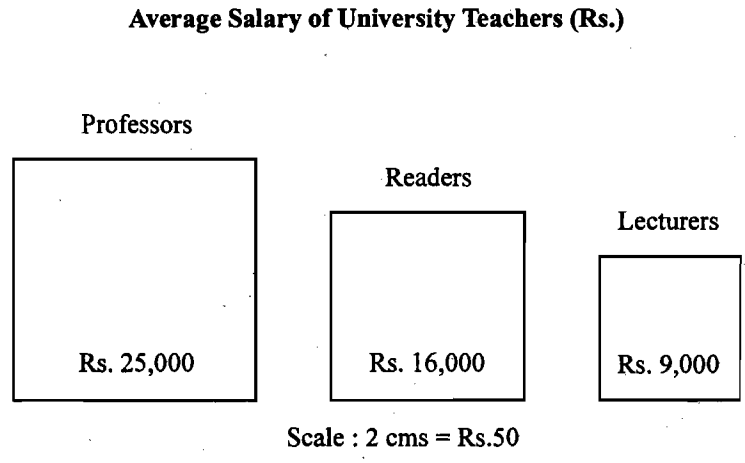
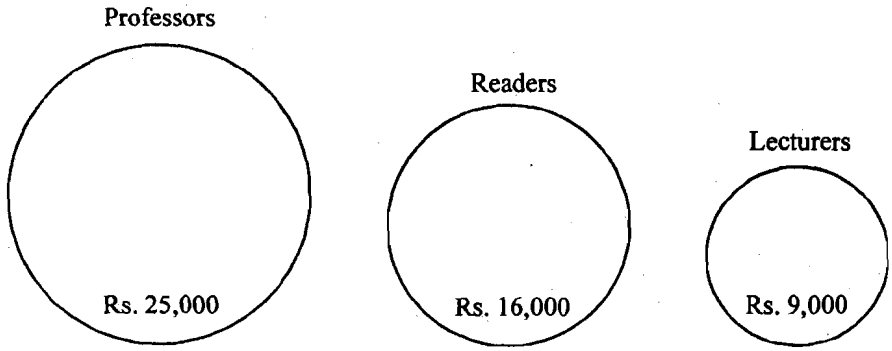


Fig. 3.11: Area Diagram (squares)

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



Scale : 1 cms = 2.5 unit.

Fig. 3.12: Area Diagram (circles)

3.6.3 Pie Diagram or Pie Chart

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

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

Table 3.23
Exports of country X to countries A, B, C and D in 1990

Country	Exports	Percentage Share	Degree
A	300	(300 × 100) ÷ 800 = 37.50	(37.5 × 360°) ÷ 100 = 135°
B	250	(250 × 100) ÷ 800 = 31.25	(31.25 × 360°) ÷ 100 = 12.5°
C	150	(150 × 100) ÷ 800 = 18.75	(18.75 × 360°) ÷ 100 = 67.5°
D	100	(100 × 100) ÷ 800 = 12.50	(12.5 × 360°) ÷ 100 = 45°
Total	**800**	**100**	**360°**

Pie Diagram Representing Exports of X

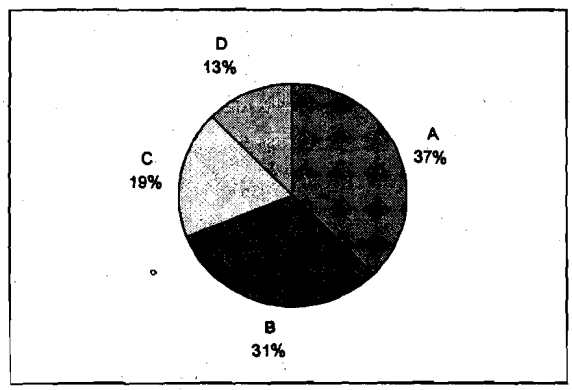


Fig. 3.13

Steps in the construction of Pie diagram

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

3.6.4 Three Dimensional Diagrams

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

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

Table 3.24

	Income class	Income (Rs.)	Cube-root	Side of cube
1.	Poor	216	$\sqrt[3]{216} = 6$	1.5 cms.
2.	Very Rich	3375	$\sqrt[3]{3375} = 15$	3.75 cms.
		Scale : 1 cm. = 4 units.		

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

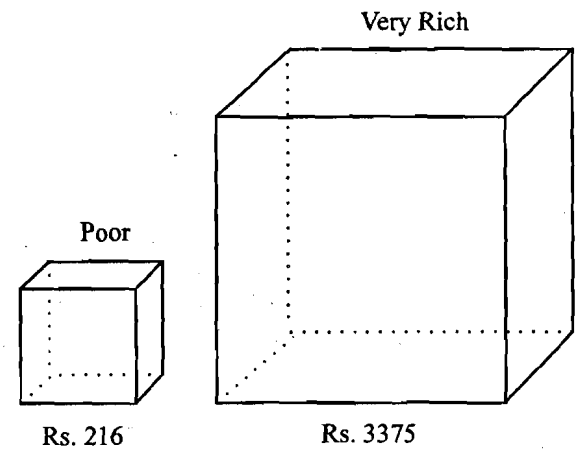


Fig. 3.14

3.6.5 Pictograms and Statistical Maps

These are also known as catrograms. Pictures are more attractive to laymen than other forms of graphic presentations. But these are not suitable everywhere. It may suit cases involving population of people of a state or number of vehicles in a metropolitan city like Delhi or Mumbai. For showing population of human beings, we draw human figures. Here also we have a scale. We may represent 1 lakh people by one human figure so that a population of three and half lakhs is shown by drawing 3½ human figures, as given in Fig. 3.15.

Fig. 3.15

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

Check Your Progress 3

1) Distinguish between the following giving at least two points of distinction.
   a) Histogram and historigram.
   b) Histogram and bar diagram.
   c) Histogram and frequency polygon.
   d) "Less-than" and "More-than" Ogives.
   e) Pie diagram and circle.

.....

.....

.....

.....

.....

.....

.....

.....

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

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

.....

.....

.....

.....

.....

.....

.....

.....

.....

3) Explain the following terms:

a) Line graph

b) Bar diagram

c) Sub-divided or component bar diagram

d) Multiple bar diagram

e) Area diagram

f) Volume diagram

.....

.....

.....

.....

.....

.....

.....

.....

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

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

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

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

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

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

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

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

1) A picture is worth a thousand words.

2) Squares and circles are examples of area diagrams.

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

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

5) A time series graph is known as historigram.

6) Histogram is same as bar diagram.

.....

.....

.....

.....

.....

.....

3.7 LET US SUM UP

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

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

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

3.8 KEY WORDS

Condensation of data: It is a process of classifying and arranging complex and unorganised mass of data to make them fit for comparison and analysis.

Array: An array is an arrangement of data in ascending or descending order. It is also called a simple array.

Frequency array: It is an array or series formed by writing various possible values of the variable along with their respective frequencies.

Discrete frequency distribution: A discrete distribution or discrete series is formed where the variable can take only discrete values like 1,2,3,..... Number of children in a family, number of students in a university, etc. are examples of discrete variable.

Continuous frequency distribution: A continuous frequency distribution is formed where the variable can take any value between two numbers. For example, height, weight, income and temperature.

Inclusive type class interval: A class interval in which all observations lying between and including the class limits are included.

Exclusive type class interval: A class interval which includes all observations that are greater than or equal to the lower limit but less than the upper limit.

Open-end class: A class in which one of the limits is not specified.

Frequency polygon: It is a broken line graph to represent a frequency distribution and can be obtained either from a histogram or directly from the frequency distribution.

Frequency curve: It is a smoothened graph of a frequency distribution obtained from frequency polygon through free hand tracing in such a way that the area under both of them is approximately the same.

Class and class limits: It is a decided group of magnitudes having two ends called class limits or *class boundaries*.

Class range: Also called *class interval.* It is the difference between two limits of a class. It is equal to upper limit minus lower limit. It is also called *class width.*

Mid-point: Also called mid-value. It is the average value of two class limits. It falls just in the middle of a class.

Relative frequency distribution: It is a frequency distribution where the frequency of each value is expressed as a *fraction* or a *percentage* of the total number of observations.

Cumulative frequency distribution: It is obtained by successive totaling of the simple frequencies of a discrete or continuous frequency distribution. This totaling can be done either from above (we get "less-than" cumulative frequency distribution) or from below (we get "more-than" cumulative frequency distribution).

Ogive: It is the graph of cumulative frequency. Graph of "less-than" cumulative frequencies gives "less-than" ogive and that of "more-than" gives "more-than" ogive.

Tabulation: It is a systematic presentation of data in rows and columns.

Caption: It is a part of a table and labels data presented in the column of a table. It is also called *box head.* It may contain one or more than one column head.

Stub: It is a part of a table. It consists of stub head and stub entries. Each stub entry labels a given data placed in the rows of the table. Both *stub head* and *stub entries* appear on the left-hand column of a table. They describe the row heads.

Main body of the table: It is certainly the most important part of the table and contains numerical information about which a hint is already made clear by the title. It is also called *field of the table.*

Line graph: It is the locus of different points obtained with the combinations of X and Y coordinates measured on X-axis and Y-axis respectively.

Historigram: The line graph of a time series is called historigram (For example, steel production since 1950).

Histogram: It is a set of adjacent rectangles presented vertically with areas proportional to the frequencies.

Bar diagram: It is often defined as a set of thick lines corresponding to various values of the variable. It is different from histogram where width of the rectangle is important.

Simple and sub-divided bar diagram: In the case of simple bar diagram only one variable can be presented. A sub-divided bar diagram is used to show various components of a phenomenon.

Pie diagram: It is a circle sub-divided into components to present proportion of different constituent parts of a total. It is also called pie chart.

Area diagrams: These are *two dimensional* diagrams. Here both the height and the base of the diagram are important. That is why they are known as area diagrams. They can be either rectangles, or squares or circles.

Volume diagrams: These are *three dimensional* diagrams. In their construction length, width and height are used. They consist of boxes, cubes, blocks, spheres and cylinders.

Pictographs: Here the data are presented in the form of pictures.

3.9 SOME USEFUL BOOKS

Elhance, D. N. and V. Elhance, 1988, *Fundamentals of Statistics,* Kitab Mahal, Allahabad.

Nagar, A. L. and R. K. Dass, 1983, *Basic Statistics,* Oxford University Press, Delhi

Mansfield, E., 1991, Statistics *for Business and Economics: Methods and Applications,* W. W. Norton and Co.

Yule, G. U. and M. G. Kendall, 1991, *An Introduction to the Theory of Statistics,* Universal Books, Delhi.

3.10 ANSWER OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) a) See Sub-section 3.3.2 and 3.3.3
   b) See Sub-section 3.3.4
   c) See Sub-section 3.3.3
   d) See Sub-section 3.3.1 and 3.3.2
2) You may give examples from your surrounding. For exact meaning of the terms refer to Section 3.3.
3) In the text we have converted the monthly income data in Table 3.2 to a frequency distribution in Table 3.6. From this you can take a clue.
4) Refer to Sub-section 3.3.3
5) Refer to Sub-section 3.3.4(c)
6) Refer to Sub-section 3.3.4(d)

Check Your Progress 2

1) Refer to Table 3.16 and Sub-Section 3.4.2 for different parts of a table.
2) Refer to Sub-section 3.4.2(2)

3) Refer to Table 3.16

4) It can be presented in more than one ways. We have given one below. Try another.

Division of Students of XY College

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

Check Your Progress 3

1) a) See Sub-section 3.5.1 and 3.5.2
   b) See Sub-section 3.5.2 and 3.6.1
   c) See Sub-section 3.5.2
   d) See Sub-section 3.5.3
   e) See Sub-section 3.6.2 and 3.6.3
2) Refer to Sub-sections 3.6.1 and 3.6.3
3) a) See Sub-section 3.5.1
   b) See Sub-section 3.6.1
   c) See Sub-section 3.6.1
   d) See Sub-section 3.6.1
   e) See Sub-section 3.6.2
   f) See Sub-section 3.6.4
4) a) angular
   b) y-axis
   c) geometric
   d) a frequency polygon
   e) ogive
   f) columns
5) True: 1,2,5
   False: 3,4,6

UNIT 4 MEASURES OF CENTRAL TENDENCY

Structure

4.0 OBJECTIVES

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

- compute numerical quantities that measure the central tendency of a set of data such as, mean, median, mode, geometric mean and harmonic mean, and
- use these measures.

4.1 INTRODUCTION

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

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

4.2 MEASURES OF CENTRAL TENDENCY

In frequency distributions of observations discussed in Unit 3 we notice that the observations tend to cluster around a central value. This phenomenon of clustering around a central value in a frequency distribution is called '*Central Tendency*'. Thus, it is of interest to locate such a value around which clustering of observations takes place. There are several measures of central tendency (or location) of a frequency distribution. These measures produce numbers that summarise a frequency distribution in terms of one of its properties, namely, central tendency.

4.2.1 Arithmetic Mean

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

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

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

Example 4.1: Compute the mean for discrete frequency distribution of Table 4.1.

Table 4.1
Frequency distribution of 100 households by size

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

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

$$\bar{X} = \frac{\sum_{i=1}^{n} f_i X_i}{N} = \frac{1\times 3 + 2\times 16 + 3\times 25 + 4\times 33 + 5\times 12 + 6\times 7 + 7\times 2 + 8\times 2}{100} = \frac{374}{100} = 3.74.$$

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

Example 4.2: Compute the mean for grouped frequency distribution of Table 4.2.

Table 4.2
Frequency distribution of 100 households by average monthly household expenditure on food

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

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

Class interval (Rs.) (0)	Mid-point (X_i) (1)	Frequency (f_i) (2)	$f_i X_i$ (3)
262.5 - 286.5	274.5	1	274.5
286.5 - 310.5	298.5	14	4179.0
310.5 - 334.5	322.5	16	5160.0
334.5 - 358.5	346.5	28	9702.0
358.5 - 382.5	370.5	26	9633.0
382.5 - 406.5	394.5	15	5917.5
Total		**100**	**34866.0**

Thus, mean of monthly average household expenditure on food is

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

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

For $i = 1, 2, \ldots, n$

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

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

We will explain this method by re-computing the mean of the monthly average household food expenditure data given in Table 4.2. We construct Table 4.3 by using A and h as explained below.

We define A = Mid-point of the class with largest frequency = 346.5 and
h = Common length of each class interval = 24.

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

Table 4.3
Computation of mean of data of Table 4.2

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

We find out that

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

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

Properties of Arithmetic Mean

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

   Let $X_1, X_2, \ldots . X_n$ be n observations with respective frequencies as f_1, $f_2, \ldots . f_n$. Mathematically, this property implies that $\sum_{i=1}^{n} f_i (X_i - \bar{X}) = 0$,

   where $X_i - \bar{X}$ is the deviation of i^{th} observation from mean.

   To prove the above property, we write

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

Hence the result.

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

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

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

To prove this property, we note that the magnitude of S will depend upon the selected value of A. Thus, we can say that S is a function of A. We want to find that value of A for which S is minimum. Using calculus, this value is given by the equation $\frac{dS}{dA} = 0$ such that $\frac{d^2S}{dA^2} > 0$.

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

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

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

This implies that

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

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

4.2.2 Median

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

When a data set has non-distinct observations — a situation more common in practice — difficulties may arise. In such situations, it may not be always possible to locate the middle most value or the central point that divides the distribution into two equal halves, For example, in the case of the data set having

5 observations 2, 9, 9, 12, 19 the value 9 is repeated twice. Thus, a formal definition of median is needed to overcome such difficulties.

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

Let us find out the median household size from the frequency distribution in Table 4.1. We notice that 77 (out of 100) households have family size of less than or equal to 4 and 56 households have family size of more than or equal to 4. Thus median family size in this case is 4.

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

This method of computing median is illustrated through the data on monthly average household expenditure on food given in Table 4.2.

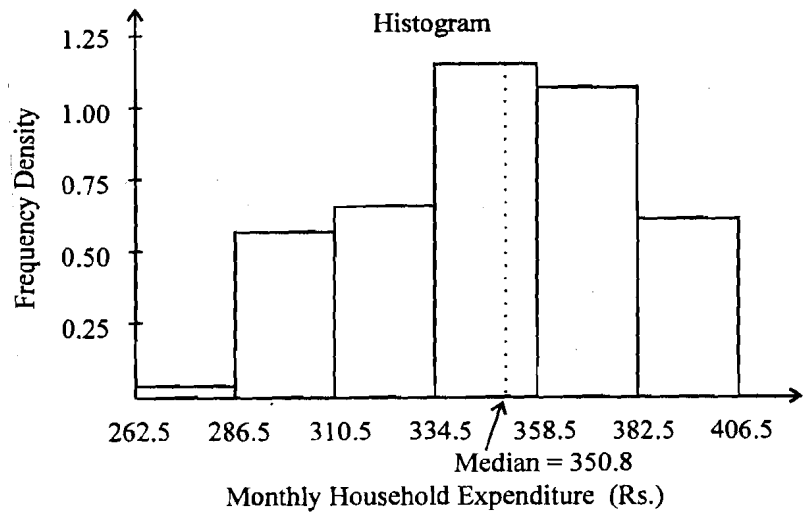


Fig. 4.1

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

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

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

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

N is the total frequency,

C is the cumulative frequency of classes preceding the median class (note that $C = 31$ in the above example),

f_m is the frequency of median class, and

h is the width of median class.

4.2.3 Mode

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

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

From Table 4.1 we find that the mode or modal value of household size is 4 as this value occurs with largest frequency of 33 among 100 households.

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

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

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

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

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

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

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

h is the width of the modal class.

Let us look back to Table 4.2. Here modal class is 334.5 - 358.5 as it has the highest frequency, 28.

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

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

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

Check Your Progress 1

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

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

Find the mean, median and mode.

.....

.....

.....

.....

.....

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

Frequency Distribution of IQ for 309 Six-Year old Children

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

.....

.....

.....

.....

.....

.....

.....

.....

4.3 OTHER MEASURES OF CENTRAL TENDENCY

Besides the arithmetic mean, median and mode there are other averages which are relatively unimportant but may be appropriate in particular situations. These are Geometric Mean and Harmonic Mean. We will discuss these in Section 4.3.1.

Often we see that all the observations do not have equal importance. In such cases we need to give differential importance to different items. Here we use weighted means — arithmetic, geometric or harmonic — instead of simple means. This we will discuss in Section 4.3.2.

4.3.1 Geometric Mean and Harmonic Mean

Often we have to deal with data that are time dependent, i.e., time series data which are unlike one-time data of Tables 4.1 and 4.2. For time dependent data, it is often of interest to find the pattern of change over time. Consider the following two data sets.

Set I : 1000 1100 1200 1300 1400 1500 1600
Set II : 1100 1210 1331 1464 1611 1772 1949

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

Second set looks more like his gross salary (in Rs.). Annual increase in the two sets are given below.

Set I : 100 100 100 100 100 100
Set II : 110 121 133 147 161 177

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

Set I : 1000 1100 1200 1300 1400 1500 1600
Set II : 1100 1241.5 1383 1524.5 1666 1807.5 1949

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

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

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

Clearly, GM is not defined unless all the n numbers are positive. By taking logarithm of GM, one has

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

which shows now GM can be computed by using a log-table. Anti-logarithm of the arithmetic mean of log X values is GM. For the second data set, gross salary increased at the rate of 11% every year. In practice, however, increase/decrease will not be at a fixed rate over the years; and it is meaningful to talk about average rate because fixed rate situation is rare. In general, GM is more appropriate average for percentage (or proportionate) rates of change than arithmetic mean as in the case of rise in various price indices, cost of living indices, etc.

Finally, we discuss about another measure of location called harmonic mean (HM). This mean comes naturally in many situations as in the following illustration. A stockist stocks Rs. 5000 worth of an item at the beginning of every month. Unit rate (in Rs.) of the item for five successive months had been 10.75, 11.80, 14.00, 11.45

and 12.00. The stockist wants to find average rate per unit of the item he has stocked for five months. Computation is presented below :

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

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

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

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

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

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

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

Note that HM is not defined when any observation is zero.

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

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

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

$$= \frac{10.75 + 11.80 + 14.00 + 11.45 + 12.00}{5} = \text{AM of the given prices.}$$

4.3.2 Weighted Means

For many practical applications weighted means (arithmetic, geometric or harmonic) reflect phenomenon more clearly than unweighted or simple means that have been

computed so far. For computation of, say, consumer price index, not all commodities are equally important. Increase in fuel cost may affect consumer price index more than an increase in agricultural prices. For stock market, stock of some key companies may be a trend setter. Weighted means are more appropriate in such situations. To find weighted mean, a weight w_i is attached to each X_i and the means are computed as if w_i's are, symbolically, frequencies of the corresponding X_i's. The computational formulae are as given below:

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

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

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

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

4.3.3 Pooled Mean

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

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

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

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

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

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

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

4.3.4 Choosing a Measure of Central Tendency

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

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

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

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

4.4 PERCENTILES

Concept of percentiles will be explained by using mainly Table 4.2 data on average monthly household expenditure. Percentiles are used in two directions, depending on the question to be answered. Direction of a question may be, what per cent of households have monthly average food expenditure upto Rs. 350.80? Or it may be, what is the maximum monthly average food expenditure of the lower 50% of the households? Note, from our earlier computation of median of Table 4.2 distribution, that the answer to one question is the figure in the other, i.e., 50% of the households have Rs. 350.80 as maximum average monthly food expenditure. Depending on interest, percentage below a cut-off point may be called for : when a poverty line is decided, it is of interest to know the percentage below the poverty line. In the other direction, it may also be of interest to find the status of lower 10% or upper 5% of the population. These are answered by using what are called percentiles.

4.4.1 Percentile: Definition and Computation

For any given percentage v, vth percentile is P_v, a value of the variable being studied, so that at least v percent of the observations are less than or equal to P_v and at least (100 - v) percent of the observations are greater than or equal to P_v.

For example, for Table 4.1, distribution of household size, $P_v = 5$ for any v from 78 to 89.

For grouped data, percentiles are more clearly understood when one looks at the cumulative distribution function. Let $F(X)$ be the proportion of observations less than or equal to X. Any given value X_0 is then the 100 $F(X_0)$th percentile. For Table 4.2, class boundaries, one has $F(286.5) = 0.01$, $F(310.5) = 0.15$, $F(334.5) = 0.31$, $F(358.5) = 0.59$ and $F(382.5) = 0.85$, and consequently Rs. 286.5 = P_{10}, Rs.310.5 = P_{15}, Rs. 334.5 = P_{31}, Rs. 358.5 = P_{59}, and Rs. 382.5 = P_{85}. Note that any amount less than Rs. 262.5 (lower boundary of first class interval) is zero-th percentile and any amount more than Rs.406.5 (upper boundary of last class interval) is 100th percentile.

4.4.2 Quartiles and Deciles

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

25th percentile = $P_{25} = Q_1$ = first quartile
50th percentile = $P_{50} = Q_2$ = second quartile
75th percentile = $P_{75} = Q_3$ = third quartile
10th percentile = $P_{10} = d_1$ = first decile
20th percentile = $P_{20} = d_2$ = second decile, etc., and
$P_{50} = Q_2 = d_5$ = median.

The formulae for Q_1 and Q_3 are similar to the formula for the median. These can be directly written as given below.

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

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

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

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

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

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

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

Check Your Progress 2

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

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

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

217.40 330.50 682.55 1263.59 2249.63

Find the Geometric Mean of the earnings.

.....

.....

.....

.....

.....

.....

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

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

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

.....

.....

.....

.....

.....

.....

.....

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

.....

.....

.....

.....

.....

.....

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

.....

.....

.....

.....

.....

.....

4.5 LET US SUM UP

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

4.6 KEY WORDS

Arithmetic Mean : Sum of observed values of a set divided by the number of observations in the set is called a mean or an average.

Frequency Distribution : The arrangement of data in the form of frequency distribution that describes the basic pattern which the data assumes in the mass.

Geometric Mean : It is the mean of n values of a variable computed as the nth root of their product.

Harmonic Mean : It is the inverse of the arithmetic mean of the reciprocals of the observations of a set.

Median : In a set of observations, it is the value of the middlemost item when they are arranged in order of magnitude.

Mode : In a set of observations, it is the value which occurs with maximum frequency.

4.7 SOME USEFUL BOOKS

Elhance, D. N. and V. Elhance, 1988, *Fundamentals of Statistics,* Kitab Mahal, Allahabad.

Nagar, A. L. and R. K. Dass, 1983, *Basic Statistics,* Oxford University Press, Delhi

Mansfield, E., 1991, *Statistics for Business and Economics: Methods and Applications,* W.W. Norton and Co.

Yule, G U. and M. G Kendall, 1991, *An Introduction to the Theory of Statistics,* Universal Books, Delhi.

4.8 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) 5.1, 5, 5
2) 108.48 ; 108.41 ; 111.42

Check Your Progress 2

1) Rs. 1.96 ; Rs. 1.95
2) Rs. 674.31 crores
3) (i) 25.83 years (ii) 24.82 years (iii) 24.86 years (iv) 27.30 years (v) 25.59 years (vi) 28.79 years
4) Arithmetic Mean, 38.25 days
5) Harmonic Mean, 12%.

UNIT 5 MEASURES OF DISPERSION

Structure

5.0 OBJECTIVES

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

- explain the concept of dispersion;
- compute numerical quantities that measure the dispersion of a set of data;
- explain Chebychev's inequality;
- compute the coefficient of variation; and
- find a measure for concentration of certain distribution of data.

5.1 INTRODUCTION

In Unit 4 we discussed various measures of central tendency, viz., arithmetic mean, median, mode, geometric mean and harmonic mean. However, in many situations these measures do not represent the distribution of data. For example, look into the following three sets of data:

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

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

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

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

In this Unit you will learn to deal with the concepts and techniques involved in reaching conclusions (making inferences) about a body of data in regard to their distribution over the range of variation of the variable.

5.2 CONCEPT OF DISPERSION

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

There are quite a few measures of dispersion. We discuss them below.

5.2.1 Range

Of all measures of dispersions, range is the simplest. It is defined as *the difference between the largest and the smallest observations*. Thus for the data given at Set A the range is 44 – 2 = 42. Similarly, for Set B the range is 17 – 17 = 0 and for Set C it is 11. Now let us look into some grouped data. For Table 4.2 data (look back to the previous Unit), the range is Rs. 406.5 – Rs. 262.5 = Rs. 144. Notice that, for grouped data, largest and the smallest observations are not identifiable. Hence we take *the difference between two extreme boundaries of the classes*.

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

5.2.2 Inter-quartile Range

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

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

Inter-quartile range is the range of the middle most 50% of the observations. If the observations are compact around median, i.e., a strong mode close to the median exists 'inter-quartile range will be smaller than half of the range. If the data are

flat, having no central tendency, this measure will be large, and its value will be close to half of the range.

Let us look in to the discrete data given in Table 4.1 of the previous Unit. Here, $P_{75} = 4$ and $P_{25} = 3$. Hence, the inter-quartile range of household size is $4 - 3 = 1$. This shows that a strong central tendency exists in the distribution of household size since range was observed to be 7 (since $8 - 1 = 7$).

For Table 4.2 data, P_{25} of the average monthly expenditure on food was seen to be Rs. 325.50; P_{75} computed similarly works out to be Rs. 377.88 and inter-quartile range is Rs. 377.88 - Rs. 325.50 = Rs. 52.38. Compared to Rs. 52.38, the range was observed to be Rs. 146.00 or 2.79 times larger. This shows not so strong central tendency for average monthly household expenditure on food.

5.2.3 Mean Deviation

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

Among many possibilities to quantify spread or dispersion of observations, one possibility is to use the deviation of observations from some central value. Since mean is the most commonly used measure of central tendency, it is often taken as the central value with reference to which the deviations are computed. These deviations are then suitably combined to get a measure of dispersion.

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

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

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

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

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

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

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

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

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

Following summary values have been computed for two data sets.

	Data Set I	**Data Set II**
Number of observations	7	7
P_{25}	7	7
Median = P_{50}	12	12
P_{75}	17	17
Range	20	20
Inter-quartile range	10	10
Mean	12	12

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

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

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

For data set I

Mean deviation =

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

$$= \frac{9+5+4+0+2+5+11}{7} = \frac{36}{7} = 5.14$$

For data set II

Mean deviation =

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

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

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

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

For household size data of Table 4.1, mean = $\overline{X} = 3.74$. Mean deviation is now computed as

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

$$= \frac{1}{100}\left(3|1-3.74| + 16|2-3.74| + \cdots + 2|8-3.74|\right) = \frac{109.12}{100} = 1.0912.$$

For Table 4.2 distribution on average household expenditure on food, mean = $\overline{X}$ = Rs. 348.66.

The mean deviation =

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

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

5.2.4 Variance and Standard Deviation

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

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

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

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

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

In the same scale of measurement, for example, observations with a variance of 2 are less dispersed than observations with variance more than 2. To talk about a distribution in terms of a measure of central tendency and a measure dispersion, it is a practical need to use both measures in the same unit. Mean and mean deviation are in the same unit. Since each deviation has been squared for

Based on variance, an equally or more popular measure of dispersion in the same unit as that of observations is *standard deviation*, abbreviated as s.d. Standard deviation is defined as the *positive square root of variance*, i.e., s.d. = σ. As it is the positive square root of variance, it cannot be negative.

Let us compute the s.d. for household size data of Table 4.1

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

$\sigma = 1.4115.$

Similarly for Table 4.2 distribution of average monthly household expenditure on food, variance in Rs.-square is given by

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

and s.d. is

σ = Rs. 30.94.

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

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

or

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

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

Variance = Mean of Squares – Square of the Mean

Using the above formulae, you may compute the variance for the data given in Tables 4.1 and 4.2 and verify the earlier results.

The computations of variance may be greatly simplified by changing X_i to

$u_i = \frac{X_i - A}{h}$, as was done in the computation of mean in Unit 4.

Note that, since

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

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

Hence,

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

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

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

Let us compute the variance by applying the above method for the data given in Table 4.2.

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

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

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

The mean of squares of u values =

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

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

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

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

In this context, the single most useful result about the spread of observations based on mean and standard deviation, irrespective of unit of measurement, is due to Chebychev (discussed below in Section 5.3.1).

Check Your Progress 1

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

.....

.....

.....

.....

.....

.....

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

.....

.....

.....

.....

.....

.....

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

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

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

.....

.....

.....

.....

.....

.....

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

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

.....

.....

.....

.....

.....

.....

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

.....

.....

5.3 RELATIONSHIP BETWEEN DISPERSION AND STANDARD DEVIATION

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

5.3.1 Chebychev's Theorem

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

Note that the theorem is not useful for any positive k less than or equal to 1, since $1-\frac{1}{k^2}$ is at the most equal to zero. For other values of k, the minimum proportion can be computed easily. For example, proportion of observations within 1.5 s.d. of the mean is certain to be at least $1-\frac{1}{1.5^2} = 0.556$ or 55.6%. The following figure indicates spread of data based on Chebychev's theorem. For the household size data of Table 4.1, $\bar{X} = 3.74$ and $s = 1.4115$. If we take $k = 2$, we can say that at least $\left[\left(1-\frac{1}{2^2}\right)\times 100\right] = 75\%$ of the households are certain to have their size between $3.74 \pm 2 \times 1.4115$, i.e., between 0.917 and 6.563.

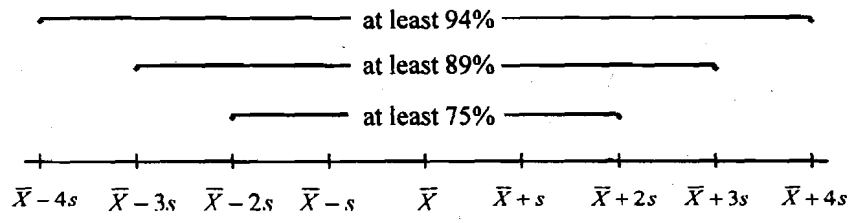


Fig. 5.1

For the Table 4.2 distribution of average monthly household expenditure on food $\bar{X}$ = Rs. 348.66 and s = Rs. 30.94, at least 55.6% (for $k = 1.5$) of households are certain to have monthly average food expenditure between Rs. 302.25 and Rs. 395.07. You can find the relevance of this theorem when we study normal distribution later in Unit 15.

5.3.2 Shape of Distribution

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

5.3.3 Coefficient of Variation

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

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

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

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

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

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

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

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

To compare the extent of inequalities, we compute

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

5.3.4 Concentration Ratio

Above was a comparison of inequality between two villages, without quantifying the level of inequality within each village. If a distribution has a long right tail, it

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

Coefficient of Inequality

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

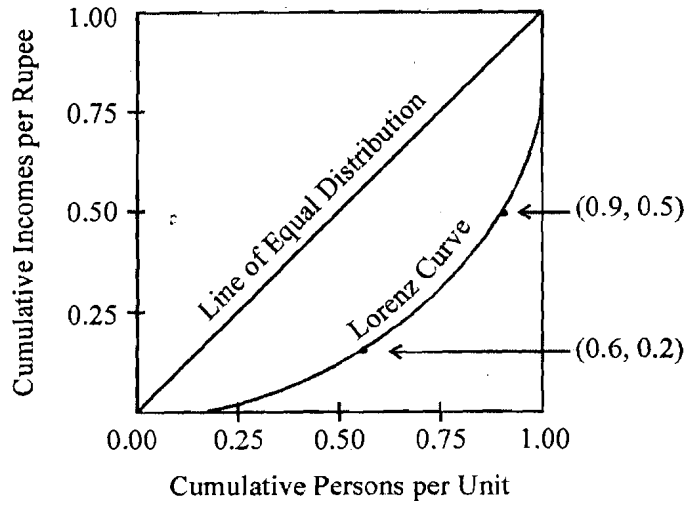


Fig. 5.2

Check Your Progress 2

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

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

Calculate the Variance, Standard Deviation and Coefficient of Variation.

.....

.....

.....

.....

.....

.....

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

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

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

.....

.....

.....

.....

.....

.....

5.4 LET US SUM UP

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

5.5 KEY WORDS

Coefficient of Variation: It is a relative measure of dispersion which is independent of the units of measurement. As opposed to this Standard Deviation is an absolute measure of dispersion.

Mean Deviation: It is the arithmetic mean of absolute deviations (i.e., the differences) from mean or median or mode.

Range: It is the difference between the largest and the smallest observations of a given set of data.

Standard Deviation: It is the positive square root of the variance.

Variance: It is the arithmetic mean of squares of deviations of observations from their arithmetic mean.

5.6 SOME USEFUL BOOKS

Elhance, D. N. and V. Elhance, 1988, *Fundamentals of Statistics,* Kitab Mahal, Allahabad.

Nagar, A. L. and R. K. Dass, 1983, *Basic Statistics,* Oxford University Press, Delhi

Mansfield, E., 1991, *Statistics for Business and Economics: Methods and Applications,* W.W. Norton and Co.

Yule, G U. and M. G Kendall, 1991, *An Introduction to the Theory of Statistics,* Universal Books, Delhi.

5.7 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Do it yourself.
2) 9.9
3) 35, 6.46, 8.85
4) 9.23, 2.49, 2.03
5) 5.0

Check Your Progress 2

1) 4.085, 2.021, 15.004%
2) 23.47%, 25 (rounded figure).

UNIT 6 MEASURES OF SKEWNESS AND KURTOSIS

Structure

6.0 OBJECTIVES

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

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

6.1 INTRODUCTION

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

6.2 CONCEPT OF SKEWNESS

The skewness of a distribution is defined as the lack of *symmetry*. In a symmetrical distribution, the Mean, Median and Mode are equal to each other and the ordinate at mean divides the distribution into two equal parts such that one

part is mirror image of the other (Fig. 6.1). If some observations, of very high (low) magnitude, are added to such a distribution, its right (left) tail gets elongated.

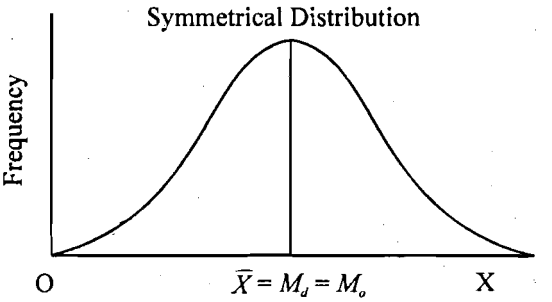


Fig. 6.1

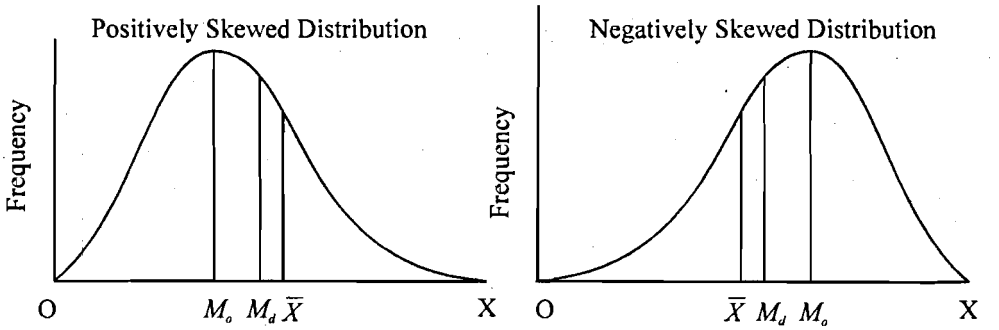


Fig. 6.2

These observations are also known as extreme observations. The presence of extreme observations on the right hand side of a distribution makes it positively skewed and the three averages, viz., mean, median and mode, will no longer be equal. We shall in fact have Mean > Median > Mode when a distribution is positively skewed. On the other hand, the presence of extreme observations to the left hand side of a distribution make it negatively skewed and the relationship between mean, median and mode is: Mean < Median < Mode. In Fig. 6.2 we depict the shapes of positively skewed and negatively skewed distributions.

The direction and extent of skewness can be measured in various ways. We shall discuss four measures of skewness in this Unit.

6.2.1 Karl Pearson's Measure of Skewness

In Fig. 6.2 you noticed that the mean, median and mode are not equal in a skewed distribution. The Karl Pearson's measure of skewness is based upon the *divergence of mean from mode* in a skewed distribution.

Since Mean = Mode in a symmetrical distribution, (Mean - Mode) can be taken as an *absolute measure of skewness*. The absolute measure of skewness for a distribution depends upon the unit of measurement. For example, if the mean = 2.45 metre and mode = 2.14 metre, then absolute measure of skewness will be 2.45 metre – 2.14 metre = 0.31 metre. For the same distribution, if we change the unit of measurement to centimetres, the absolute measure of skewness is 245

centimetre – 214 centimetre = 31 centimetre. In order to avoid such a problem Karl Pearson takes a relative measure of skewness.

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

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

The sign of S_k gives the direction and its magnitude gives the extent of skewness.

If $S_k > 0$, the distribution is positively skewed, and if $S_k < 0$ it is negatively skewed.

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

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

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

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

Example 6.1: Compute the Karl Pearson's coefficient of skewness from the following data:

Table 6.1

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

Table for the computation of mean and s.d.

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

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

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

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

Height (in inches)	No. of persons
57.5 - 58.5	10
58.5 - 59.5	18
59.5 - 60.5	30
60.5 - 61.5	42
61.5 - 62.5	35
62.5 - 63.5	28
63.5 - 64.5	16
64.5 - 65.5	8

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

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

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

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

Thus the distribution is positively skewed.

6.2.2 Bowley's Measure of Skewness

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

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

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

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

The Bowley's coefficient for the data on heights given in Table 6.1 is computed below.

Height (in inches)	No. of persons (f)	Cumulative Frequency
57.5 - 58.5	10	10
58.5 - 59.5	18	28
59.5 - 60.5	30	58
60.5 - 61.5	42	100
61.5 - 62.5	35	135
62.5 - 63.5	28	163
63.5 - 64.5	16	179
64.5 - 65.5	8	187

Computation of Q_1 :

Since $\frac{N}{4} = 46.75$, the first quartile class is 59.5 – 60.5. Thus

$l_{Q_1} = 59.5,\ C = 28,\ f_{Q_1} = 30$ and $h = 1$.

$\therefore \quad Q_1 = 59.5 + \frac{46.75 - 28}{30} \times 1 = 60.125.$

Computation of M_d (Q_2) :

Since $\frac{N}{2} = 93.5$, the median class is 60.5 – 61.5. Thus

$l_m = 60.5,\ C = 58,\ f_m = 42$ and $h = 1$.

$\therefore \quad M_d = 60.5 + \frac{93.5 - 58}{42} \times 1 = 61.345.$

Computation of Q_3 :

Since $\frac{3N}{4} = 140.25$, the third quartile class is 62.5 – 63.5. Thus

$l_{Q_3} = 62.5,\ C = 135,\ f_{Q_3} = 28$ and $h = 1$.

$\therefore \quad Q_3 = 62.5 + \frac{140.25 - 135}{28} \times 1 = 62.688.$

Hence, Bowley's coefficient $S_Q = \frac{62.688 - 2 \times 61.345 + 60.125}{62.688 - 60.125} = 0.048$.

6.2.3 Kelly's Measure of Skewness

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

Kelly's coefficient of skewness, denoted by S_p, is given by

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

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

Note that $P_{50} = M_d$ (median).
The value of S_p, for the data given in Table 6.1, can be computed as given below.

Computation of P_{10} :

Since $\frac{10N}{100} = \frac{10 \times 187}{100} = 18.7$, 10th percentile lies in the class 58.5 – 59.5. Thus

$l_{P_{10}} = 58.5, C = 10, f_{P_{10}} = 18$ and $h = 1$.

$\therefore \quad P_{10} = 58.5 + \frac{18.7 - 10}{18} \times 1 = 58.983.$

Computation of P_{90} :

Since $\frac{90N}{100} = \frac{90 \times 187}{100} = 168.3$, 90th percentile lies in the class 63.5 – 64.5. Thus

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

$P_{90} = 63.5 + \frac{168.3 - 163}{16} \times 1 = 63.831.$

Hence, Kelly's coefficient $S_P = \frac{63.831 - 2 \times 61.345 + 58.983}{63.831 - 58.983} = 0.026.$

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

Check Your Progress 1

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

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

.....

.....

.....

.....

.....

.....

.....

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

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

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

.....

.....

.....

.....

.....

.....

3) The following measures were computed for a frequency distribution :

Mean = 50, coefficient of Variation = 35% and

Karl Pearson's Coefficient of Skewness = - 0.25.

Compute Standard Deviation, Mode and Median of the distribution.

.....

.....

.....

.....

.....

.....

6.3 MOMENTS

The rth moment about mean of a distribution, denoted by μ_r, is given by

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

Thus, rth moment about mean is the mean of the rth power of deviations of observations from their arithmetic mean. In particular,

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

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

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

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

These moments are also known as *central moments*.

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

Let A denote an arbitrary mean, then rth moment about A is defined as

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

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

Moment Measure of Skewness

The moment measure of skewness is based on the property that, for a symmetrical distribution, all odd ordered central moments are equal to zero.

We note that $\mu_1 = 0$, for every distribution, therefore, the lowest order moment that can provide an absolute measure of skewness is μ_3.

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

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

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

Example 6.2: Compute the Moment coefficient of skewness (β_1) from the following data.

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

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

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

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

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

$$\mu_2 = \frac{260}{100} \times 100 = 260, \text{ and}$$

$$\text{s.d. } \sigma = \sqrt{260} = 16.12.$$

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

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

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

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

Conversion of Raw Moments into Central Moments

We can write

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

$$= \frac{1}{N}\sum_{i=1}^{n} f_i\left[\left(X_i - A\right) - \mu_1'\right]^r \qquad \left(\text{Since } \mu_1' = \frac{1}{N}\sum f_i\left(X_i - A\right) = \bar{X} - A\right)$$

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

$$= \frac{1}{N}\sum_{i=1}^{n} f_i\left[{}^rC_0\left(X_i - A\right)^r \mu_1'^0 - {}^rC_1\left(X_i - A\right)^{r-1}\mu_1' + {}^rC_2\left(X_i - A\right)^{r-2}\mu_1'^2 - \cdots\cdots\right]$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i\left(X_i - A\right)^r - {}^rC_1\frac{1}{N}\sum_{i=1}^{n} f_i\left(X_i - A\right)^{r-1}\mu_1' + {}^rC_2\frac{1}{N}\sum_{i=1}^{n} f_i\left(X_i - A\right)^r\mu_1'^2 - \cdots\cdots$$

From the above, we can write

$$\mu_r = \mu_r' - {}^rC_1\mu_{r-1}'\mu_1' + {}^rC_2\mu_{r-2}'\mu_1'^2 - {}^rC_3\mu_{r-3}'\mu_1'^3 + \cdots\cdots$$

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

$$\mu_2 = \mu_2' - {}^2C_1\mu_1'^2 + {}^2C_2\mu_0'\mu_1'^2 = \mu_2' - \mu_1'^2 \quad (\text{since } \mu_0' = 1)$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 3\mu_1'^3 - \mu_1'^3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

Example 6.3: Compute the first four moments about mean from the following data.

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

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

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

From the above table, we can write

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

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

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

$$\mu_4' = \frac{21\times 10^4}{10} = 21000$$

Moments about Mean

By definition,

$$\mu_1 = 0,$$

$$\mu_2 = 90 - 9 = 81,$$

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

$$\mu_4 = 21000 - 4\times(-900)\times(-3) + 6\times 90\times(-3)^2 - 3\times(-3)^4$$

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

Check Your Progress 2

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

Marks	:	0 - 20	20 - 40	40 - 60	60 - 80	80 - 100
Frequency	:	8	28	35	17	12

.....

.....

.....

.....

.....

.....

.....

.....

2) The first three moment of a distribution about the value 3 of a variable are 2, 10 and 30 respectively. Obtain $\overline{X}$, μ_2, μ_3 and hence β_1. Comment upon the nature of skewness.

.....

.....

.....

.....

.....

.....

.....

.....

6.4 CONCEPT AND MEASURE OF KURTOSIS

Kurtosis is another measure of the shape of a distribution. Whereas skewness measures the lack of symmetry of the frequency curve of a distribution, kurtosis is a measure of the relative peakedness of its frequency curve. Various frequency curves can be divided into three categories depending upon the shape of their peak. The three shapes are termed as Leptokurtic, Mesokurtic and Platykurtic as shown in Fig. 6.3.

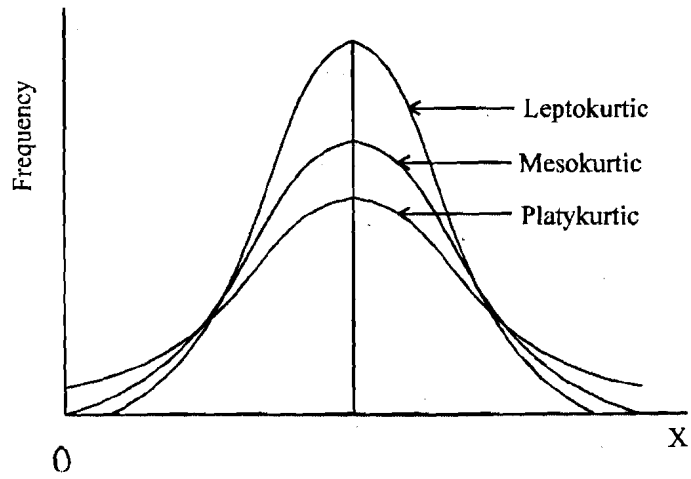


Fig. 6.3

A measure of kurtosis is given by $\beta_2 = \frac{\mu_4}{\mu_2^2}$, a coefficient given by Karl Pearson.

The value of $\beta_2 = 3$ for a mesokurtic curve. When $\beta_2 > 3$, the curve is more peaked than the mesokurtic curve and is termed as leptokurtic. Similarly, when $\beta_2 < 3$, the curve is less peaked than the mesokurtic curve and is called as platykurtic curve.

Example 6.4: The first four central moments of a distribution are 0, 2.5, 0.7 and 18.75. Examine the skewness and kurtosis of the distribution.

To examine skewness, we compute β_1.

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

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

Kurtosis is given by the coefficient $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{18.75}{(2.5)^2} = 3.0$.

Hence the curve is mesokurtic.

Check Your Progress 3

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

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

.....

.....

.....

.....

.....

.....

.....

.....

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

.....

.....

.....

.....

.....

.....

.....

6.5 LET US SUM UP

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

6.6 KEY WORDS

Skewness: Departure from symmetry is skewness.

Moment of Order *r*: It is defined as the arithmetic mean of the *r*th power of deviations of observations.

Coefficient of Kurtosis: It is a measure of the relative peakedness of the top of a frequency curve.

6.7 SOME USEFUL BOOKS

Elhance, D. N. and V. lhance, 1988, *Fundamentals of Statistics,* Kitab Mahal, Allahabad.

Nagar, A. L. and R. K. Dass, 1983, *Basic Statistics,* Oxford University Press, Delhi

Mansfield, E., 1991, *Statistics for Business and Economics: Methods and Applications,* W.W. Norton and Co.

Yule, G. U. and M. G. Kendall, 1991, *An Introduction to the Theory of Statistics,* Universal Books, Delhi.

6.8 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) 0.237
2) – 0.12, - 0.243
3) 17.5, 54.38, 51.46

Check Your Progress 2

1) 0,499.64, 2579.57, 589111.61, 0.053, skewness is positive.
2) 5, 6, -14, 0.907, since μ_3 is negative the distribution is negatively skewed.

Check Your Progress 3

1) 0, 59.99, – 50.18, 8356.64, 0.012 (negatively skewed), 2.32 (platykurtic).
2) 0, 3. Thus the distribution is symmetrical and mesokurtic. Such a distribution is also known as a Normal Distribution.

UNIT 7 PRESENTATION OF BIVARIATE DATA

Structure

7.0 OBJECTIVES

After going through this unit you will be in a position to:

- distinguish between various types of variables;
- present bivariate data in the form of frequency distributions; and
- explain the concepts of marginal and conditional distributions.

7.1 INTRODUCTION

The word 'bivariate' is used to describe situations in which two characteristics are measured on each individual or item, the characteristics being represented by two variables. For example, the measurement of height (X_i) and weight (Y_i) of students in a school. The subscript i in this case represents the student concerned. Thus, for example, X_5, Y_5 represent the height and weight of the fifth student.

Statistical data relating to simultaneous measurement of two variables are called bivariate data. The observation on each individual are paired, one for each variable (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n). In bivariate data, when a large number of pairs of observation are available, it becomes necessary to condense the data in the form of a two-way table, called the *bivariate frequency distribution.*

We discuss below the procedures of preparing bivariate frequency distributions for different types of variables discussed in the previous section.

7.2 TYPES OF VARIABLES

The manner in which we deal with a variable (representing data or any summary measure) depends on the nature of the variable. We distinguish three types of variables for statistical purposes — nominal, ordinal and numerical variables.

A ***nominal*** variable is one which takes qualitative values, which do not have any ordering relationships among them. For example, gender is a nominal variable taking only the qualitative values, male and female; there is no ordering in 'male' and 'female' status. A nominal variable is also called an *attribute*. Here we can divide the observations into categories. However, we cannot say that one category is higher than the other.

An ***ordinal*** variable is one which takes qualitative values having an ordering relation among them. For example, education is an ordinal variable taking, for instance, the qualitative values, illiterate, literate but below senior secondary, senior secondary, graduate, postgraduate. There is an ordering or ranking among them, the educational level being considered *higher* as we move from the first category mentioned to the last.

A ***numerical*** variable is one which takes quantitative values. Numerical variables can be of two types: *discrete* and *continuous*. Discrete variable is one which takes only values at certain isolated points. For example, the number of children in a family is a discrete variable, taking values 0, 1, 2, ... The number of isolated values that a discrete variable can take need not be *finite* in number. On the other hand, a *continuous* variable can take any value in an interval. For example, height is a continuous variable, which conceptually can take any value in the interval, say 0 to 200 centimeters.

7.3 PRESENTATION OF NOMINAL AND ORDINAL VARIABLES

You have already studied how to represent data on a single variable. Let us now consider two variables. Let us consider students of a college and find out two variables: gender and mother tongue of each student. Note that both the variables considered here are nominal: gender taking values male (M) and female (F); and mother tongue taking values, say Hindi (H), Bengali (B), Tamil (T) and Other (O). Then the data set is of the form:

(H, M), (B, M), (T, M), (O, M), (H, F), (B, F), (T, F),(O, F),

where (H, M) denotes Hindi Male. We can summarise these data, without losing any information, in the form of a table as in Table 7.1.

Table 7.1 presents the *joint frequency distribution* of two variables—mother tongue and gender. In a table of this sort (whether the variables are nominal, ordinal or numerical), a combination of the levels or values of the two variables is called a *cell* of the table and the frequency corresponding to it is called the *cell frequency*. For example, you can see from Table 7.1 that (Male, Tamil) is a cell with frequency 367. It means that 367 students of the college are Tamilian males. The joint distribution here gives much more information than the distribution of each variable separately.

The distribution of each variable separately is called the *marginal distribution*. If you look into the last column which adds the male and female frequencies, it gives the marginal distribution of the variable mother tongue. Similarly the marginal distribution of the variable gender is given by the last row. The joint distribution enables us to study the relationship between these two variables.

Table 7.1
Bivariate Frequency Distribution of Mother Tongue and Gender

Mother Tongue	Gender		Total
	Male	Female	
Hindi	456	523	979
Bengali	234	221	455
Tamil	367	387	754
Others	350	401	751
Total	**1407**	**1532**	**2939**

In this example, we may be interested to see if different mother tongue groups have different gender ratios. For this, you have to compute the proportions of males and females for each mother tongue. Such distributions are called *conditional distributions*. Here we find out the distribution of one variable given a particular value of the other variable. For example, the conditional distribution of gender for a given mother tongue Hindi is: Male $0.4658\left(=\frac{456}{979}\right)$; Female $0.5342\left(=\frac{523}{979}\right)$. Similarly the other conditional distributions are: For Bengali: 0.5143; 0.4857, For Tamil: 0.4867; 0.5133; For Others: 0.4660; 0.5340.

Now, let us compare the 'conditional distributions of mother tongue' with the 'marginal distribution of gender'. From the last row of Table 7.1 we see that the marginal distribution of gender is $0.4787\left(=\frac{1407}{2939}\right)$; $0.5213\left(=\frac{1532}{2939}\right)$. Compare this with the conditional distributions obtained earlier. This helps us in understanding the difference in gender distribution between various mother tongue groups.

One may also consider a similar set of computations reversing the roles of two variables. The conditional distribution of mother tongue may be computed for given category of gender, i.e., Male or Female. For Male this turns out to be: Hindi 0.3241 $\left(=\frac{456}{1407}\right)$; Bengali 0.1663; Tamil 0.2608; Others 0.2488. Similarly for Female: $0.3414\left(=\frac{523}{1532}\right)$; 0.1443; 0.2526; 0.2617. These may be compared with the marginal distribution of mother tongue, i.e.,

$0.3331\left(=\frac{979}{2939}\right)$; 0.1548; 0.2565; 0.2555.

These concepts and methods are similar even if one of the variables is ordinal.

If we are interested in finding a cause effect relationship, then the conditional distribution of the dependent variable for each value of the independent variable may be of interest. For instance, if a study is made to explore the relationship between educational level and occupation, one may regard educational level as the cause and occupation as the effect. Hence, we may study the conditional distributions of occupation for each of the educational levels and compare these conditional distributions.

Thus when only nominal and ordinal variables are involved, preparation and presentation of joint frequency distribution in the form of a table are fairly

straightforward; useful information is obtained from marginal and conditional distributions associated with these tables. However, if there are many levels in a variable, tidy presentation in the form a table is difficult. Further, if frequencies corresponding to a level of variable are small, the marginal and conditional relative frequencies associated with these levels are not reliable and hence may not be worth computing, presenting and analysing. In these situations, certain level of variables are pooled and presented, computed and analysed together. In the above example, the level 'others' in the mother tongue variable is an instance of this sort, where all the languages whose frequencies are small are pooled into 'others'.

Check Your Progress 1

1) In each of the following cases, decide if the variable is of nominal, ordinal or numerical type:

   a) The country of citizenship of a tourist coming to India.

   b) The grade obtained by a student in an examination, classified as A^+, A, B, C or D.

   c) The age of an individual.

   d) Price of share of a public limited company at a stock exchange.

   e) Impression of a foreign tourist about India classified as Fantastic, Good, Average, Bad, Terrible.

.....

.....

.....

.....

.....

.....

.....

.....

2) With a suitable example explain the concepts of marginal and conditional distributions.

.....

.....

.....

.....

.....

.....

.....

.....

7.4 PRESENTATION OF NUMERICAL VARIABLES

Let us consider the case of constructing bivariate frequency table from numerical variables. To begin with, we consider two variables: one nominal and one numerical. Later on we will consider the case where both variables are numerical.

a) may not like to give the information, or

b) may not simply be available even after repeated visits.

3) **Error in recording**

This type of error may arise at the stage when the investigator records the answers or even at the tabulation stage. A major reason for such error is the carelessness on the part of the investigator.

4) **Error due to inherent bias of the investigator**

Every individual suffers from personal prejudices and biases. Despite the provision of the best possible training to the investigators, their personal biases may come into play when they interpret the questions to be put to the respondents or record the answers to these questions.

In complete enumeration the extent of non-sampling error tends to be significantly large because, generally, a large number of individuals are involved in the data collection process. We try to minimise this error through:

i) a careful planning of the survey,

ii) providing proper training to the investigators,

iii) making the questionnaire simple.

However, we would like to emphasize that complete enumeration is always prone to large non-sampling errors.

16.4.2 Sampling Error

By now it should be clear that in the sampling method also, non-sampling error may be committed. It is almost impossible to make the data absolutely free of such errors. However, since the number of respondents in a sample survey is much smaller than in census, the non-sampling error is generally less pronounced in the sampling method. Besides the non-sampling errors, there is sampling error in a sample survey. Sampling error is the absolute difference between the parameter and the corresponding statistic, that is, $|T-\theta|$.

Sampling error is not due to any lapse on the part of the respondent or the investigator or some such reason. It arises because of the very nature of the procedure. It can never be completely eliminated. However, we have well developed sampling theories with the help of which the effect of sampling error can be minimised.

16.5 ADVANTAGES OF SAMPLE SURVEY

There are important advantages of a sample survey over complete enumeration or census method. Some of these advantages are mentioned below.

i) **Practicability**

Sometimes, a census may not be practicable due to the enormity of the task required in the collection of data of a large population. In such a situation, a sample survey may be quite practicable.

ii) **Speed**

The data may be collected and summarised faster in a sample survey than in a

census. This may be an important advantage, particularly, when the information is urgently needed.

iii) **Accuracy**

In any survey, census or sample, the required information is obtained by filling in the questionnaires. It has been observed that more accurate results are achieved when the investigators themselves fill in the questionnaire instead of the respondents filling it. Again, personal interviews may result in more accurate information than sending the questionnaires to the respondents by post and requesting them to fill in these questionnaire. Normally, the number of investigators involved in an inquiry varies directly with the number of respondents covered in the inquiry. As a result, personal interviews prove to be easier in the case of a sample survey than in a census. In fact, a sample survey has a greater scope to employ more efficient and better-trained investigators. In the case of a sample survey, the investigators can devote more time to each respondent. Thus, although a sample survey can have less coverage than a census, it may have greater accuracy of the results.

iv) **Cost**

It is obvious that a sample survey results in less expenditure than a complete enumeration. After all, in a survey only part of the population is involved. The cost components of an inquiry are:

a) Overhead cost of the organisation conducting the survey,

b) Cost of collecting the data,

c) Cost of processing and tabulating the data, and

d) Cost of publication of results of the survey.

In this cost break-up, items (b) and (c) are in the nature of variable costs, whereas (a) and (d) are the fixed cost items. As a result, items (b) and (c) will definitely be much smaller in a sample survey than for a census. We should note that the designing of a proper sample survey and the selection of an appropriate sample may entail considerable expenditure. However, generally it has been observed that a sample survey is less costly than complete enumeration.

Check Your Progress 1

1) Define the following concepts:
   a) Population
   b) Sample
   c) Parameter
   d) Statistic

   .....

   .....

   .....

   .....

2) Distinguish between the following:
   a) Estimator and Estimate
   b) Census and Sample Survey
   c) Sampling error and non-sampling error

.....

.....

.....

.....

3) What are the advantages of sampling over a census?

.....

.....

.....

.....

16.6 TYPES OF SAMPLING

The method of selecting a sample from a given population is called *sampling*. Basically there are two types of sampling, viz., probability sampling and non-probability sampling. In probability sampling the sampling units are selected according to some chance mechanism or probability of selection. On the other hand, non-probability sampling is based on judgement or discretion of the person making a choice. Thus in non-probability sampling certain units may be selected because of convenience or they serve a purpose or the researcher feels that these units are representative of the population. No random selection on the basis of chance mechanism is involved here.

16.6.1 Probability Sampling

It is also called random sampling. It is a procedure in which every member of the population has a chance or probability of being selected in the sample. It is in this probabilistic sense that the sample is random. The word 'random' does not mean that the sample is obtained in a haphazard manner without following any rule.

Random sampling is based on the well-established principles of probability theory. There are quite a few variants of the random sampling, viz., simple random sampling, systematic random sampling and stratified random sampling. We discuss these types below.

a) **Simple Random Sampling**

If there is not much variation in the characteristics of the members of a population, we can follow the method of simple random sampling. In this method, we consider the population in its entirety as a homogeneous group and follow the principle of random sampling to choose the members for the sample.

There are two variants of simple random sampling, viz., simple random sampling with replacement (SRSWR) and simple random sampling without replacement (SRSWOR). This difference pertains to the way the sample units are selected. According to the procedure of simple random sampling with replacement (SRSWR), we draw one unit from the population, note down its features and put it back to the whole lot in the sense that the unit again becomes eligible for selection. In this way, the total number of units in the population always remains the same. In other words, the composition of the population remains unchanged, and each member of the population has the *same chance* or probability of being selected

in the sample. In fact, if N is the size of the population, this probability is $\frac{1}{N}$.

On the other hand, in the case of simple random sampling without replacement, the unit once selected is not returned to the population in the sense that it becomes ineligible for selection again. As a result, after each successive draw, the composition of the population changes. Therefore, for subsequent draw from the population the probability of any particular unit being picked up also gets changed. Let us try to understand this. Suppose, the population size is N and we want to draw a sample of size n from it by the principle of SRSWOR. Before the first unit is drawn, each unit of the population has the *same* chance $(\frac{1}{N})$ of being selected in the sample. Once the first member of the sample is selected, each of the *remaining N-1* members of the population has an equal chance of $\frac{1}{N-1}$ of selection in the sample. Finally, before the n^{th} member of the sample is chosen, each of the *remaining* members of the population has an equal chance of $\frac{1}{N-(n+1)} = \frac{1}{N-n-1}$ of being included in the sample.

We should note that from a population of size N, the number of samples of size n that can be drawn with replacement is N^n and the number of samples that can be drawn without replacement is ${}^{N}c_{n}$.

Example 16.1

Suppose a population consists of the following 5 units (4, 5, 7, 9, 10). How many samples of size 2 can be drawn from it?

i) If we follow the procedure of SRSWR the number of samples that can be selected is

$= N^n = 5^2 = 25$.

The possible samples are given by

(4, 4), (4, 5), (4, 7), (4, 9), (4, 10), (5, 4), (5, 5), (5, 7), (5, 9), (5, 10), (7, 4), (7, 5), (7, 7), (7, 9), (7, 10), (9, 4), (9, 5), (9, 7), (9, 9), (9, 10), (10, 4), (10, 5), (10, 7), (10, 9), (10, 10).

We should note that in sampling with replacement, the order in which the units are selected also matters. Thus, (4, 10) and (10, 4) are considered as two different samples.

ii) If we follow the procedure of SRSWOR the number of samples that can be selected is

$$= {}^{N}c_{n} = {}^{5}c_{2} = \frac{5!}{2!(5-2)!} = \frac{5!}{2!\,3!} = \frac{5\times4}{2\times1} = \frac{20}{2} = 10.$$

The possible samples are given by

(4, 5), (5, 7), (7, 9), (9, 10), (4, 7), (4, 9), (4, 10), (5, 9), (5, 10), (7, 10).

We should note that in sampling without replacement, once a member is selected, it cannot be selected again. Thus, samples like (4, 4), (5, 5) etc. cannot be selected. Similarly, if a sample like (4, 5) is selected, then another sample like (5, 4) cannot be selected.

b) **Systematic Random Sampling**

In this variant of random sampling, only the first unit of the sample is selected at random from the population. The subsequent units are then selected by following some definite rule. For example, suppose, we have to choose a sample of agricultural plots. In systematic random sampling, we begin with selecting one plot *at random* and then every 10^{th} plot may be selected.

c) **Stratified Random Sampling**

Stratified random sampling is the appropriate method if the population under consideration consists of heterogeneous units. Here, first we divide the population into certain homogeneous groups or strata. Secondly, from each stratum some units are selected by simple random sampling. Thirdly, after selecting the units from each stratum, they are mixed together to obtain the final sample.

Let us consider an example. Suppose, we want to estimate the per capita income of Delhi by a sample survey. It is common knowledge that Delhi is characterised by rich localities, middle class localities and poor localities in terms of the income groups of the people living in these localities. Now, each of these different localities can constitute a stratum from which some people may be selected by adopting simple random sampling procedure.

d) **Multi-Stage Random Sampling**

Let us consider a situation where we want to obtain information from a sample of households in a large city, say, Delhi. Sometimes, it may not be possible to directly take a sample of households because a list of all the households may not be easily obtained. In such a situation, one may resort to take samples in various stages. Generally, the city is divided into certain geographical areas for administrative purposes. These areas may be termed as city blocks. So in the first stage, some of such blocks may be selected by random sampling. In the next stage, from each of the selected blocks in the first stage, some households may be selected again by the principle of random sampling. In this way, ultimately a sample of households from a large city may be obtained. The above-mentioned example is the case of a two-stage random sampling. However, if the nature of the inquiry so demands, the method of sampling can be extended to more than two stages.

16.6.2 Non-Probability Sampling

We have considered the method of random sampling and some of its variants above. It should be clear that the basic objective of the principle of random sampling is to eliminate or at least minimise the effect of the subjective bias of the investigator in the selection of the population sample. But for certain purposes, there is a need for using discretion. For example, suppose a teacher has to choose 4 participants from a class of 30 students in a debate competition. Here, the teacher may select the top 4 debaters on the basis of her own conscious judgement about the top debaters in the class. This is an example of purposive sampling. In this method, the purpose of the sample guides the choice of certain members or units of the population.

16.6.3 Mixed Sampling

In mixed sampling, we have some features of both non-probability sampling and random sampling. Suppose, an institute has to send 5 students for managerial training in a company during the summer vacation. Initially, it may shortlist about 20 students who are considered to be suitable for the training by applying its own discretion. Then from these 20 students, 5 students may finally be selected by random sampling.

16.7 SAMPLING DISTRIBUTIUON

By now it should be clear that generally the size of a sample is much smaller than the parent population. Consequently, many samples can be selected from the same population which are different from one another. Since an estimate of a parameter depends upon the sample values, and these values may change from one sample to another, there can be different estimates or values of a statistic for the same parameter. This variation in values is called *sampling fluctuation*. Suppose, a number of samples, each of size n, are drawn from a population of size N and for each sample, the value of the statistic is computed. If the number of samples is large, these values can be arranged in the form of a relative frequency distribution. When the number of samples tends to infinity, the resultant relative frequency distribution of the values of a statistic is called the *sampling distribution* of the given statistic.

Suppose, we are interested in estimating the population mean (which is a parameter), denoted by μ. A random sample of size n is drawn from this population (of size N). The sample mean $\bar{x} = \frac{1}{n}\Sigma x_i$ is a statistic corresponding to the population mean μ. We should note that $\bar{x}$ is a random variable as its value changes from one sample to another in a probabilistic manner.

Example 16.2

Consider a population consisting of the following 5 units: 2, 4, 6, 8, and 10. Suppose, a sample of size 2 is to be selected from it by the method of simple random sampling without replacement. We want to obtain the sampling distribution of the sample mean and its standard error.

The number of samples that can be selected without replacement.

$$= {}^{N}C_{n} = {}^{5}C_{2} = \frac{5!}{2!(5-2)!} = \frac{5!}{2!\,3!} = \frac{5\times 4}{2\times 1} = \frac{20}{2} = 10.$$

The possible samples along with the corresponding sample means $(\bar{x})$ are presented in Table 16.1.

Table 16.1 : Possible Samples and Sample Means

Sample	Sample Mean $(\bar{x})$
(2, 4)	3
(2, 6)	4
(2, 8)	5
(2, 10)	6
(4, 6)	5
(4, 8)	6
(4, 10)	7
(6, 8)	7
(6, 10)	8
(8, 10)	9

Now, we can have a frequency distribution of the sample means:

Table 16.2: Frequency Distribution of Sample Means

Sample Mean	Frequency
$(\bar{x})$	(f)
3	1
4	1
5	2
6	2
7	2
8	1
9	1

From the frequency distribution given in Table 16.2, we can present the probability distribution of the sample mean as given in Table 16.3.

Table 16.3: Sampling Distribution of Sample Means

Sample Mean $(\bar{x})$	Probability $\left(\frac{f}{\Sigma f}\right)$
3	$\frac{1}{10}$
4	$\frac{1}{10}$
5	$\frac{2}{10}$
6	$\frac{2}{10}$
7	$\frac{2}{10}$
8	$\frac{1}{10}$
9	$\frac{1}{10}$

We note here that Σf, which, from the frequency distribution of the sample mean presented earlier, is equal to 10. In Table 16.3, we have used the relative frequency for the calculation of the probabilities.

16.8 STANDARD ERROR OF A STATISTIC

In the previous Section we learnt that we can draw a number of samples depending upon the population and sample sizes. From each sample we get a different value for the statistic we are looking for. These values can be arranged in the form of a probability distribution, which is called the sampling distribution of the concerned statistic. The statistic is also similar to a random variable since a probability is attached to each value it takes. In Table 16.3 in the previous Section we have presented the statistic along with its probability.

We have learnt in Unit 14 that mathematical expectation of a random variable is equal to its arithmetic mean. Let us find out the mathematical expectation and standard deviation of the sampling distribution.

We notice two important properties of the sampling distribution.

1) The expectation of the sampling distribution of the statistic is equal to the population parameter. Thus if we have the sampling distribution of sample means, then its expected value is equal to population mean. Symbolically, $E(\bar{x}) = \mu$.

2) The standard deviation of the sampling distribution is called 'standard error' of the concerned statistic. Thus if we have sampling distribution of sample means, then its standard deviation is called the 'standard error of sample means'. Thus standard error indicates the spread of the sample means away from the population mean. In Block 7 we would see that standard error is used for hypothesis testing and statistical estimation.

Example 16.3

Find out the standard error of the sampling distribution given in Table 16.3

We know that standard error of the sample mean is standard deviation of the sampling distribution. Thus,

$$\sigma_{\bar{x}} = \sqrt{E(\bar{x})^2 - [E(\bar{x})]^2}$$

Now,

$$E(\bar{x}) = 3\times\frac{1}{10}+4\times\frac{1}{10}+5\times\frac{2}{10}+6\times\frac{2}{10}+7\times\frac{2}{10}+8\times\frac{1}{10}+9\times\frac{1}{10}=\frac{60}{10} = 6$$

and

$$E(\bar{x})^2 = 9\times\frac{1}{10}+16\times\frac{1}{10}+25\times\frac{2}{10}+36\times\frac{2}{10}+49\times\frac{2}{10}+64\times\frac{1}{10}+81\times\frac{1}{10}=\frac{390}{10}$$

$$= 39.$$

$$\therefore \sqrt{E(\bar{x})^2 - [E(\bar{x})]^2} = \sqrt{39-36} = \sqrt{3} = 1.73.$$

Thus, the standard error of the sample mean in this case is 1.73.

Now a question may be shaping up in your mind.

Do we have to draw all possible samples to find out standard error? In Example 16.3 above we first noted down all the possible samples, arranged these in a relative frequency distribution form and thereafter calculated the standard deviation. In Example 16.3 the population size and sample size were quite small, and thus the task was manageable. But, can you imagine what would happen when we have much larger population and sample sizes? It is too difficult and cumbersome a task. In fact the entire advantages of sampling disappears if we start selecting all possible samples!

Secondly, is it possible to fit a theoretical probability distribution (discussed in Block 5) to the sampling distribution? In fact, the *Central Limit Theorem* says that, "if samples of size n are drawn from any population, the sample means are approximately normally distributed for large values of n". Thus whatever be the distribution of the population, the sampling distribution of $\bar{x}$ will be approximately normal for large enough sample sizes. If the population is normal, then sampling

distribution of $\bar{x}$ is normal for any sample size. If population is approximately normally distributed than sampling distribution of $\bar{x}$ is approximately normal even for small sample size. Moreover, even if population is *not* normally distributed, sampling distribution of $\bar{x}$ is approximately normal for large sample sizes.

Thirdly, what is the relationship between standard deviation of the population from which the sample is drawn and the standard error of $\bar{x}$? Obviously, the spread of $\bar{x}$ will be less than the spread of the population units. The standard error of $\bar{x}$ is given by

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ where $\sigma_{\bar{x}}$ is standard error of $\bar{x}$ and σ is standard deviation of the original population.

Thus standard error is always smaller in value than standard deviation of the population, because standard error is equal to the standard deviation of the population divided by square root of the sample size.

The above is true for simple random sampling with replacement. When sampling is without replacement in that case we have to make some finite population correction and standard error is given by $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times \frac{N-n}{N-1}$.

When the ratio $\frac{n}{N}$ is very small both the procedures give almost similar results. But when sample size is not negligible compared to population size the correction factor needs to be applied.

How do we interpret the standard error? As mentioned earlier it shows the spread of the statistic. Thus, if standard error is smaller then there is a greater probability that the estimate is closer to the concerned parameter.

Example 16.4

Consider the population: 2,5,8,13

i) Calculate the population mean and the population standard deviation.

ii) Construct a sampling distribution of the sample mean when random samples of size 2 are selected from the population
   a) with replacement, and
   b) without replacement. Find the mean and the standard error of the distribution in each case.

iii) Verify that in the case of random sampling with replacement, $E(\bar{x}) = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ and in the case of random sampling without replacement, $E(\bar{x}) = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\left(\sqrt{1-\frac{n-1}{N-1}}\right)$.

Answer:

We have the population: 2,5,8,13; population size $N = 4$; sample size $n = 2$.

i) Population Mean:

$$\mu = \frac{1}{N}\sum_{k=1}^{N} X_k = \frac{2+5+8+13}{4} = 7$$

Population standard deviation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(X_k - \mu)^2} = \sqrt{\frac{(2-7)^2+(5-7)^2+(8-7)^2+(13-7)^2}{N}}$$

$$= \sqrt{\frac{25+4+1+36}{4}} = \sqrt{\frac{66}{4}} = \sqrt{16.5} = 4.06$$

ii) (a) Number of possible samples with replacement = $N^n = 16$.

The samples:

(2,2), (2,5), (2,8), (2,13),
(5,2), (5,5), (5,8), (5,13),
(8,2), (8,5), (8,8), (8,13),
(13,2), (13,5), (13,8), (13,13).

The sample means:

2, 3.5, 5, 7.5,
3.5, 5, 6.5, 9,
5, 6.5, 8, 10.5,
7.5, 9, 10.5, 13.

Sampling Distribution of Sample Means:

$\bar{x}$	f	$\frac{f}{N} = P$(Probability)
2	1	$\frac{1}{16}$
3.5	2	$\frac{2}{16}$
5	3	$\frac{3}{16}$
6.5	2	$\frac{2}{16}$
7.5	2	$\frac{2}{16}$
8	1	$\frac{1}{16}$
9	2	$\frac{2}{16}$
10.5	2	$\frac{2}{16}$
13	1	$\frac{1}{16}$

Mean of the sampling distribution:

$$E(\bar{x}) = \sum_{i=1}^{n} P_i \bar{x}_i \quad \text{(where } \bar{x}_i \text{ is the mean of } i^{th} \text{ sample)}$$

$$= \frac{1}{16}(2+7+15+13+15+8+18+21+13)$$

$$= \frac{1}{16} \times 112 = 7$$

Standard error of the distribution:

$$\sigma_{\bar{x}} = \sqrt{E(\bar{x}^2) - \{E(\bar{x})\}^2}$$

Now,

$$E(\bar{x}^2) = \sum_{i=1}^{n} P_i \bar{x}_i^2$$

$$= \frac{1}{16}(1\times 2^2 + 2\times 3.5^2 + 3\times 5^2 + 2\times 6.5^2 + 2\times 7.5^2 + 1\times 8^2 + 2\times 9^2 + 2\times 10.5^2 + 1\times 13^2)$$

$$= \frac{1}{16}(1 \times 4 + 2 \times 12.5 + 3 \times 25 + 2 \times 42.25 + 2 \times 56.25 + 1 \times 64 + 2 \times 81 + 2 \times 110.5 + 1 \times 169)$$

$$= \frac{1}{16}(4+25+75+84.5+112.5+64+81+221+169)$$

$$= \frac{1}{16} \times 748 = 57.31.$$

And,

$$\{E(\bar{x})\}^2 = 7^2 = 49$$

$$\sigma_{\bar{x}} = \sqrt{57.31-49} = \sqrt{8.31} = 2.83.$$

Thus, the mean and the standard error of the sampling distribution in the case of random sampling with replacement are 7 and 2.83 respectively.

b) Number of possible samples without replacement = ${}^{N}C_n = {}^{4}C_2 = 6$

The samples:

(2,5), (2,8), (2,13), (5,8), (5,13), (8,13).

Sample means:

3.5, 5, 7.5, 6.5, 9, 10.5.

Sampling Distribution of Sample Means:

$\bar{x}$	f	$\frac{f}{N} = P$(Probability)
3.5	1	$\frac{1}{6}$
5	1	$\frac{1}{6}$

7.5	1	$\frac{1}{6}$
6.5	1	$\frac{1}{6}$
9	1	$\frac{1}{6}$
10.5	1	$\frac{1}{6}$

Mean of the population:

$$E(\bar{x}) = \sum_{i=1}^{n} P_i \bar{x}_i$$

$$= \frac{1}{6}(3.5+5+7.5+6.5+9+10.5)$$

$$= \frac{1}{6} \times 42 = 7.$$

Standard error of the distribution

$$\sigma_{\bar{x}} = \sqrt{E(\bar{x}^2) - \{E(\bar{x})\}^2}$$

Now,

$$E(\bar{x}^2) = \sum_{i=1}^{n} P_i x_i^2$$

$$= \frac{1}{6}(3.5^2 + 5^2 + 7.5^2 + 6.5^2 + 9^2 + 10.5^2)$$

$$= \frac{1}{6}(12.25 + 25 + 56.25 + 42.25 + 81 + 110.25)$$

$$= \frac{1}{6} \times 327 = 54.5$$

And we already know,

$$\{E(\bar{x})\}^2 = 7^2 = 49$$

$$\sigma_{\bar{x}} = \sqrt{54.5 = 49} = \sqrt{5.5} = 2.35$$

Thus, the mean and the standard error of the sampling distribution in the case of random sampling without replacement are 7 and 2.35 respectively.

iii) In the case of random sampling with replacement,

$$E(\bar{x}) = 7 = \mu$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4.06}{\sqrt{2}} = \frac{4.06}{1.414} = 2.87 \cong 2.83,$$

as we have independently obtained from the sampling distribution of the sample means.

In the case of random sampling without replacement,

$E(\bar{x}) = 7 = \mu$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{\sigma^2}{n} \times \frac{N-n}{N-1}} = \sqrt{\frac{16.48}{2} \times \frac{2}{3}} = \sqrt{8.24 \times 0.67} = \sqrt{5.52} = 2.35,$$

as we have independently obtained from the sampling distribution of the sample mean

Hence, our results are all verified.

16.9 DESIRABLE PROPERTIES OF AN ESTIMATOR

Suppose, θ is an unknown population *parameter* that we are interested in. We may want to estimate θ on the basis of a random sample drawn from the population. For this purpose we may use a statistic T (which is a function of the sample values). Here T is an *estimator* of θ and the value of T that is obtained from the given sample is an *estimate* of θ. In fact, the value is known as a *point estimate* in the sense that it is one particular value of the estimator (see Unit 18 for details).

Earlier, we have discussed the concepts of sampling and non-sampling errors. We recapitulate here that the absolute difference (ignoring the sign) between a sample statistic and the population parameter, i.e., $|T-\theta|$ measures the extent of the sampling error. We may note here that an estimator is essentially a formula for computing an estimate of the population parameter and there can be several potential estimators (alternative formulae) that may be used for this purpose. So, there should be some desirable properties on the basis of which we can select a particular estimator for estimating the population parameter. A very simple requirement for T to be a good estimator of θ is that the difference $|T-\theta|$ should be as small as possible. Various approaches have been suggested to ensure this.

16.9.1 Unbiasedness

We have already noted that the value of a statistic varies from sample to sample due to sampling fluctuation. Although the individual values of a statistic may be different from the unknown population parameter, on an average, the value of a statistic should be equal to the population parameter. In other words, the sampling distribution of T should have a central tendency towards θ. This is known as the property of unbiasedness of an estimator. It means that although an individual value of a given estimator may be higher or lower than the unknown value of the population parameter, there is no bias on the part of the estimator to have values that are always greater or smaller than the unknown population parameter. If we accept that mean (here, expectation) is a proper measure for central tendency, then T is an *unbiased estimator* for θ if $E(T) = \theta$.

16.9.2 Minimum-Variance

It is also desirable that the average spread of all the possible values of an unbiased estimator around the population parameter is as small as possible. It will reduce the chance of an estimate being far away from the parameter. If we accept that variance is a proper measure for average spread (dispersion), we want that among all the unbiased estimators, T should have the smallest variance. Symbolically, $V(T) \leq V(T')$ where, V stands for variance and T' is any other unbiased estimator.

An estimator T, which is unbiased and among all the unbiased estimator has the minimum variance, is known as a *minimum-variance unbiased estimator*. Let us consider an example. Suppose, we have a random sample of size n from a given population of size N. In this case, the sample mean is given by $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ where x_i is the i^{th} member of the sample. It can be proved that it is an unbiased estimator of the population mean μ. Symbolically

$E(\bar{x}) = \mu$

However, it can be shown that the sample variance defined as $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is not an unbiased estimator of the population variance σ^2. Symbolically,

$E(s^2) \neq \sigma^2$

On the contrary, if we define the sample variance as $s'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, then s'^2 is an unbiased estimator of $\sigma'^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$.

Suppose further that the sample values are not only random but also independent (random sample with replacement) and the underlying population is normal. It can be shown that the sample mean $\bar{x}$ is not only an unbiased estimator of the population mean μ but also it has the minimum variance among all the unbiased estimators of μ.

16.9.3 Consistency and Efficiency

Another approach may be to suggest that the estimator T should approximate the unknown population parameter θ as the sample size n increases. Since T itself is a random variable, we may express this requirement in probabilistic or stochastic terms as the statistic T should converge to the parameter θ stochastically (i.e., in probability) as $n \to \infty$. A statistic T with this property is called a *consistent* estimator of θ.

In real life, a large number of consistent estimators of the same parameter q have often been found. In such a situation, obviously, some additional criterion is needed to choose among these consistent estimators. One such criterion may be to demand that not only T should converge stochastically to θ but also it should do so quite rapidly. Without going into the details, we may mention here that some times an estimator assumes the form of a normal distribution when the sample size n increases indefinitely. Such estimators are called *asymptotically normal*. If we focus on consistent estimators that are asymptotically normal, the rapidity of their convergence is indicated by their respective asymptotic variances. In fact, the convergence is the fastest for the estimator that has the *lowest asymptotic variance*. Such kind of an estimator is known as an *efficient estimator* among all the asymptotically normal consistent estimators of a population parameter.

Check Your Progress 2

1) Define the following concepts:
   a) Simple Random Sample

b) Sampling Distribution

c) Standard Error

.....

.....

.....

.....

2) Distinguish between the following:

a) Simple random sampling with replacement and Simple random sampling without replacement

b) Simple random sampling and stratified random sampling

.....

.....

.....

.....

3) Given a population: 1, 2, 5, 6. Bring out all possible samples of size 2

i) with replacement, and

ii) without replacement.

.....

.....

.....

.....

4) Given a population: 2, 4, 6. Suppose a sample of size 2 is to be selected from this population by the method of random sampling without replacement.

a) Present the sampling distribution of sample mean.

b) Compute the standard error.

.....

.....

.....

.....

16.10 LET US SUM UP

In this unit, we distinguished between the census method and the sample method of conducting a statistical inquiry. We have seen that on account of various resource constraints, census method cannot be undertaken always. Moreover, due to the enormity of the task involved, the chances of committing non-sampling errors in a census are at times quite high. A sample survey, on the other hand, has some definite advantages. A properly conducted sample survey is generally less error prone. A sample survey has a sound scientific basis. As a result, the sampling distribution of the relevant statistic (obtained from random samples) forms an objective basis of assessment about a population parameter.

16.11 KEY WORDS

Estimate : It is the particular value that can be obtained from an estimator.

Estimator : It is the specific functional form of a statistic or the formula involved in its calculation. Generally, the two terms, statistic and estimator, are used interchangeably.

Parameter : It is a measure of some characteristic of the population.

Population : It is the entire collection of units of a specified type in a given place and at a particular point of time.

Random Sampling : It is a procedure where every member of the population has a definite chance or probability of being selected in the sample. It is also called probability sampling.

Sample : It is a sub-set of the population. Therefore, it is a collection of some units from the population.

Sampling Distribution : It refers to the probability distribution of a statistic.

Sampling Error : The absolute difference between population parameter and relevant sample statistic.

Sampling Fluctuation : It is the variation in the values of a statistic computed from different samples.

Simple Random Sampling : This is a sampling procedure, in which, each member of the population has the *same chance* of being selected in the sample.

Standard Error : It is the standard deviation of the sampling distribution of a statistic.

Statistic : It is a function of the values of the units that are included in the sample. The basic purpose of a statistic is to estimate some population parameter.

Statistical Inference : It is the process of drawing conclusions about an unknown population characteristic on the basis of a known sample drawn from it.

16.12 SOME USEFUL BOOKS

Bhardwaj. R. S., 1999, *Business Statistics* (First Edition), Excel Books, New Delhi, Chapter 20.

Nagar, A. L. and Das, R. K., 1988, *Basic Statistics,* Oxford University Press, Delhi, Chapter 9.

Goon, A. M., Gupta, M. K. and Dasgupta, B., 1971, *Fundamentals of Statistics,*

Volume 1 (Fourth Edition), The World Press Private Limited, Calcutta, Chapters 14, 15 and 16.

16.13 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Read the text and define these terms in one or two sentences each.
2) Read the text and distinguish in a few sentences.
3) Read the text and answer in a few sentences.

Check Your Progress 2

1) Read the text and define these terms in a few sentences.
2) Read the text and distinguish in a few sentences.
3) Go through Example 16.2 in the text and attempt yourself.
4) 0.82.

UNIT 8 CORRELATION ANALYSIS

Structure

8.0 OBJECTIVES

After going through this unit, you will be in a position to:

- plot scatter diagram;
- measure covariance between two variables;
- compute correlation coefficient;
- compute rank correlation coefficient; and
- determine whether two variables are correlated.

8.1 INTRODUCTION

In the previous unit we discussed the methods of presentation of bivariate data in the form of frequency distributions. In this unit we deal with the concept of correlation which measures the strength of relationship between two variables. When we compute measures of correlation from a set of bivariate data, our interest focuses on the *degree* and *direction* of the association between the variables.

In statistical studies with several variables, there are generally two types of problems. In some problems it is of interest to study how the variables are interrelated; such problems are tackled using *correlation* techniques. For instance, an economist may be interested in studying the relationship between the stock prices of various companies; for this he may use correlation techniques.

In other problems there is a variable y of basic interest and the problem is to find out what information the other variable provides on Y, such problems are tackled using *regression* techniques. For instance, an economist may be interested in studying what factors determine the pay of an employed person and in particular, he may be interested in exploring what role the factors such as education, experience, market demand, etc. play in determining the pay. In the above situation he may use regression techniques to set up a prediction formula for pay based on education, experience, etc.

While correlation is dealt in the present unit, regression analysis will be covered in the next unit.

8.2 SCATTER DIAGRAM

We first illustrate how the relationship between two variables is studied. A teacher is interested in studying the relationship between the performance in Statistics and Economics of a class of 20 students. For this he compiles the scores on these subjects of the students in the last semester examination. Some data of this type are presented in Table 8.1.

Table 8.1
Scores of 20 Students in Statistics and Economics

Serial Number	*Score in*		*Serial Number*	*Score in*	
	Statistics	*Economics*		*Statistics*	*Economics*
1	82	64	11	76	58
2	70	40	12	76	66
3	34	35	13	92	72
4	80	48	14	72	46
5	66	54	15	64	44
6	84	56	16	86	76
7	74	62	17	84	52
8	84	66	18	60	40
9	60	52	19	82	60
10	86	82	20	90	60

A representation of data of this type on a graph is a useful device which will help us to understand the nature and form of the relationship between the two variables, whether there is a discernible relationship or not and if so whether it is linear or not. For this let us denote score in Economics by X and the score in Statistics by Y and plot the data of Table 8.1 on the x-y plane. It does not matter which is called X and which Y for this purpose. Such a plot is called *Scatter Plot* or *Scatter Diagram.* For data of Table 8.1 the scatter diagram is given in Fig. 8.1.

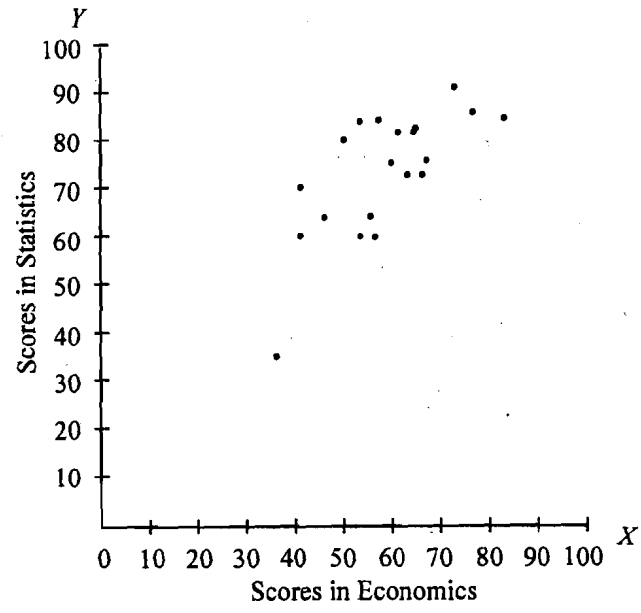


Fig. 8.1: Scatter Diagram of Scores in Statistics and Economics

An inspection of Table 8.1 and Fig. 8.1 shows that there is a *positive relationship* between x and y. This means that larger values of x are associated with larger values of y and smaller values of x with smaller values of y. Further, the points seem to lie scattered around both sides of a straight line. Thus it appears that a linear relationship exists between x and y. However, this relationship is not *perfect* in the sense that there are deviations from such a relationship. It would indeed be useful to get a measure of the strength of this linear relationship.

8.3 COVARIANCE

In the case of a single variable we have learnt the concept of variance, which is defined as

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad \ldots(8.1)$$

In the above we use a subscript x to specify that σ_x^2 represents the variance in x. In a similar manner we can represent σ_y^2 as the variance in y, and σ_x and σ_y as the standard deviation in x and y respectively.

As you know, variance measures the dispersion from mean. In the case of bivariate data we have to reach a single figure which will present the deviation in both the variables from their respective means. For this purpose we use a concept termed covariance, which is defined as follows:

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \qquad \ldots(8.2)$$

You may recall that standard deviation is always positive since it is defined as the positive square root of variance. In the case of covariance there are two terms $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$ which represent the deviations in x from $\bar{X}$ and Y from $\bar{Y}$. Moreover, $(X_i - \bar{X})$ can be positive or negative depending on whether x_i is less than or greater than $\bar{X}$. Similarly $(Y_i - \bar{Y})$ can be positive or negative. It is not necessary that whenever $(X_i - \bar{X})$ is positive $(Y_i - \bar{Y})$ will also be positive. Therefore, the product $(X_i - \bar{X})(Y_i - \bar{Y})$ can be either positive or negative. A positive value for $(X_i - \bar{X})(Y_i - \bar{Y})$ implies that whenever $X_i > \bar{X}$, we have $Y_i > \bar{Y}$. Thus a higher value of x_i is associated with a relatively higher value in y_i. On the other hand, $(X_i - \bar{X})(Y_i - \bar{Y}) < 0$ implies that a lower value in X_i is associated with a relatively higher value in y_i. When we sum it over all the observations and divide by the number of observations, we may obtain a negative or positive value. Therefore, covariance can assume both positive and negative values.

When covariance between x and y is negative ($\sigma_{xy} < 0$) we can say that the relationship could be inverse. Similarly, ($\sigma_{xy} > 0$) implies a positive relationship between x and y. A major limitation of covariance is that it is not independent of unit of measurement. It means that if we change the unit of measurement of the variables we will get a different value for σ_{xy}.

The computation of σ_{xy} as given in (8.2) often involves large numbers. Therefore, it is derived further as

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n}\sum_{i=1}^{n}(X_iY_i - \bar{X}Y_i - \bar{X}\bar{Y})$$

By further simplification we find that

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}X_iY_i - \frac{1}{n}\sum_{i=1}^{n}\bar{X}Y_i - \frac{1}{n}\sum_{i=1}^{n}X_i\bar{Y} + \frac{1}{n}\sum_{i=1}^{n}\bar{X}\bar{Y}$$

Since $\frac{1}{n}\sum_{i=1}^{n}\bar{X}Y_i = \frac{1}{n}\sum_{i=1}^{n}X_i\bar{Y} = \bar{X}\bar{Y}$ we have

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}X_iY_i - \bar{X}\bar{Y} \qquad \ldots(8.3)$$

8.4 CORRELATION COEFFICIENT

The task before us is to measure the linear relationship between x and y. It is desirable to have this measure of strength of linear relationship independent of the scale chosen for measuring the variables. For instance, if we are measuring the relationship between height and weight, we should get the same measure whether height is measured in inches or centimetres and weight in pounds or kilograms. Similarly, if a variable is temperature, it should not matter whether it is recorded in Celsius or Fahrenheit. This can be achieved by standardising each variable, that is by considering $\frac{X-\bar{X}}{\sigma_x}$ and $\frac{Y-\bar{Y}}{\sigma_y}$ where $\bar{X}$ and $\bar{Y}$ are the means of X and Y respectively and σ_x and σ_y are standard deviations.

Let us denote these standardised variables by u and v respectively. Let us also use the notation (X_i, Y_i) to denote the score i^{th} student in Economics and Statistics respectively, i ranging from 1 to n, the number of students, n being 20 in our example. Similarly, let (u_i, v_i) denote the standardised scores of i^{th} student. Then recall the following formulae for mean and standard deviation:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n}\bar{X}_i; \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(\bar{X}_i - \bar{X})^2;$$

$$Y = \frac{1}{n}\sum_{i=1}^{n}X_i; \sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

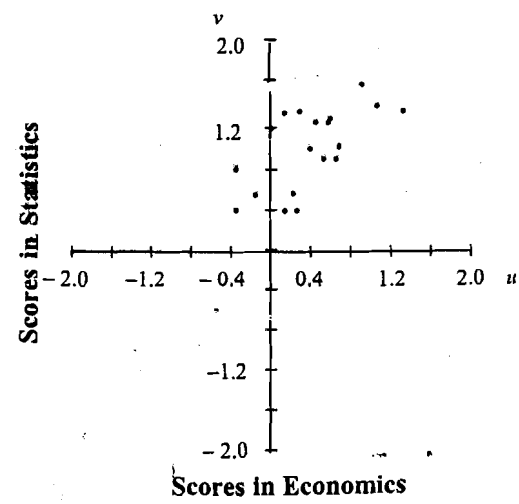


Fig. 8.2: Scatter Diagram of Standardised Scores in Statistics and Economics

Fig. 8.2 is the scatter diagram in terms of standardised variables u and v. Let us observe that in this example there is a positive association between the two scores. The larger one score is, the larger the other score also is; the smaller one score is the smaller the other score is, on the whole. In view of this, most of the points are either in the *first quadrant* or in the *third quadrant*. The first quadrant represents the cases where both scores are above their respective means and third quadrant represents the cases where both scores are below their respective means. There are only a very few points in second and fourth quadrants, which represent the cases where one score is above its mean and the other is below its mean. Thus the product of the u, v values is a suitable indicator of the strength of the relationship; this product is positive in the first and third quadrants and negative in the second and fourth. Thus the product of u, v averaged over all the points may be considered to be suitable measure of the strength of linear relationship between X and Y. This measure is called the *correlation coefficient* between X and Y and is usually denoted by r_{xy} or simply by r, when it is clear what x and y in the context are. This is also called the *Pearson's Product-Moment Correlation Coefficient* to distinguish it from other types of correlation coefficients.

Thus the formula for r is

$$r = \frac{1}{n}\sum_{i=1}^{n} u_i v_i \qquad \ldots(8.4)$$

If we substitute the variables x and y in (8.4) above

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma_x}\right)\left(\frac{Y_i - \bar{Y}}{\sigma_y}\right) = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sigma_x \sigma_y}$$

In the above expression, the term

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

is the *covariance* between x and y (σ_{xy}).

Thus the formula for correlation coefficient is

$$r = \frac{\sigma_{xy}}{\sigma_x \times \sigma_y} \qquad \ldots (8.5)$$

Incorporating the formulae for $\bar{x}, \bar{y}$, σ_x, σ_y it becomes

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \qquad \ldots (8.6)$$

or alternatively

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{\left[n\sum_{i=1}^{n} X^2{}_i - \left(\sum_{i=1}^{n} X_i\right)^2\right]\left[n\sum_{i=1}^{n} Y^2{}_i - \left(\sum_{i=1}^{n} Y_i\right)^2\right]}} \qquad \ldots (8.7)$$

Let us go back to the data given in Table 8.1 and work out the value of r. You can use any of the formulae (8.4), (8.5), (8.6) or (8.7) to get the value of r. Since

all the above formulae are derived from the same concept we obtain the same value for r whichever formulae we use. For the data set in Table 8.1 we have calculated it by using (8.4) and (8.7). We construct Table 8.2 for this purpose.

Table 8.2: Calculation of Correlation Coefficient

Observation No.	X	Y	X^2	Y^2	XY
1	82	64	6724	4096	5248
2	70	40	4900	1600	2800
3	34	35	1156	1225	1190
4	80	48	6400	2304	3840
5	66	54	4356	2916	3564
6	84	56	7056	3136	4704
7	74	62	5476	3844	4588
8	84	66	7056	4356	5544
9	60	52	3600	2704	3120
10	86	82	7396	6724	7052
11	76	58	5776	3364	4408
12	76	66	5776	4356	5016
13	92	72	8464	5184	6624
14	72	46	5184	2116	3312
15	64	44	4096	1936	2816
16	86	76	7396	5776	6536
17	84	52	7056	2704	4368
18	60	40	3600	1600	2400
19	82	60	6724	3600	4920
20	90	60	8100	3600	5400
Total	**1502**	**1133**	**116292**	**67141**	**87450**

From Table 8.2 we note that

$$\sum_{i=1}^{20} X_i = 1502; \bar{X} = 75.1;$$

$$\sum_{i=1}^{20} Y_i = 1133; \bar{Y} = 56.65;$$

$$\sum_{i=1}^{20} X_i^2 = 116292; \sigma_x^2 = \frac{1}{20}\left[116292 - \frac{1502^2}{20}\right] = 174.59; \sigma_x = 13.21;$$

$$\sum_{i=1}^{20} Y_i^2 = 67141; \sigma_y^2 = \frac{1}{20}\left[67141 - \frac{1133^2}{20}\right] = 147.83; \sigma_y = 12.16;$$

$$\sum X_i Y_i = 87450; \sigma_{xy} = \frac{1}{20}\left[87450 - \frac{1502 \times 1133}{20}\right] = 118.09$$

Thus using formula 8.4, we have

$$r = \frac{118.09}{13.21 \times 12.16} = 0.735$$

Now let us use the formula 8.7. We have

$$r = \frac{20 \times 87450 - 1502 \times 1133}{\sqrt{(20 \times 116292 - 1502^2)(20 \times 67141 - 1133^2)}} = 0.735$$

Thus we see that both the formulae provide the same value of the correlation coefficient r. You can check yourself that the same value of r is obtained by using the formula (8.5). For this purpose you will need values on

$\sum (X_i - \bar{X})^2$, $\sum (Y_i - \bar{Y})^2$ and $\sum (X_i - \bar{X})(Y_i - \bar{Y})$

Hence you can have five columns on

$(X_i - \bar{X})$, $(Y_i - \bar{Y})$, $(X_i - \bar{X})^2$, $(Y_i - \bar{Y})^2$ and $(X_i - \bar{X})(Y_i - \bar{Y})$ in a table and find the totals.

8.5 INTERPRETATION OF CORRELATION COEFFICIENT

It is a mathematical fact that the value of r as defined above lies between –1 and + 1. The extreme values of –1 and + 1 are obtained only in situations where there is a *perfect linear relationship* between X and Y. The value –1 is obtained when this relationship is perfectly negative (i.e., inverse) and + 1 when this is perfect positive (i.e., direct). The value of 0 is obtained when there is no linear relationship between x and y.

We can make some guess work about the sign and degree of the correlation coefficient from the scatter diagram. Fig. 8.3 gives example of scatter diagrams for various values of r. Fig. 8.3(a) is a scatter diagram for the case $r = 0$; here there is no *linear relationship* between x and y. Fig. 8.3(b) is also an example of scatter diagram for the case $r = 0$; here there is discernible relationship between X and Y but it is not of the linear type. Here, initially, Y increases with X but later Y decreases as X increases resulting in a definitive quadratic relationship. But the correlation coefficient in this case is zero. Thus the correlation coefficient is only a measure of linear relationship. This sort of scatter diagram is obtained, if we plot, for instance, body weight (Y) of individuals against their age (X). Fig. 8.3(c) is an example of a scatter diagram where there is a perfect positive linear relationship between X and Y. We get this sort of scatter diagram if we plot, for instance, height of individuals in inches (X) against their heights in centimeters (Y); in that case $Y = 2.54X$, which is a deterministic and perfect linear relationship. Figures 8.3(d) to 8.3(k) are scatter diagrams for other values of r. From these scatter diagrams we get an idea of the nature of relationship and associated values of r.

From these it would seem that a value of 0.81 indicates a fair degree of linear relationship between scores in Statistics and Economics of these candidates. Such a quantification of relationship or association between variables is helpful for natural and social scientists to understand the phenomena they are investigating and explore these phenomena further. In an example of this sort, an educational psychologist may compute correlation coefficients between scores in various subjects and by further statistical analysis of the correlation coefficients and using psychological techniques may be able to form a theory as to what mental and other faculties are involved in making students good in various disciplines.

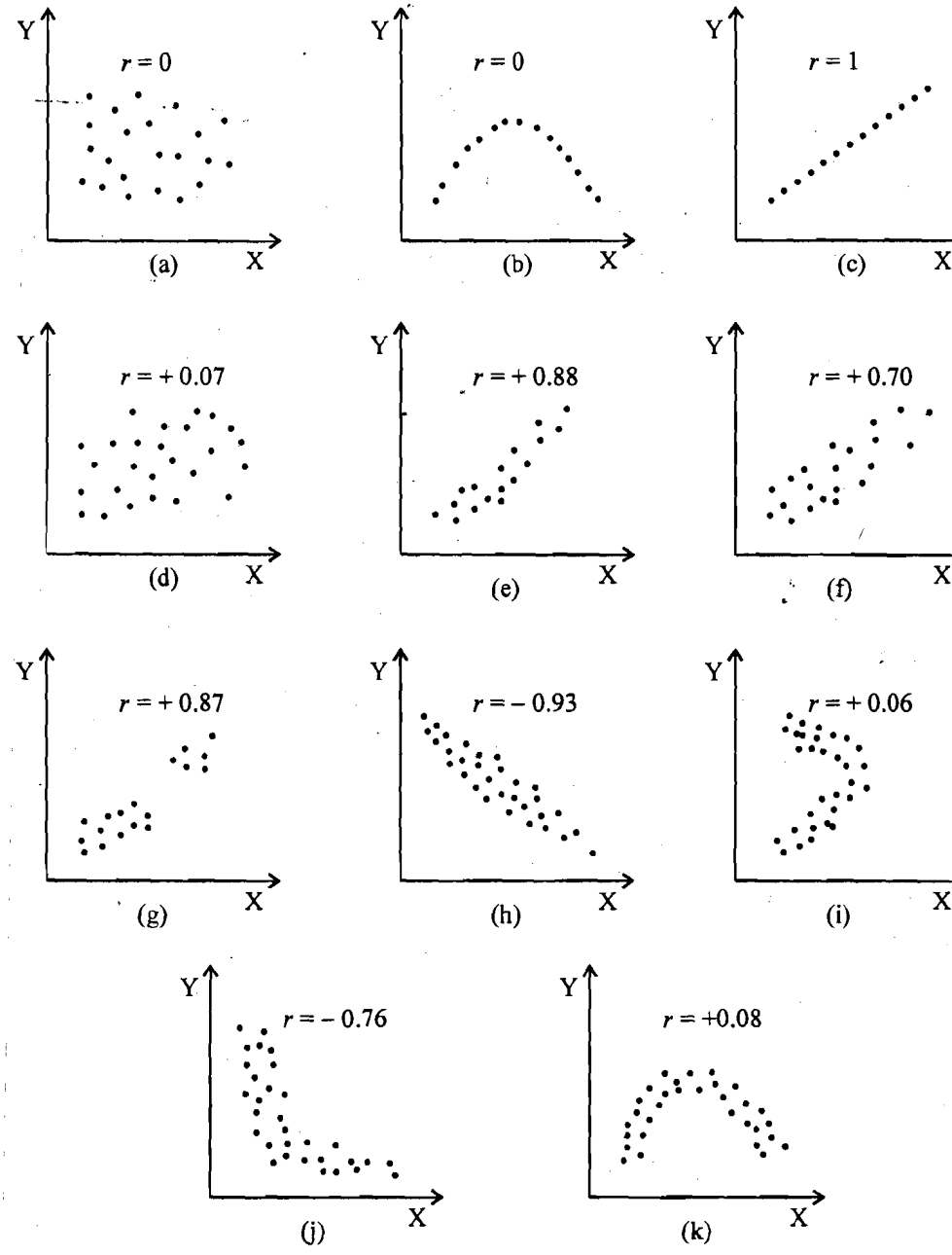


Fig. 8.3: Scatter Plots for Various Values of Correlation Coefficient

Remember that

- Correlation coefficient shows the linear relationship between X and Y. Thus, even if there is a strong non-linear relationship between X and Y, correlation coefficient may be low.
- Correlation coefficient is independent of scale and origin. If we subtract some constant from one (or both) of the variables, correlation coefficient will remain unchanged. Similarly, if we divide one (or both) of the variables by some constant, correlation coefficient will not change.
- Correlation coefficient varies between -1 and $+1$. This means r cannot be smaller than -1 and cannot be greater than $+1$.

The existence of a linear relationship between two variables is not to be interpreted to mean a cause-effect relationship between the two. For instance, if you work out the correlation between family expenditures on petrol and chocolates, you may find it to be fairly high indicating a fair degree of linear relationship. However, this

Both are luxury items and richer families can afford them and poorer ones cannot. Thus the high correlation here is caused by the high correlation of each of the variables with family income. To consider another example, suppose for each of the last twenty years, you work out the average height of an Indian and the average time per week an Indian watches television; you are likely to find a positive correlation. This does not, however, imply that watching television increases one's height or that taller people tend to watch television longer. Both these variables have an increasing trend over time and this is reflected in the high correlation. This kind of correlation between two variables is caused by the effect of a third variable on each of them rather than a direct linear cause-effect relationship between them is called *spurious correlation*.

Another aspect of the computation of correlation coefficient that we should be aware of is that the correlation coefficient like any other quantity computed from sample, varies from sample to sample and these sample fluctuations should be taken into account in making use of the computed coefficient. We do not discuss these techniques here.

Whether the presence of a linear relationship between two variables and hence a high correlation between them is genuine or spurious, such a situation is helpful to *predict* one variable from the other. We examine these prediction techniques in Unit 9.

Check Your Progress 1

1) Calculate r from the following given results :

   $n = 10;\ \sum X = 125, \sum X^2 = 1585, \sum Y = 80, \sum Y^2 = 650, \sum XY = 1007.$

   .....

   .....

   .....

   .....

   .....

2) Calculate the coefficient of correlation for the ages of husband and wife :

Age of husband	:	23	27	28	29	30	31	33	35	36	39
Age of wife	:	18	22	23	24	25	26	28	29	30	32

   .....

   .....

   .....

   .....

   .....

3) Specimens of similarly treated alloy steel containing various percentages of nickel are tested for toughness with the following results :

   Toughness (arbitrary units):

   47 50 52 52 54 56 58 59 60 60 62 64 65 66

   Percentage of Nickel :

   2.7 2.7 2.8 2.8 2.9 3.2 3.2 3.3 3.4 3.5 3.6 3.7 3.7 3.8

   Find the correlation coefficient between toughness and nickel content and comment on the result.

.....

.....

.....

.....

.....

.....

4) Determine the correlation coefficient between x and y —

x :	5	7	9	11	13	15
y :	1.7	2.4	2.8	3.4	3.7	4.4

.....

.....

.....

.....

.....

.....

5) The following table gives the saving bank deposits in billions of dollars and strikes and lock-outs, in thousands, over a number of years. Compute the correlation coefficient and comment on the result.

Saving deposits :	5.1	5.4	5.5	5.9	6.4	6.0	7.2
Strikes and lock-outs :	3.8	4.4	3.3	3.6	3.3	2.3	1.0

.....

.....

.....

.....

.....

.....

8.6 RANK CORRELATION COEFFICIENT

The Pearson's product moment correlation coefficient (or simply, the correlation coefficient) described above is suitable if both the variables involved are measurable (numerical) and the relationship between the variables is linear. However, there are situations where variables are not numerical but various items can be ranked according to the characteristics (i.e., ordinal). Sometimes even when the original variables are measurable, they are converted into ranks and a measure of association is computed. Consider for instance the situation when two examiners are asked to judge ten candidates on the basis of an oral examination. In this case, it may be difficult to assign scores to candidates, but the examiners find it reasonably easy to rank the candidates in order of merit. Before using the results, it may be advisable to find out if rankings are in reasonable concordance. For this, a measure of association between the ranks assigned by the two examiners may be computed. The Karl Pearson's correlation coefficient is not suitable in this situation. One may use the following measure called *Spearman's Rank Correlation Coefficient* for this purpose.

Table 8.3: Ranks of 10 Candidates by two Examiners

S.No.	Rank given by		Difference	
	Examiner I	Examiner II	D_i	D_i^2
1	6.0	6.5	- 0.5	0.25
2	2.0	3.0	- 1.0	1.00
3	8.5	6.5	2.0	4.00
4	1.0	1.0	0.0	0.00
5	10.0	2.0	8.0	64.00
6	3.0	4.0	- 1.0	1.00
7	8.5	9.5	- 1.0	1.00
8	4.0	5.0	- 1.0	1.00
9	5.0	8.0	- 3.0	9.00
10	7.0	9.5	- 2.5	6.25
			$\sum D_i = 0$	$\sum D_1^2 = 87.50$

Let us consider the data of Table 8.3. Here there are some ties; the tied cases are given the same rank in such a way that their total is the same as when there are no tie. For example, when there are two cases with rank 6, each is given a rank of 6.5 and there is no case with rank either 6 or 7. Similarly, if there are three cases with rank 5, then each is given a rank of 6 and there is no case with rank 5 or 7. Spearman's rank correlation coefficient, called Spearman's Rho, denoted by ρ, is based on the difference D_i (i for ith observation) between the two rankings. If the two rankings completely coincide, then D_i is zero for every case. The larger the value of D_i, the greater is the difference between the two rankings and smaller is the association. Thus the association can be measured by considering the magnitudes of D_i. Since the sum of D_i is always zero, to find a single index on the basis of D_i values, we should remove the sign of D_i and consider only the magnitude. In Spearman's ρ, this is done by taking D_i^2.

However, the largeness or smallness of $\sum_{i=1}^{n} D_i^2$, where n is the number of cases, will depend on n. Thus, in order to be able to interpret this value, we could create a ratio by dividing this sum by the largest possible value, which depends only on n, which is $\frac{n(n^2-1)}{6}$. However, $\frac{6\times\sum_{i=1}^{n} D_i^2}{n(n^2-1)}$ is zero for perfect association and 2 for lack of association, i. e., perfect negative association, while we would like it to be other way around. So we subtract this ratio from 1. Thus

$$\rho = 1 - \frac{6\times\sum_{i=1}^{n} D_i^2}{n(n^2-1)} \qquad \text{..... (8.8)}$$

is defined as Spearman's rank correlation.

Let us calculate the value of ρ from the data given in Table 8.3.

$$\rho = 1 - \frac{6\times 87.5}{10(10^2-1)} = 1 - \frac{525}{990} = 1 - 0.53 = 0.47.$$

Like Karl Pearson's coefficient of correlation the Spearman's rank correlation has a value + 1 for perfect matching of ranks, –1 for perfect mismatching of ranks and 0 for the lack of relation between the ranks.

There are other measures of association suitable for use when the variables are of nominal, ordinal and other types. We do not discuss them here.

Check Your Progress 2

1) In a contest, two judges ranked eight candidates A, B, C, D, E, F, G and H in order of their preference, as shown in the following table. Find the rank correlation coefficient.

	A	B	C	D	E	F	G	H
First Judge	5	2	8	1	4	6	3	7
Second Judge	4	5	7	3	2	8	1	6

.....

.....

.....

.....

.....

.....

2) Compute the correlation coefficient of the following ranks of a group of students in two examinations. What conclusion do you draw from the result?

Roll Nos.	1	2	3	4	5	6	7	8	9	10
Rank in B.Com. Exam.	1	5	8	6	7	4	2	3	9	10
Rank in M. Com Exam.	2	1	5	7	6	3	4	8	10	9

.....

.....

.....

.....

.....

.....

3) Ten competitors in a musical contest were ranked by 3 judges A, B and C in the following order :

Ranks by A :	1	6	5	10	3	2	4	9	7	8
Ranks by B :	3	5	8	4	7	10	2	1	6	9
Ranks by C :	6	4	9	8	1	2	3	10	5	7

Using Rank Correlation method, discuss which pair of judges has the nearest approach to common liking in music.

.....

.....

.....

.....

.....

.....

4) Ten students obtained the following marks in Mathematics and Statistics. Calculate the rank correlation coefficient.

Student (Roll No.)	1	2	3	4	5	6	7	8	9	10
Marks in Mathematics	78	36	98	25	75	82	90	62	65	39
Marks in Statistics	84	51	91	60	68	62	86	58	53	47

.....

.....

.....

.....

.....

8.7 LET US SUM UP

In this unit you have learnt about scatter diagram and covariance. Also you learnt about the coefficient of correlation and the coefficient of rank correlation that will indicate the closeness of the linear association or correlation between two variables. However, correlation does not imply a cause-effect relationship.

8.8 KEY WORDS

Correlation Analysis : Refers to a measure of association between two random variables. If two random variables have been such that when one gets changed the other will do so in a related manner, they are regarded to be correlated. Variables which are independent are not correlated. The correlation coefficient is a number between –1 and + 1. It could be calculated from a number of pairs of observations which are normally referred to as points (X, Y). A coefficient of 1 implies perfect positive correlation, –1 perfect negative correlation and 0 no correlation.

Covariance : The first product moment of two variables about their means is called covariance. The formula for the calculation of covariance is $\frac{1}{n}\sum(X_i - \bar{X})(Y_i - \bar{Y})$ or $\frac{1}{n}\left(\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}\right)$ where X and Y are corresponding values of each variable and n is the number of observations.

Rank Correlation Coefficient : There happen to be many occasions when it may not be convenient, economic or even possible to give values to variables. However, various items can be ranked. In such cases, a rank correlation coefficient may be used.

Scatter Diagram : A diagram showing the joint variation of two variables X and Y. Each member is represented by a point whose coordinates, on ordinary rectangular axes, are the values of the variables. A set of n observations thus provides n points on the diagram and the scatter or clustering of the points exhibits the relationship between X and Y.

8.9 SOME USEFUL BOOKS

Nagar, A.L. and R.K. Das, 1989 : *Basic Statistics*, Oxford University Press, Delhi.

Goon, A.M., M.K. Gupta and B. Dasgupta, 1987 : *Basic Statistics*, The World Press Pvt. Ltd., Calcutta.

8.10 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) + 0.47
2) + 0.996
3) + 0.98
4) + 0.995
5) – 0.84

Check Your Progress 2

1) 2/3
2) + 0.64
3) – 0.21, + 0.64, – 0.30
4) + 0.82

UNIT 9 REGRESSION ANALYSIS

Structure

9.0 OBJECTIVES

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

- explain the concept of regression;
- explain the method of least squares;
- identify the limitations of linear regression;
- apply linear regression models to given data; and
- use the regression equation for prediction.

9.1 INTRODUCTION

In the previous Unit we noted that correlation coefficient does not reflect cause and effect relationship between two variables. Thus we cannot predict the value of one variable for a given value of the other variable. This limitation is removed by regression analysis. In regression analysis, to be discussed in this Unit, the relationship between variables are expressed in the form of a mathematical equation. It is assumed that one variable is the cause and the other is the effect. You should remember that regression is a statistical tool which helps understand the relationship between variables and predicts the unknown values of the dependent variable from known values of the independent variable.

9.2 THE CONCEPT OF REGRESSION

In regression analysis we have two types of variables: i) dependent (or explained) variable, and ii) independent (or explanatory) variable. As the name (explained and explanatory) suggests the dependent variable is explained by the independent variable.

In the simplest case of regression analysis there is one dependent variable and one independent variable. Let us assume that consumption expenditure of a household is related to the household income. For example, it can be postulated that as household income increases, expenditure also increases. Here consumption expenditure is the dependent variable and household income is the independent variable.

Usually we denote the dependent variable as Y and the independent variable as X. Suppose we took up a household survey and collected n pairs of observations in X and Y. The next step is to find out the nature of relationship between X and Y.

The relationship between X and Y can take many forms. The general practice is to express the relationship in terms of some mathematical equation. The simplest of these equations is the linear equation. This means that the relationship between X and Y is in the form of a straight line and is termed linear regression. When the equation represents curves (not a straight line) the regression is called non-linear or curvilinear.

Now the question arises, 'How do we identify the equation form?' There is no hard and fast rule as such. The form of the equation depends upon the reasoning and assumptions made by us. However, we may plot the X and Y variables on a graph paper to prepare a scatter diagram. From the scatter diagram, the location of the points on the graph paper helps in identifying the type of equation to be fitted. If the points are more or less in a straight line, then linear equation is assumed. On the other hand, if the points are not in a straight line and are in the form of a curve, a suitable non-linear equation (which resembles the scatter) is assumed.

We have to take another decision, that is, the identification of dependent and independent variables. This again depends on the logic put forth and purpose of analysis: whether 'Y depends on X' or 'X depends on Y'. Thus there can be two regression equations from the same set of data. These are i) Y is assumed to be dependent on X (this is termed 'Y on X' line), and ii) X is assumed to be dependent on Y (this is termed 'X on Y' line).

Regression analysis can be extended to cases where one dependent variable is explained by a number of independent variables Such a case is termed multiple regression. In advanced regression models there can be a number of both dependent as well as independent variables.

You may by now be wondering why the term 'regression', which means 'reduce'. This name is associated with a phenomenon that was observed in a study on the relationship between the stature of father (x) and son (y). It was observed that the average stature of sons of the tallest fathers has a tendency to be less than the average stature of these fathers. On the other hand, the average stature of sons of the shortest fathers has a tendency to be more than the average stature of these fathers. This phenomenon was called *regression towards the mean*. Although this appeared somewhat strange at that time, it was found later that this is due to natural variation within subgroups of a group and the same phenomenon occurred in most problems and data sets. The explanation is that many tall men come from families with average stature due to vagaries of natural variation and they produce sons who are shorter than them on the whole. A similar phenomenon takes place at the lower end of the scale.

9.3 LINEAR RELATIONSHIP: TWO VARIABLE CASE

The simplest relationship between X and Y could perhaps be a linear *deterministic* function given by

$$Y_i = a + bX_i \qquad \ldots(9.1)$$

In the above equation X is the independent variable or explanatory variable and Y is the dependent variable or explained variable. You may recall that the subscript i represents the observation number, i ranges from 1 to n. Thus Y_1 is the first observation of the dependent variable, X_5 is the fifth observation of the independent variable, and so on.

Equation (9.1) implies that Y is completely determined by X and the parameters a and b. Suppose we have parameter values $a = 3$ and $b = 0.75$, then our linear equation is Y = 3 + 0.75 X. From this equation we can find out the value of Y for given values of X. For example, when X = 8, we find that Y = 9. Thus if we have different values of X then we obtain corresponding Y values on the basis of (9.1). Again, if X_i is the same for two observations, then the value of Y_i will also be identical for both the observations. A plot of Y on X will show no deviation from the straight line with intercept 'a' and slope 'b'.

If we look into the deterministic model given by (9.1) we find that it may not be appropriate for describing economic interrelationship between variables. For example, let Y = consumption and X = income of households. Suppose you record your income and consumption for successive months. For the months when your income is the same, do your consumption remain the same? The point we are trying to make is that economic relationship involves certain randomness.

Therefore, we assume the relationship between Y and X to be *stochastic* and add one error term in (9.1). Thus our stochastic model is

$$Y_i = a + bX_i + e_i \qquad \ldots(9.2)$$

where e_i is the error term. In real life situations e_i represents randomness in human behaviour and excluded variables, if any, in the model. Remember that the right hand side of (9.2) has two parts, viz., i) deterministic part (that is, $a + bX_i$), and ii) stochastic or randomness part (that is, e_i). Equation (9.2) implies that even if X_i remains the same for two observations, Y_i need not be the same because of different e_i. Thus, if we plot (9.2) on a graph paper the observations will not remain on a straight line.

Example 9.1

The amount of rainfall and agricultural production for ten years are given in Table 9.1.

Table 9.1: Rainfall and Agricultural Production

Rainfall (in mm.)	*Agricultural production (in tonne)*
60	33
62	37
65	38
71	42
73	42
75	45
81	49
85	52
88	55
90	57

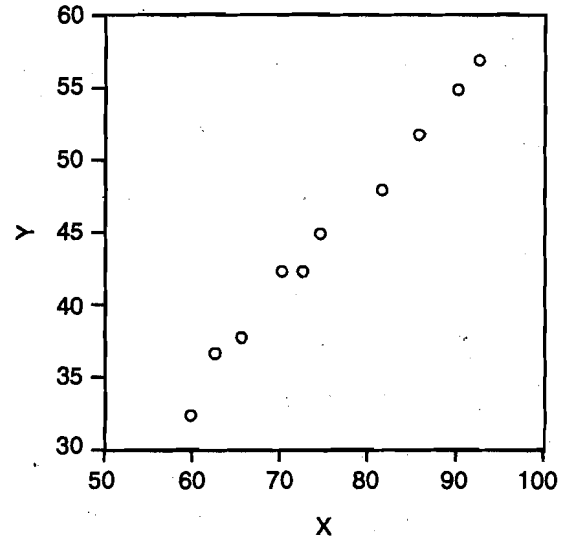


Fig. 9.1: Scatter Diagram

We plot the data on a graph paper. The scatter diagram looks something like Fig. 9.1. We observe from Fig. 9.1 that the points do not lie strictly on a straight line. But they show an upward rising tendency where a straight line can be fitted. Let us draw the regression line along with the scatter plot.

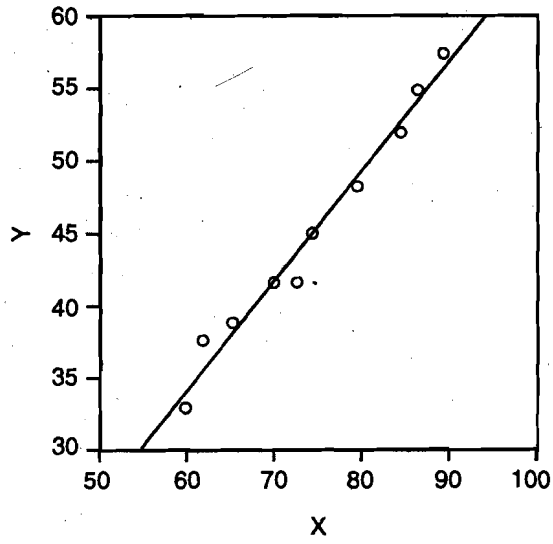


Fig. 9.2: Regression Line

The vertical difference between the regression line and the observations is the error e_i. The value corresponding to the regression line is called the predicted value or the expected value. On the other hand, the actual value of the dependent variable corresponding to a particular value of the independent variable is called the observed value. Thus 'error' is the difference between predicted value and observed value.

A question that arises is, 'How do we obtain the regression line? The procedure of fitting a straight line to the data is explained below.

9.4 MINIMISATION OF ERRORS

As mentioned earlier, a straight line can be represented by

$$Y_i = a + bX_i$$

where b is the *slope* and a is the *intercept* on y-axis. The location of a straight line depends on the value of a and b, called *parameters*. Therefore, the task before us is to *estimate* these parameters from the collected data. (You will learn more about the concept of estimation in Block 7). In order to obtain the line of best fit to the data we should find estimates of a and b in such a way that the error e_i is minimum.

In Fig. 9.1 these differences between observed and predicted values of Y are marked with straight lines from the observed points, parallel to y-axis, meeting the regression line. The lengths of these segments are the errors at the observed points.

Let us denote the n observations as before by (X_i, Y_i), $i = 1, 2, ..., n$. In Example 9.1 on agricultural production and rainfall, n=10. Let us denote the predicted value of Y_i at X_i by $\hat{Y}_i$ (the notation $\hat{Y}_i$ is pronounced as 'Y_i-cap' or 'Y_i-hat'). Thus

$\hat{Y}_i = a + bX_i$, $i = 1, 2, ..., n$.

The error at the i^{th} point will then be

$$e_i = Y_i - \hat{Y}_i \qquad \ldots (9.3)$$

It would be nice if we can determine a and b in such a way that each of the e_i, $i = 1, 2, ..., n$ is zero. But this is impossible unless it so happens that all the n points lie on a straight line, which is very unlikely. Thus we have to be content with minimising a combination of e_i, $i = 1, 2, ..., n$. What are the options before us?

- It is tempting to think that the total of all the e_i, $i = 1, 2, ..., n$, that is, $\sum_{i=1}^{n} e_i$ is a suitable choice. But it is not. Because, for points above the line are positive and below the line are negative. Thus by having a combination of large positive and large negative errors, it is possible for $\sum_{i=1}^{n} e_i$ to be very small.
- A second possibility is that if we take $a = \bar{y}$ (the arithmetic mean of the Y_i's) and $b = 0$, $\sum_{i=1}^{n} e_i$ could be made zero. In this case, however, we do not need

the value of X at all for prediction! The predicted value is the same irrespective of the observed value of X. This evidently is wrong.

- What then is wrong with the criterion $\sum_{i=1}^{n} e_i$? It takes into account the sign of e_i. What matters is the magnitude of the error and whether the error is on the positive side or negative side is really immaterial. Thus, the criterion $\sum_{i=1}^{n} |e_i|$ is a suitable criterion to minimise. Remember that $|e_i|$ means the absolute value of e_i. Thus, if $e_i = 5$ then $|e_i| = 5$ and also if $e_i = -5$ then $|e_i| = 5$. However, this option poses some computational problems.
- For theoretical and computational reasons, the criterion of *least squares* is preferred to the absolute value criterion. While in the absolute value criterion the sign of e_i is removed by taking its absolute value, in the *least squares criterion* it is done by squaring it. Remember that the squares of both 5 and –5 are 25. This device has been found to be mathematically and computationally more attractive.

We explain in detail the least squares method in the following Section.

9.5 METHOD OF LEAST SQUARES

In the least squares method we minimise the sum of squares of the error terms, that is, $\sum_{i=1}^{n} e_i^2$.

From (9.3) we find that $e_i = Y_i - \hat{Y}_i$

which implies $e_i = Y_i - (a + bX_i) = Y_i - a - bX_i$.

Hence, $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - a - bX_i)^2$ ...(9.4)

The next question is: How do we obtain the values of a and b to minimise (9.3)?

- Those of you who are familiar with the concept of differentiation will remember that the value of a function is minimum when the first derivative of the function is zero and second derivative is positive. Here we have to choose the value of a and b. Hence, $\sum_{i=1}^{n} e_i^2$ will be minimum when its partial derivatives with respect to a and b are zero. The partial derivatives of $\sum_{i=1}^{n} e_i^2$ are obtained as follows:

$$\frac{\partial \sum_i e_i^2}{\partial a} = \frac{\partial \sum_i (Y_i - a - bX_i)^2}{\partial a} = 2 \cdot (-1) \cdot \sum_i (Y_i - a - bX_i) \quad \ldots(9.5)$$

$$\frac{\partial \sum_i e_i^2}{\partial b} = \frac{\partial \sum_i (Y_i - a - bX_i)^2}{\partial b} = 2 \cdot (-X_i) \cdot \sum_i (Y_i - a - bX_i) \quad \ldots(9.6)$$

By equating (9.5) and (9.6) to zero and re-arranging the terms we get the following two equations:

$$\sum_{i=1}^{n} Y_i = na + b\sum_{i=1}^{n} X_i \quad \ldots(9.7)$$

$$\sum_{i=1}^{n} X_i Y_i = a\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^2 \quad \ldots(9.8)$$

These two equations, (9.7) and (9.8), are called the *normal equations* of least squares. These are two simultaneous linear equations in two unknowns. These can be solved to obtain the values of a and b.

Those of you who are not familiar with the concept of differentiation can use a rule of thumb (We suggest that you should learn the concept of differentiation, which is so much useful in Economics). We can say that the normal equations given at (9.7) and (9.8) are derived by multiplying the coefficients of a and b to the linear equation and summing over all observations. Here the linear equation is $Y_i = a + bX_i$. The first normal equation is simply the linear equation $Y_i = a + bX_i$ summed over all observations (since the coefficient of a is 1).

$\sum Y_i = \sum a + \sum bX_i$ or $\sum Y_i = na + b\sum X_i$

The second normal equation is the linear equation multiplied by X_i (since the coefficient of b is X_i)

$\sum X_i Y_i = \sum aX_i + \sum bX_i^2$ or $\sum X_i Y_i = a\sum X_i + b\sum X_i^2$

After obtaining the normal equations we calculate the values of a and b from the set of data we have.

Example 9.2: Assume that quantity of agricultural production depends on the amount of rainfall and fit a linear regression to the data given in Example 9.1.

In this case dependent variable (Y) is quantity of agricultural production and independent variable (X) is amount of rainfall. The regression equation to be fitted is $Y_i = a + bX_i + e_i$

For the above equation we find out the normal equations by the method of least squares. These equations are given at (9.7) and (9.8). Next we construct a table as follows:

Table 9.2: Computation of Regression Line

X_i	Y_i	X_i^2	X_iY_i	$\hat{Y}_i$	e_i
60	33	3600	1980	33.85	-0.85
62	37	3844	2294	35.34	1.66
65	38	4225	2470	37.57	0.43
71	42	5041	2982	42.03	-0.03
73	42	5329	3066	43.51	-1.51
75	45	5625	3375	45.00	0.00
81	49	6561	3969	49.46	-0.46
85	52	7225	4420	52.43	-0.43
88	55	7744	4840	54.66	0.34
90	57	8100	5130	56.15	0.85
$\sum_i X_i = 750$	$\sum_i Y_i = 450$	$\sum_i X_i^2 = 57294$	$\sum_i X_iY_i = 34526$	$\sum_i \hat{Y}_i = 450$	$\sum_i e_i = 0$

By substituting values from Table 9.2 in the normal equations (9.7) and (9.8) we get the following:

$$\begin{aligned} 450 &= 10a + 750b \\ 34526 &= 750a + 57294b \end{aligned}$$

By solving these two equations we obtain $a = -10.73$ and $b = 0.743$.

So the regression line is $\hat{Y}_i = -10.73 + 0.743X_i$.

Notice that the sum of errors $\sum_i e_i$ for the estimated regression equation in zero (see the last column of Table 9.2).

The computation given in Table 9.2 often involves large numbers and poses difficulty. Hence we have a short-cut method for calculating the values of a and b from the normal equations.

Let us take

$x = X - \bar{X}$ and $y = Y - \bar{Y}$ where $\bar{X}$ and $\bar{Y}$ are the arithmetic means of X and Y respectively.

Hence $xy = (X - \bar{X})(Y - \bar{Y})$

By re-arranging terms in the normal equations we find that

$$b = \frac{\sum_{i=1}^{n} xy}{\sum_{i=1}^{n} x^2} \qquad \ldots(9.9)$$

$$a = \bar{Y} - b\bar{X} \qquad \ldots(9.10)$$

You may recall from Unit 8 that *covariance* is given by $\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$

$= \frac{1}{n}\sum_{i=1}^{n} x_i y_i$. Moreover, variance of X is given by $\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$

Since $b = \frac{\sum_{i=1}^{n} xy}{\sum_{i=1}^{n} x^2}$ we can say that $b = \frac{\sigma_{xy}}{\sigma_x^2}$ $\qquad \ldots(9.11)$

Since these formulae are derived from the normal equations we get the same values for a and b in this method also. For the data given in Table 9.1 we compute the values of a and b by this method. For this purpose we construct Table 9.3.

Table 9.3: Computation of Regression Line (short-cut method)

	X_i	Y_i	x_i	y_i	x_i^2	$x_i\ y_i$
	60	33	-15	-12	225	180
	62	37	-13	-8	169	104
	65	38	-10	-7	100	70
	71	42	-4	-3	16	12
	73	42	-2	-3	4	6
	75	45	0	0	0	0
	81	49	6	4	36	24
	85	52	10	7	100	70
	88	55	13	10	169	130
	90	57	15	12	225	180
Total	**750**	**450**	**0**	**0**	**1044**	**776**

On the basis of Table 9.3 we find that

$$\bar{X} = \frac{750}{10} = 75 \text{ and } \bar{Y} = \frac{450}{10} = 45$$

$$b = \frac{\sum_{i=1}^{n} xy}{\sum_{i=1}^{n} x^2} = \frac{776}{1044} = 0.743$$

$$a = \bar{Y} - b\bar{X} = 45 - 0.743 \times 10 = -10.73$$

Thus the regression line in this method also $\hat{Y}_i = -10.73 + 0.743X_i$ ...(9.12)

Coefficient b in (9.12) is called the regression coefficient. This coefficient reflects the amount of increase in Y when there is a unit increase in X. In regression equation (9.12) the coefficient $b = 0.743$ implies that if rainfall increase by 1 mm. agricultural production will increase 0.743 thousand tonne.

Regression coefficient is widely used. It is also an important tool of analysis. For example, if Y is aggregate consumption and X is aggregate income, b represents marginal propensity to consume (MPC).

9.6 PREDICTION

A major interest in studying regression lies in its ability to forecast. In Example 9.1 in the previous Section we assumed that the quantity of agricultural production is dependent on the amount of rainfall. We fitted a linear equation to the observed data and got the relationship

$$\hat{Y}_i = -10.73 + 0.743X_i$$

From this equation we can predict the quantity of agricultural output given the amount of rainfall. Thus when rainfall is 60 mm. agricultural production is $(-10.73 + 0.74 \times 60) = 33.85$ thousand tonnes. This figure is the *predicted value* on the basis of regression equation. In a similar manner we can find the predicted values of Y for different values of X.

Compare the predicted value with the observed value. From Table 9.1 where observed values are given we find that when rainfall is 60 mm. agricultural production is 33 thousand tonnes. In fact, the predicted values $\hat{Y}_i$ for observed values of X are given in the fifth column of Table 9.2. Thus when rainfall is 60 mm. predicted value is 33.85 thousand tonnes. Thus the error value is –0.85 thousand tonne.

Now a question arises, 'Which one, between observed and predicted values, should we believe?' In other words, what will be the quantity of agricultural production if there is a rainfall of 60 mm. in future? On the basis of our regression line it is given to be 33.85 tonnes. And we accept this value because it is based on the overall data. The error of –0.85 is considered as a random fluctuation which may not be repeated.

The second question that comes to our mind is, 'Is the prediction valid for any value of X?' For example, we find from the regression equation that when rainfall is zero, agricultural production is –10.73 thousand tonne. But common sense tells us that agricultural production cannot be negative! Is there anything wrong with our regression equation? In fact, the regression equation here is estimated on the basis of rainfall data in the range of 60-90 mm. Thus prediction is be valid in this range of X. Our prediction should not be for far off values of X.

A third, question that arises here is, 'Will the predicted value come true?' This depends upon the *coefficient of determination*. If the coefficient of determination is closer to one, there is greater likelihood that the prediction will be realised. However, the predicted value is constrained by elements of randomness involved with human behaviour and other unforeseen factors.

9.7 RELATIONSHIP BETWEEN REGRESSION AND CORRELATION

In regression analysis the status of the two variables (X, Y) are different such that Y is the variable to be predicted and X is the variable, information on which is to be used. In the rainfall-agricultural production problem, it makes sense to predict agricultural production on the basis of rainfall and it would not make sense to try and predict rainfall on the basis of agricultural production. However, in the case of scores in Economics and Statistics (see Example 8.1 in the previous Unit), either one could be X and the other Y. Hence we consider the two prediction problems: (i) predicting Economics score (Y) from Statistics score (X); and (ii) predicting Statistics score (X) from Economics score (Y).

Thus we can have two regression coefficients from a given set of data depending upon the choice of dependent and independent variables. These are:

a) Y on X line, $Y_i = a + bX_i$

b) X on Y line, $X_i = \alpha + \beta Y_i$

You may ask, 'What is the need for having two different lines? By rearrangement of terms of the Y on X line we obtain $X_i = -\frac{a}{b} + \frac{1}{b}Y_i$. Thus we should have $\alpha = -\frac{a}{b}$ and $\beta = \frac{1}{b}$. However, the observations are not on a straight line and

the relation between X and Y is not a mathematical one. You may recall that estimates of the parameters are obtained by the method of least squares. Thus the regression line $\hat{Y}_i = a + bX_i$ is obtained by minimising $\sum_i (Y_i - a - bX_i)^2$ whereas the regression line $\hat{X}_i = \alpha + \beta Y_i$ is obtained by minimising $\sum_i (X_i - \alpha - \beta Y_i)^2$.

However, there is a relationship between the two regression coefficients b and β.

We have noted earlier that $b = \frac{\sigma_{xy}}{\sigma_x^2}$. By a similar formula by interchanging the roles of X and Y we find $\beta = \frac{\sigma_{xy}}{\sigma_y^2}$. But by definition we notice that $\sigma_{xy} = \sigma_{yx}$.

Thus $b \times \beta = \frac{\sigma_{xy}^2}{\sigma_x^2 \times \sigma_y^2}$, which is the same as r^2.

This r^2 is called the *coefficient of determination*. Thus the product of the two regression coefficients of Y on X and X on Y is the square of the correlation coefficient. This gives a relationship between correlation and regression. Notice, however, that the coefficient of determination of either regression is the same, i.e., r^2; this means that although the two regression lines are different, their predictive powers are the same. Note that the coefficient of determination r^2 ranges between 0 and 1, i.e., the maximum value it can assume is unity and the minimum value is zero; it cannot be negative.

From the previous discussion, two points emerge clearly:

1) If the points in the scatter lie close to a straight line, then there is a strong relationship between X and Y and the correlation coefficient is high.
2) If the points in the scatter diagram lie close to a straight line, then the observed values and predicted values of Y by least squares are very close and the prediction errors $(Y_i - \hat{Y}_i)$ are small.

Thus, the prediction errors by least squares seem to be related to the correlation coefficient. We explain this relationship here. The sum of squares of errors at the various points upon using the least squares linear regression is $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$.

On the other hand, if we had not used the value of observed X to predict Y, then the prediction would be a constant, say, a. The best value of a by least squares criterion is such an a that minimises $\sum_{i=1}^{n} (Y_i - a)^2$; the solution to this a is seen to be $\bar{Y}$. Thus the sum of squares of errors of prediction at various points without using X is $\sum_{i=1}^{n} (Y_i - \bar{Y})^2$.

The ratio, $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \Big/ \sum_{i=1}^{n} (Y_i - \bar{Y})^2$ can then be used as an index of how much has been gained by the use of X. In fact, this ratio is the coefficient of determination and same as r^2 mentioned above. Since both the numerator and denominator of this ratio are non-negative, the ratio is greater than or equal to zero.

Check Your Progress 1

1) From the following data find the coefficient of linear correlation between X and Y. Determine also the regression line of Y on X, and then make an estimate of the value of Y when $X = 12$,

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

.....

.....

.....

.....

.....

.....

2) Obtain the lines of regression for the following data:

(X)	1	2	3	4	5	6	7	8	9
(Y)	9	8	10	12	11	13	14	16	15

.....

.....

.....

.....

.....

.....

3) Find the two lines of regression from the following data :

Age of Husband (X)	25	22	28	26	35	20	22	40	20	18
Age of Wife (Y)	18	15	20	17	22	14	16	21	15	14

Hence estimate (i) age of husband when the age of wife is 19, (ii) age of wife when the age of husband is 30.

.....

.....

.....

.....

.....

.....

4) From the following data, obtain the two regression equations :

Sales :	91	97	108	121	67	124	51	73	111	57
Purchases :	71	75	69	97	70	91	39	61	80	47

.....

.....

.....

.....

.....

.....

17.9 NON-PROBABILITY SAMPLING PROCEDURES

There are different types of non-probability sampling such as:

1) Convenience Sampling
2) Judgment Sampling
3) Quota Sampling
4) Snowball Sampling

We discuss the procedure of drawing a non-probability sampling below.

17.9.1 Convenience Sampling

This is one of the most commonly used methods of non-probability sampling. In this method the researcher's convenience forms the basis for selection of the sample. Especially for a exploratory research there is a pressing need for data. In such situations the selection of sampling units is left to the interviewer. The population units are included in the sample simply because they are in the right place at the right time. This method is often used during preliminary research efforts to get a gross estimate of the results, without incurring the cost or time required to select a sample. For example, during the budget session or when the price of a product is increased or a new government is formed, convenience samples are used by the researchers/journalists to reflect public opinion. Convenience samples are extensively used in marketing research.

The advantage of convenience sampling is that it is less expensive and less time-consuming. The limitations of convenience sampling are: (a) it involves sample selection bias, and (b) it does not provide a representative sample of the population and therefore we cannot generalise the results.

17.9.2 Judgment Sampling

This is another commonly used non-probability sampling procedure. This procedure is often referred to as *purposive sampling*. In this procedure the researcher selects the sample based on his/her judgment. The researcher believes that the selected sample elements are representative of the population. For example, the calculation of consumer price index is based on judgment sampling. Here the sample consists of a basket of consumer items and other goods and services which are expected to reflect a representative sample. The prices of these items are collected from selected cities that are viewed as typical cities with demographic profiles matching the national profile.

The advantage of judgment sampling is that it is low cost, convenient and quick. The disadvantage is that it does not allow direct generalisations to population. The quality of the sample depends upon the judgment of the researcher.

17.9.3 Quota Sampling

In this procedure the population is divided into groups based on some characteristics such as gender, age, education, religion, income group, etc. A quota of units from each group is determined. The quota may be either proportional or non-proportional. The proportional quota sampling is based on the proportion of each characteristic in the population so that the proportion in the sample represents the population proportion. For example, if you know that there are 80% of the households whose income is below say Rs.100000 per annum and 20% households

have income above Rs.100000 per annum in a city. You want to take a sample of size 100 households. Then you include 80 households from below Rs.100000 income and 20 households from above Rs.100000 income. The objective here is to meet the proportional quota of sampling from each characteristic in the population.

The non-proportional quota sampling is a bit less restrictive. In this procedure, you specify the minimum number of sampled units from each group. You are not concerned with having proportions in the population. For instance, in the above example you may simply interview 50 households from each income group instead of 80% and 20%. The interviewer is instructed to fill the quota for each group based on convenience or judgment. The very purpose of quota sampling is that various groups in the population are represented to the extent the investigator desires.

Do not confuse the quota sampling with stratified sampling that you have learned earlier. In stratified sampling you select random samples from each stratum or group whereas in quota sampling the interviewer has a fixed quota. For example, in a city there are five market centres. A company wants to assess the demand for its new product and sends 5 investigators to assess the demand by interviewing 50 prospective customers from each market. It is left to the investigator whom he/she will interview at each market centre. If the product is targeted to women, this way you cannot elicit the information among various groups of women customers like housewives or employed women or young or old. In this sampling you are simply fixing a quota for each investigator.

The quota sampling has the advantage over others if the sample meets the characteristics of the population that you are looking into. In addition, the cost and time involved in collecting the data are greatly reduced. However, there are many disadvantages as well. In quota sampling, the samples are selected according to the convenience of the investigator instead of selecting random samples. Therefore, the selected samples may be biased. If there are a large number of characteristics on the basis of which the quotas are fixed, then it becomes very difficult to fix the quotas/sub-quotas for each group/sub-group. Also the investigators have the tendency to collect information only from those who are willing to provide information and avoid unwilling respondents.

17.9.4 Snowball Sampling

In snowball sampling, we begin by identifying someone who meets the criteria for inclusion in our study. We then ask him/her to recommend others who also meets the criteria. Although this method would hardly lead to representative samples, there are times when it may be the best method available. Snowball sampling is especially useful when we are trying to reach populations that are inaccessible or hard to find. For example, if we are studying the homeless, we are not likely to find good lists of homeless people within a specific geographical area. However, if we go to that area and identify one or two, we may find that they know very well who the other homeless people in their vicinity are and how we can find them.

17.10 DETERMINING THE SAMPLE SIZE

The use of appropriate sampling procedure is necessary for a representative sample. However, this condition is not sufficient. In addition to the above, we should

determine the sample size. The question of how large a sample should be is a difficult one. Sample size can be determined by various considerations. The following are some of the considerations in determining the sample size:

a) Sampling error
b) Number of comparisons to be made
c) Response rates
d) Funds available

a) **Sampling Error:** In Unit 16 you have learned that smaller samples have greater sampling error than large samples. On the other hand, larger samples have larger non-sampling errors than smaller samples. The sampling error is a number that describes the precision of an estimate of the sample. It is usually expressed as a margin of error associated with a statistical level of confidence. For example, for a prime minister preferential poll you may say that the incumbent is favored by 65% of votes, with a margin of error (precision) of plus or minus 5 percentage points at a 95% confidence level. This means that if the same surveys were conducted with 100 different samples of voters, 95 of the surveys would be expected to show the incumbent favoured by between 60% and 70% of the voters (65% ± 5%). Remember as you increase the precision level of your results you need larger sample size.

b) **Number of Comparisons to Make:** Sometimes we may be interested in making comparisons of two or more groups (strata) in the sample. For example, we may want to make the comparison between male and female respondents or between urban and rural respondents. Or we may want to compare the results for 4 geographical regions of the country say north, south, west and east. Then we need an adequate sample size in each region or stratum of the population. Therefore, the heterogeneity of population characteristics plays a significant role in deciding the sample size.

c) **Response Rates:** In mail surveys, we know that all those questionnaires mailed to the respondents may not reach us back after filling the questionnaires. As per the experiences on mail survey, the response rate ranges between 10% to 50%. Then, if you are expecting a 20% response rate, for example, you will have to mail 5 times the number of sample size required.

d) **Funds Available:** The funds available may influence the sample size. If the funds available for the study are limited then you may not be able to spend more than a certain amount of the total money available with you on collecting the data.

It is even more difficult to decide the sample size, when you use the non-probability sampling procedures. This is because there are no definite rules to be followed in non-probability sampling procedures. It all depends upon on what you want to know, the purpose of inquiry, what will be useful, what will have credibility and what can be done with available time and resources. In purposive sampling, the sample should be judged on the basis of purpose. In non-probability sampling procedures, the validity, meaningfulness, and insights generated have more to do with the information-richness of the sample units selected rather than the sample size.

Some Formulae to Determine the Sample Size

Technical considerations suggest that the required sample size is a function of the precision of the estimates you wish to achieve, the variance of the population and

the confidence level you wish to use. If you want more precision and confidence level then you may need larger sample size. The more frequently used confidence levels are 95% and 99%. And the more frequently used precision levels are 1% and 5%. There are different formulae used to determine the sample size depending upon various considerations discussed above. In this section we will discuss three of them.

i) If we wish to report the results as percentages (proportions) of the sample responding, we use the following formula:

$$n_i = \frac{P_i(1-P_i)}{\frac{\alpha^2}{z^2} + \frac{P_i(1-P_i)}{N_i}}$$

Where, n_i = sample size of the i^{th} attribute required

P_i = estimated proportion of the population possessing i^{th} attribute of interest (for example, proportion of males, females, urban, rural, etc.)

α = precision required (0.01, 0.05 etc.)

z = standardized value indicating the confidence level (z=1.96 at 95% confidence level and z=2.58 at 99% confidence level)

N_i = population size of the i^{th} attribute (known or estimated)

Example 17.8: A population consists 80% rural and 20% urban people. Given that the population size is 50000, determine the sample size required. Assume that the desired precision and confidence levels are 1% and 99% respectively. In this example,

P_1 = proportion of rural people = 0.80

P_2 = proportion of urban people = 0.20

N_1 = rural population size = 50000 × 0.80 = 40000

N_2 = urban population size = 50000 × 0.20 = 10000

α = 0.01

z = 2.58 (at 99% confidence level)

The required sample size is

$$n_1 = \text{rural sample} = \frac{P_1(1-P_1)}{\frac{\alpha^2}{z^2} + \frac{P_1(1-P_1)}{N_1}}$$

$$= \frac{0.80(1-0.80)}{\frac{0.01^2}{2.58^2} + \frac{0.80(1-0.80)}{40000}}$$

$$= \frac{0.80(0.20)}{\frac{0.0001}{6.6564} + \frac{0.80(0.20)}{40000}}$$

$$= \frac{0.16}{0.000019 + \frac{0.16}{40000}}$$

$$= \frac{0.16}{0.000019 + 0.000004}$$

$$= \frac{0.16}{0.000023} = 8410.8 \text{ or say } 8411$$

n_2 = urban sample = $\dfrac{P_2(1-P_2)}{\dfrac{\alpha^2}{z^2} + \dfrac{P_2(1-P_2)}{N_2}}$

$$= \frac{0.20(1-0.20)}{\dfrac{0.01^2}{2.58^2} + \dfrac{0.20(1-0.20)}{10000}}$$

$$= \frac{0.20(0.80)}{\dfrac{0.0001}{6.6564} + \dfrac{0.20(0.80)}{10000}}$$

$$= \frac{0.16}{0.000019 + \dfrac{0.16}{10000}}$$

$$= \frac{0.16}{0.000019 + 0.000016}$$

$$= \frac{0.16}{0.000035} = 4568.4 \text{ or say } 4568$$

Therefore we need to have a sample of size 8411 + 4568 = 12979 units.

ii) If we wish to report the results as means (averages) of the sample responding, we use the following formula:

$$n_i = \frac{P_i^2}{\dfrac{\alpha^2}{z^2} + \dfrac{P_i^2}{N_i}}$$

Where, n_i = sample size of the i^{th} attribute required

P_i = estimated standard deviation of the i^{th} attribute of interest (for example, average income of high income group, low income group etc.)

α = precision required (0.01 or 0.05 as the case may be)

z = standardized value indicating the confidence level (z=1.96 at 95% confidence level and z=2.58 at 99% confidence level)

N_i = population size of the i^{th} attribute (known or estimated)

Example 17.9: It is planned to conduct a study to know the average income of households. Given that the standard deviation of households is 2.5 and the population size is 10000, determine the sample size required. Assume that the desired precision and confidence levels are 5% and 95% respectively.

In this example,

P_i = standard deviation of income = 2.5

N_i = number of households = 10000

α = 0.05
z = 1.96 (at 95% confidence level)

The required sample size is

$$n_1 = \frac{P_1^2}{\frac{\alpha^2}{z^2} + \frac{P_1^2}{N_1}}$$

$$= \frac{2.5^2}{\frac{0.05^2}{1.96^2} + \frac{2.5^2}{10000}}$$

$$= \frac{6.25}{\frac{0.0025}{3.8416} + \frac{6.25}{10000}}$$

$$= \frac{6.25}{0.000651 + 0.000625}$$

$$= \frac{6.25}{0.001276} = 4898$$

iii) If we wish to report the results in a variety of ways or we have the difficulty in estimating the proportion or standard deviation of the attribute of interest, we use the following formula:

$$n = \frac{0.25}{\frac{\alpha^2}{z^2} + \frac{0.25}{N}}$$

Where, n = sample size required
α = precision required (0.01 or 0.05 as the case may be)
z = standardized value indicating the confidence level (z=1.96 at 95% confidence level and z=2.58 at 99% confidence level)
N = population size (known or estimated)

Example 17.10: Given that the population size is 10000, determine the sample size required when desired precision and confidence levels are 5% and 99% respectively.

In this example,
N = 10000
α = 0.05
z = 2.58 (at 99% confidence level)

The required sample size is

$$n = \frac{0.25}{\frac{0.05^2}{2.58^2} + \frac{0.25}{10000}}$$

$$n = \frac{0.25}{\frac{0.0025}{6.6564} + \frac{0.25}{10000}}$$

$$n = \frac{0.25}{0.0003756 + 0.000025} = \frac{0.25}{0.000401} = 624$$

Check Your Progress 2

1) Say whether the following statements are true or false.
   a) When the units included in the sample are based on judgment of the investigator, the sampling is said to be random.
   b) With increasing sample size the sampling error decreases.
   c) Convenience sampling has the disadvantage that it may not be representative sample.
2) One of the major disadvantage of judgment sampling is
   a) The procedure is very cumbersome
   b) The sample selection depends on the individual judgment of the investigator
   c) It gives small sample size
   d) It is very expensive.

17.11 LET US SUM UP

The most commonly used probability sampling procedure is the simple random sampling which allows a chance to all population units to be included in the sample. The sample units are chosen using random number tables. A systematic random sample uses the first sample unit at random as a starting point and the subsequent sample units are chosen systematically. A stratified sample guarantees inclusion of units from each stratum. A cluster sample involves complete enumeration of one or more randomly selected clusters.

The non-probability sampling procedures include convenience sampling, judgment sampling, quota sampling and snowball sampling. These sampling procedures are not independent from sampling bias but still popular in some situations particularly marketing research.

A number of factors decide the sample size. It may be the number of groups in the population, the heterogeneity of population, funds and time available, etc.

Using a sample saves a lot of money, time and manpower. If a suitable sampling procedure is used in selecting units, appropriate sample size is selected and necessary precautions are taken to reduce sampling errors, then a sample should yield a valid and reliable information about the population.

17.12 KEY WORDS

Cluster Sampling : It is a sampling procedure where the entire population is divided into groups called clusters and then a random number of clusters are selected. All observations in the selected clusters are included in the sampling.

Convenience Sampling : It refers to the method of obtaining a sample that is most conveniently available to the researcher.

Judgment Sampling : In this sampling procedure the selection of sample is based on the researcher's judgment about some appropriate characteristic required of the sample units.

Multistage Sampling : The sample selection is done in a number of stages.

Quota Sampling : In this sampling procedure the samples are selected on the basis of some parameters such as age, gender, geographical region, education, income, religion, etc.

Random Sampling Random sampling is a sampling technique where we select sample from a population. Here, each unit of the population has a chance of being included in the sample.

Simple Random Sampling : It is the basic sampling procedure when we select samples using lottery method or using random number tables.

Snowball Sampling : Snowball sampling relies on referrals from initial sampling units to generate additional sampling units.

Stratified Sampling : In this sampling procedure the population is divided into groups called strata and then the samples are selected from each stratum using a random sampling method.

Systematic Sampling : A sampling procedure in which units are selected from the population at uniform interval that is measured in time, order or space.

17.13 SOME USEFUL BOOKS

Kothari, C.R.(1985) *Research Methodology : Methods and Techniques*, Wiley Eastern, New Delhi.

Levin, R.I. and D.S. Rubin. (1999) *Statistics for Management*, Prentice-Hall of India, New Delhi

Mustafi, C.K.(1981) *Statistical Methods in Managerial Decisions*, Macmillan, New Delhi.

Plane, D.R. and E.B. Oppermann. (1986) *Business and Economic Statistics*, Business Publications, Inc: Plano.

Zikmund, William G. (1988) *Business Research Methods*, The Dryden Press, New York.

17.14 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) d
2) c
3) a) False
   b) True
4) c)

Check Your Progress 2

1) a) False
   b) True
   c) True
2) b)

UNIT 10 INDEX NUMBERS

Structure

10.0 OBJECTIVES

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

- define index numbers; and
- construct and calculate them.

10.1 INTRODUCTION

An "index" in the common sense of the word is an "indicator" and no more than that. "Index numbers" or "indices" are forms of the plural, but they all mean the same thing.

An index number represents the general level of magnitude of the **changes** between two (or more) periods of time or places, in a number of variables taken as a whole. In this definition, the word "variable" refers to numerical variables which can be measured in quantity, such as the prices of commodities. For example, we may like to compare the price level of an article between 1980 and 1990 or between

Mumbai and Kolkata. Let us consider the yield of rice in 1985 and in 1990 as 50,000 and 60,000 tons respectively. The year 1985 is taken as base for comparison of yields, that is 1985 = 100. The corresponding figure for 1990 will be $\frac{60,000}{50,000} \times 100 = 120$. This is a single-commodity index number in its simplest form, being just a relative number. In practice, however, we deal usually with a number of commodities for the construction of an index.

Index numbers are ratios that are usually expressed as percentage in order to avoid awkward decimals. Thus if one commodity costs 45 paise in 1970 and Rs. 1.50 in 1974 the ratio would be

$$\frac{150}{45} \text{ or } 3.33$$

If instead of this we express the ratio into a percentage

$$\frac{150}{45} \times 100$$

we say that the index is 333, based on 1970, which is 100.

10.2 STEPS IN CONSTRUCTION OF INDEX NUMBERS

Many government and private agencies are engaged in computation of index numbers or indices as they are often required for the purpose of forecasting business and economic conditions, providing general information, etc.

It is not always the case that the comparison should be over time, but most common types of index numbers measure changes over time. Similarly, index numbers may be constructed for studying changes in any variable, such as intelligence, aptitude, efficiency, production, etc., but the time series of prices is perhaps most frequently used. Our subsequent discussion on index numbers will therefore be made with special reference to prices of commodities. The principles of construction are, however, quite general in nature, and may thus be applied to other areas of interest.

There are various uses of price index numbers. The *wholesale price index number* indicates the price changes taking place in wholesale markets. On the other hand, the *consumer price index number* or the *cost of living index number* tells us about the changes in the prices faced by an individual consumer. Its major application is in the calculation of dearness allowance so that real wage does not decrease; or in comparing the cost of living in, say, different regions. It is also used to measure changes in purchasing power of money. The reciprocal of a general price index is known as *purchasing power of money* with reference to the base period. For example, if the price index number goes up to 150, it means that the same amount of money will be able to purchase 100/150 = 0.67 times or 67% of the volume of goods being purchased in the base period.

10.2.1 Selection of Base Period

Since index numbers measure relative changes, they are expressed with one selected situation (e.g., period, place, etc.) as 100. This is called the *base* or the starting point of the series of index numbers. For example, a date is first chosen and all changes are measured from it. The base may be one day such as with index of retail prices, the average of a year or the average of a period.

While selecting a base period the following aspects should be taken into consideration:

1) The base date must be "normal" in the sense that the data chosen are not affected by any irregular or abnormal situations such as natural calamities, war, etc. It is desirable to restrict comparisons to stable periods for achieving accuracy.
2) It should not be too back-dated as the patterns of trade, imports or consumer preferences may change considerably if the time-span is too long. A ten to twenty year interval is likely to be suitable for one base date, and after that the index becomes more and more outdated. Greater accuracy is attained for moderate short-run indices than for those covering greater span of time.
3) For indices dealing with economic data, the base period should have some economic significance.

10.2.2 Choice of a Suitable Average

An index number is basically the result of averaging a series of data (e.g., *price-relatives* of several commodities). There are, however, several ways of averaging a series: mean (i.e., arithmetic mean), mode, median, geometric mean and harmonic mean.

The question naturally arises as to which average to chose. The mode has the merit of simplicity, but may be indefinite. The median suffers from the same limitations. Moreover, neither of them takes into account the size of the items at each end of a distribution. The harmonic mean has very little practical application to index numbers. As a result, mode, median and harmonic mean are not generally used in the calculation of index numbers. Thus, the arithmetic mean is most commonly used. However, the geometric mean is sometimes used despite its slight difficulty in calculation.

10.2.3 Selection of Items and their Numbers

The number and kinds of commodities to be included in the construction of an index number depend on the particular problem to be dealt with, economy and ease of calculations. Various practical considerations determine the number and kinds of items to be taken into account. For a wholesale price index, the number of commodities should be as large as possible. On the other hand, for an index meant to serve as a predictor of price movement rather than an indicator of changes over time, a much smaller number of items may be adequate. Care should, however, be taken to ensure that items chosen are not too few which make the index unrepresentative of the general level. A fixed set of commodities need not also be used for a very long period as some items lose their importance with the passage of time and some new items gain in significance. In general, the commodities should be sensitive and representative of the various elements in the price system.

10.2.4 Collection of Data

As prices often vary from market to market, they should be collected at regular intervals from various representative markets. It is desirable to select shops which are visited by a cross section of customers. The reliability of the index depends greatly on the accuracy of the quotations given for each constituent item.

10.3 METHOD OF CONSTRUCTION OF INDEX NUMBER

Various methods of construction of index numbers are as follows:

1) Relative methods
   a) Simple average of relatives
   b) Weighted average of relatives
2) Aggregative methods
   a) Simple aggregative formula
   b) Weighted aggregative formula
      i) Laspeyres' index
      ii) Paasche's index
      iii) Edgeworth-Marshell's index
      iv) Fisher's Ideal index.

10.3.1 Relative Methods

If we record prices of a variety of commodities at a given date and at a later date record the prices of similar items, the change in price can be simply expressed as a percentage of the new compared with the old for each commodity. This provides us with price relatives and if weights are available the next step will be to multiply the relatives by the weights. Finally, an index number can be produced if we add together the weighted relatives and calculate an average.

It is unrealistic to assume that the consumption of each commodity has been equal. So most indices take account of the proportions of each item actually used. This method of weighting shows the relative importance of each in the series.

Given k commodities with base year prices of

$$p_{01}, p_{02}, \ldots\ldots p_{0k}$$

and current prices of

$$p_{n1}, p_{n2}, \ldots\ldots p_{nk}$$

the price relative for the ith commodity will be $\frac{p_{ni}}{p_{0i}}$ where $i = 1, 2, \ldots\ldots, k$ and the subscript 0 refers to the base year and subscript n refers to the current year.

a) **Simple average of relatives**

The arithmetic mean of the price relative is given by

$$\text{index} = 100 \sum_{i=1}^{k} \frac{\left(p_{ni} / p_{0i}\right)}{k} \qquad \ldots\ldots(10.1)$$

For simplicity we can omit the subscript 'i' and write

$$\text{index} = 100 \sum \left(\frac{p_n / p_0}{k} \right)$$

b) **Weighted average of relatives**

The most suitable weights to use are the value of each item, which is denoted by w_i for the i-th commodity. One may use the value of base year quantities sold at the base year prices ($w_{0i} = p_{0i}q_{0i}$) or current year quantities sold at current prices ($w_{1i} = p_{1i}q_{1i}$) or any other value as weights. The weights can also be a set of constant factors derived rationally.

A weighted arithmetic mean of price relatives using **base year values as weights** is given by

$$\text{index} = \frac{\sum \frac{p_n}{p_0} \times w_0}{\sum w_0} \times 100 \qquad \text{.....(10.2)}$$

omitting suffix i for simplicity. It may be noted that base year weighting preserves continuity, but loses "up-to-dateness" in the course of time.

Example 10.1: The table below presents the average fares per railway journey. Using 1948 average = 100, calculations are made according to base year weights.

Class of ticket	No. of passenger journeys in 1948 in millions (q_o)	Fare (Rs.) 1948 (p_o)	Fare (Rs.) 1969 (p_n)	Weights $w_o = p_o q_o$	Price relative $P=(p_n/p_o)\times 100$	$P.w_o$
Full fare	23	12	60	276	500	138000
Excursions	25	6	30	150	500	75000
Festival	20	4	15	80	375	30000
Season tickets	32	5	14	160	280	44800
Total				**666**		**287800**

Applying formula (10.2), we get

$$\text{index for 1969} = \frac{287800}{666} = 432.13.$$

Using **current year values** ($w_n = p_n q_n$) **as weights**, the index is given by

$$\text{index} = \frac{\sum \frac{p_n}{p_o} \times w_n}{\sum w_n} \times 100 \qquad \text{.....(10.3)}$$

Example 10.2: The table below shows the average fares per railway journey. Using 1948 average = 100, calculations are made according to current year weights.

Class of ticket	No. of passenger journeys in 1948 in millions (q_n)	Fare (Rs.) 1948 (p_o)	Fare (Rs.) 1969 (p_n)	Weights $w_n = p_n q_n$	Price relative $P=(p_n/p_o)\times 100$	$P.w_n$
Full fare	25	12	60	1500	500	750000
Excursions	26	6	30	780	500	390000
Festival	9	4	15	135	375	50630
Season tickets	27	5	14	378	280	105800
Total				**2793**		**1296430**

Applying formula (10.3), we get

$$\text{index} = \frac{1296430}{2793} = 464.17$$

10.3.2 Aggregative Methods

In this method, the aggregate (sum-total) of the prices of all commodities in the current or given year is expressed as a percentage of the same in the base year. Thus, in the case of **simple aggregative index,** we have:

$$\text{Index number} = \frac{\text{aggregate prices in the current year}}{\text{aggregate prices in the base year}} \times 100$$

$$= \frac{p_{n1} + p_{n2} + \ldots\ldots\ldots + p_{nk}}{p_{01} + p_{02} + \ldots\ldots\ldots + p_{0k}} \times 100$$

$$= \frac{\sum p_{ni}}{\sum p_{0i}} \times 100 = \frac{\sum p_n}{\sum p_0} \times 100 \qquad \ldots (10.4)$$

where the summation $\left(\sum_{i=1}^{k}\right)$ extends over all selected commodities numbering k.

On the other hand, in the case of **weighted aggregative index** we have,

$$\text{General index} = \frac{p_{n1}q_1 + p_{n2}q_2 + \ldots\ldots\ldots + p_{nk}q_k}{p_{01}q_1 + p_{02}q_2 + \ldots\ldots\ldots + p_{0k}q_k} \times 100$$

$$= \frac{\sum p_{ni}q_i}{\sum p_{0i}q_i} \times 100$$

or simply

$$= \frac{\sum p_n q}{\sum p_0 q} \times 100 \qquad \ldots (10.5)$$

The weights used should be actual quantities bought or sold, and these are kept unchanged until such time as the index requires to be revised.

There are many formulae for weighted aggregative index, but depending on the type of weights used, we discuss four indices which are commonly used.

a) **Laspeyres' index**

If we use base period quantities (q_0) as the weights in the general weighted aggregative index formula (10.5), we get what is known as Laspeyre's formula (L).

$$L = \frac{\sum p_n q_0}{\sum p_0 q_0} \times 100 \qquad \ldots (10.6)$$

It can be seen that this index has fixed base year quantity as weights (q_0) and is equivalent to a arithmetic mean of price relatives given at formula (10.2). Thus, we can also write (10.6) as

$$L = \frac{\sum \frac{p_n}{p_0} \times p_0 q_0}{\sum p_0 q_0} \times 100$$

b) **Paasche's index**

If we use current year quantities (q_n) as weights in the general aggregative index formula (10.5), we get what is known as Paasche's formula (P).

$$P = \frac{\sum p_n q_n}{\sum p_0 q_n} \times 100 \qquad \ldots (10.7)$$

where q_n (actually $q_{n1}, q_{n2}, \ldots q_{nk}$) are the quantities bought or sold in the current period.

c) **Fisher's Ideal Index**

An index number obtained as geometric mean (i.e., square root of the product) of indices obtained by Laspeyres' and Paasche's formulae, satisfies certain important properties (to be discussed later), is known as Fisher's ideal formula

$$F = \sqrt{L \times P} = \sqrt{\frac{\sum p_n q_0}{\sum p_0 q_0} \times \frac{\sum p_n q_n}{\sum p_0 q_n}} \times 100 \qquad \ldots (10.8)$$

d) **Edgeworth-Marshall Index**

If the mean of the base period and the current period quantities is used as weight, i.e.,

$w = \frac{1}{2}(q_0 + q_n)$, we get what is known as a compromise formula of Edgeworth-Marshall index.

$$I = \frac{\sum p_n (q_0 + q_n)/2}{\sum p_0 (q_0 + q_n)/2} \times 100$$

$$= \frac{\sum p_n (q_0 + q_n)}{\sum p_0 (q_0 + q_n)} \times 100 \qquad \ldots (10.9)$$

Table 10.1:
Illustrative calculations of Laspeyres', Paasche's, Edgeworth-Marshall's and Fisher's indices

Item	Base Year (1970)		Current Year (1980)					
	Price (p_0)	Quantity (q_0)	Price (p_n)	Quantity (q_n)	p_0q_0	p_nq_0	p_0q_n	p_nq_n
A	20	7	25	9	140	175	180	225
B	42	6	40	8	252	240	336	320
C	30	17	25	4	510	425	120	100
D	8	15	14	10	120	210	80	140
E	10	8	13	5	80	104	50	65
Total					**1102**	**1154**	**766**	**850**

1) Laspeyres' price index $= \frac{\sum p_nq_0}{\sum p_nq_0} \times 100 = \frac{1154}{1102} \times 100 = 104.72 = 105$

2) Paasche's price index $= \frac{\sum p_nq_n}{p_0q_n} \times 100 = \frac{850}{766} \times 100 = 110.97 = 111$

3) Edgeworth-Marshall's index $= \frac{\sum p_nq_0 + \sum p_nq_n}{\sum p_0q_0 + \sum p_0q_n} \times 100$

$= \frac{1154+850}{1102+766} \times 100$

$= \frac{2004}{1868} \times 100 = 107.28 = 107$

4) Fisher's ideal index $= \sqrt{\frac{\sum p_nq_0}{\sum p_0q_0} \frac{\sum p_nq_n}{\sum p_0q_n}} \times 100$

$= \sqrt{[(L) \times (P)]} = \sqrt{(104.72 \times 110.97)}$

$= 107.8 = 108$

Note that for the same price change different formulae provide different values. Moreover, when prices are increasing, Laspeyres' index gives the lowest value while Paasche's index gives the highest value. Therefore, it is often said that Laspeyres' index is an under-estimate while Paasche's index is an over-estimate of true price change.

10.3.3 Quantity or Volume Index Numbers

We can get a quantity or volume index number, which measures and permits comparison of quantities of goods, from corresponding price index number formulae simply by replacing p by q and q by p.

1) Quantity relative $= \frac{q_n}{q_0} \times 100$

2) Arithmetic Mean (A.M.) of quantity of relatives $= 100 \sum \left(\frac{q_n}{q_0} \right) / k$

3) Weighted A.M. of quantity relative index:

a) Base year weights: $\frac{\sum(q_n/q_0)\times w_0}{\sum w_0}\times 100$ (where $w_0 = p_0q_0$)

b) Current year weights: $\frac{\sum(q_n/q_0)\times w_n}{\sum w_n}\times 100$ (where $w_n = p_nq_n$)

4) Simple aggregative quantity index = $\frac{\sum q_n}{\sum q_0}\times 100$

5) Laspeyres' quantity index = $\frac{\sum q_nP_0}{\sum q_0P_0}\times 100$

6) Paasche's quantity index = $\frac{\sum q_nP_n}{\sum q_0P_n}\times 100$

7) Fisher's ideal index = $\sqrt{\frac{\sum q_np_0}{\sum q_0p_0}\frac{\sum q_np_n}{\sum q_0p_n}}\times 100$

8) Edgeworth-Marshall's index = $\frac{\sum q_n(p_0+p_n)}{\sum q_0(p_0+p_n)}\times 100$

Check Your Progress 1

1) What do index numbers seek to measure?

.....

.....

.....

.....

.....

2) Discuss the various problems involved in construction of index numbers with particular reference to price indices.

.....

.....

.....

.....

.....

3) The following are the prices of six different commodities for 1983 and 1984. Compute the price index by (a) aggregative method, (b) average of price relatives method by using arithmetic mean.

Commodities	Price in 1983 (Rs.)	Price in 1984 (Rs.)
A	40	50
B	50	60
C	20	30
D	50	70
E	80	80
F	100	110

.....

.....

4) Calculate Fisher's Ideal Index Number from the following group of items.

Item No.	Base Year		Current Year	
	Price (in Rs.)	Quantity (in kg)	Price (in Rs.)	Quantity (in kg)
1	4	1.0	3	4
2	8	1.5	7	5

.....

.....

.....

.....

.....

.....

5) Calculate Laspeyres' and Paasche's Index Numbers from the following data:

Item	Base Year		Current Year	
	Quantity	Price per pound	Quantity	Price per pound
Bread	6.0	40 paise	7.0	30 paise
Meat	4.0	45 paise	5.0	50 paise
Tea	0.5	90 paise	1.5	40 paise

.....

.....

.....

.....

.....

.....

10.4 MERITS OF THE VARIOUS AGGREGATIVE MEASURES

The different index numbers serve different purposes and, therefore, the appropriateness of a particular index number depends on the purpose at hand.

The Laspeyres' index calculation is simpler, since this uses the base period quantities as weights which are not difficult to get and the denominator needs calculating only once. But in this index a rise in prices tends to be *overstated,* since it does not take into account corresponding falls in demand or changes in output. Indices such as Paasche's, on the other hand, use current period quantities as weights which are difficult to get and the weights need to be constructed afresh for every year. Moreover, Paasche's index tends to *understate* the rise in prices because it uses current weights.

The Laspeyres' index is probably more commonly used, since it is convenient to employ fixed weights. But with the passage of time the weights are rendered out of date. For example, in 1970 the number of TVs in Calcutta was nil. In 1990, there are more TVs than refrigerators. The Paasche's index uses the preferable current weights, but since up-to-date information on quantity of goods produced or consumed or marketed or distributed are not readily obtained, the Laspeyres' index has a great advantage.

10.5 TESTS FOR INDEX NUMBERS

A perfect index number, which measures the change in the level of a phenomenon from one period to another, should satisfy certain tests. There are three major tests of index numbers: (1) Time reversal test, (2) Factor reversal test, and (3) Circular test.

10.5.1 The Time Reversal Test

According to this test, if we reverse the time subscripts (such as 0 and n) of a price (or quantity) index the result should be the reciprocal of the original index.

Symbolically,

$$I_{0n} \times I_{n0} = 1$$

where I_{0n} = index number for period n with the base period 0

I_{n0} = index number for period 0 with the base period n.

If from 1975 to 1982 the price changes from Rs. 4 to Rs. 16, the price in 1982 is 400 percent of the price in 1975, and the price in 1975 is 25 percent of the price in 1982. The product of the two price relatives is $4 \times 0.25 = 1$. The test is based on the analogy that the principle, which holds good for a single commodity, should also be true for the index number as a whole.

There are five methods which do satisfy the time reversal test. These are:

1) Simple geometric mean of price relatives
2) Aggregative indices with fixed weights
3) Edgeworth-Marshall formula
4) Weighted geometric mean of price relatives if fixed weights are used
5) Fisher's ideal index

Fisher's ideal index $F = \sqrt{\frac{\sum p_n q_0}{\sum p_0 q_0} \times \frac{\sum p_n q_n}{\sum p_0 q_n}}$

If time subscripts are reversed,

$$F' = \sqrt{\frac{\sum p_0 q_n}{\sum p_n q_n} \times \frac{\sum p_0 q_0}{\sum p_n q_0}}$$

Since $F \times F' = 1$, the test is satisfied.

10.5.2 The Factor Reversal Test

With the usual notations, a "value index" formula is given by

$$I_v = \frac{\sum p_n q_n}{\sum p_0 q_0}$$

Now, for example, Laspeyres' index for prices and quantities are given respectively by

$$I_p = \frac{\Sigma p_n q_0}{\Sigma p_0 q_0}$$

and $$I_q = \frac{\Sigma q_n p_0}{\Sigma q_0 p_0}$$

The factor reversal test desires that $I_p . I_q = I_v$

But for Laspeyres' index

$$Ip.Iq = \frac{\Sigma(p_n q_0)(\Sigma q_n p_0)}{\Sigma(p_0 q_0)^2} \neq I_v$$

On the other hand, Fisher's ideal index satisfies this test, as shown below.

$$I_p = \sqrt{\frac{\Sigma p_n q_0}{\Sigma p_n q_0} \times \frac{\Sigma p_n q_n}{\Sigma p_0 q_n}}$$

$$I_q = \sqrt{\frac{\Sigma q_n p_0}{\Sigma q_n p_0} \times \frac{\Sigma q_n p_n}{\Sigma q_0 p_n}}$$

$$I_p I_q = \sqrt{\frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_n q_n}{\Sigma p_0 q_n} \times \frac{\Sigma q_n p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_n p_n}{\Sigma q_0 p_n}}$$

$$= \sqrt{\frac{\Sigma p_n q_n}{\Sigma p_0 q_0} \times \frac{\Sigma q_n p_n}{\Sigma q_0 p_0}} = \frac{\Sigma p_n q_n}{\Sigma p_0 q_0} = I_v$$

To understand this principle further, we take the following example.

If the price and quantity per unit of an item changed in 1990, as compared to 1970, from Rs. 16 to Rs. 32 and from 100 units to 200 units respectively, then the price and quantity in 1990 would both be 200% or 2.00 times the price and quantity in 1970. The values (product of price and quantity) would be Rs. 1600 in 1970 and Rs. 6400 in 1990, so that the value ratio is 6400/1600 = 4.00. Thus, we verify that 2.00 ′ 2.00 = 4.00, that is, the product of price ratio and quantity ratio is equal to the value ratio.

Only the Fisher's ideal index satisfies this test.

Example 10.3: We show with the following data that the Fisher's ideal index satisfies the factor reversal test:

Item	Price (Rs.)		No. of units		p_0q_0	p_nq_0	p_0q_n	p_nq_n
	1983 (p_0)	1989 (p_n)	1983 (q_0)	1989 (q_n)				
I	6	10	50	56	300	500	336	560
II	2	2	100	120	200	200	240	240
III	4	6	60	60	240	360	240	360
IV	10	12	30	24	300	360	240	288
V	8	12	40	36	320	480	288	432
Total					**1360**	**1900**	**1344**	**1880**

Price Ratio: $I_p = \sqrt{\frac{\sum p_n q_0}{\sum p_n q_0} \times \frac{\sum p_n q_n}{\sum p_0 q_n}} = \sqrt{\frac{1900}{1360} \times \frac{1880}{1344}}$

Quantity Ratio: $I_p = \sqrt{\frac{\sum q_n p_0}{\sum q_n p_0} \times \frac{\sum q_n p_n}{\sum q_0 p_n}} = \sqrt{\frac{1344}{1360} \times \frac{1880}{1900}}$

Value Ratio: $I_v = \frac{\sum p_n q_n}{\sum p_0 q_0} = \frac{1880}{1360}$

$$I_p . I_q = \sqrt{\frac{1900}{1360} \times \frac{1880}{1344} \times \frac{1344}{1360} \times \frac{1880}{1900}} = \sqrt{\frac{1880}{1360} \times \frac{1880}{1360}}$$

$$= \frac{1880}{1360}$$

$= I_v$ which shows that the test is satisfied.

10.5.3 Chain Index Number and Circular Test

Two types of base periods are used for the construction of index numbers, namely, (a) fixed base, (b) chain base. Most commonly used indices use fixed base method. This method cannot take into account any changes in price or quantity in any other year. It fails to include new commodities gaining importance at a later date or exclude commodities losing significance in course of time. These problems can be overcome by chain index numbers.

Using a suitable index number formula (say, Laspeyres' index), link indices, defined as follows, are first calculated: Link index = Index number with previous period as base. The chain index is obtained by multiplying link indices progressively. Thus, the chain index number I_{0n} for period n with base period 0 is given by

$I_{01} = I_{01}$
$I_{02} = I_{01} \times I_{12}$
$I_{03} = I_{01} \times I_{12} \times I_{23} = I_{02} \times I_{23}$
.....
.....
$I_{0n} = I_{01} \times I_{12} \times ..... \times I_{(n-2)(n-1)} \times I_{(n-1)n}$

Example 10.4: The calculation of chain index numbers is illustrated with reference to the following data:

Year	Link index	Chain index (Base 1970 = 100)
1970	100	100
1971	$I_{01} = 80$	$100 \times \frac{80}{100} = 80$
1972	$I_{12} = 120$	$80 \times \frac{120}{100} = 96$
1973	$I_{23} = 75$	$96 \times \frac{75}{100} = 72$

Thus, the chain index numbers for the years 1971 to 1973 with 1970 as the base are 80, 96 and 72 respectively.

Circular Test: The circular test is an extension of time reversal test over a number of years. It states that the chain index for the year 1973, calculated above, starting from the base year 1970 will be same as the index number directly calculated with fixed base period of 1970. In symbols,

$I_{01} \times I_{12} \times ..... \times I_{(n-1)n} \times I_{n0} = 1$. (Notice that $I_{0n} = \frac{1}{I_{n0}}$)

Considering an aggregate index with fixed weights

$$\frac{\sum p_1 q}{\sum p_0 q}$$

we can illustrate the test as follows:

With base period 0, we can trace the above formula from 1 to 3 years:

$$\frac{\sum p_1 q}{\sum p_0 q} \times \frac{\sum p_2 q}{\sum p_1 q} \times \frac{\sum p_3 q}{\sum p_2 q} \times \frac{\sum p_0 q}{\sum p_3 q} = 1$$

The formulae satisfying the requirements of circular test are:

1) Simple aggregative index
2) Simple geometric mean of relatives
3) Weighted aggregative index (such as Laspeyres' index with constant weights)
4) Weighted geometric mean of relatives with constant weights.

Fisher's ideal index does not satisfy this test. It has been proved that no index satisfies both the factor reversal and the circular tests.

Check Your Progress 2

1) Compute the chain index number with 1980 prices as base from the following table giving the average wholesale prices of commodities A, B and C for years 1980 - 84

Commodity	Average whole sale Price (in Rs.)				
	1980	1981	1982	1983	1984
A	20	16	28	35	21
B	25	30	24	36	45
C	20	25	30	24	30

.....

.....

.....

.....

.....

.....

.....

2) Construct Fisher's Ideal Index number from the following data and show that it satisfies Factor and Time Reversal Tests.

Commodities	Base Year		Current Year	
	Price per unit	Expenditure (Rs.)	Price per unit	Expenditure (Rs.)
A	2	40	5	75
B	4	16	8	40
C	1	10	2	24
D	5	25	10	60

.....

.....

.....

.....

.....

.....

10.6 COST OF LIVING INDEX NUMBER (CLI) OR CONSUMER PRICE INDEX NUMBER (CPI)

This is an index of changes in the prices of goods and services commonly consumed by a homogeneous group of people, such as families of industrial workers. The major items of consumption that are considered for the construction of CLI are:

1) Food
2) Fuel and Light
3) Clothing
4) House rent
5) Miscellaneous

The common method for obtaining the consumption basket is to conduct a family living survey among the population group for which the index is to be constructed. Prices of selected items are also collected from various retail markets used by consumers in question. It may be noted that each of the above broad groups contains several sub groups. Thus, 'food' includes cereals, pulses, oils, meat, fish, egg, spices, vegetables, fruits, non-alcoholic beverages, etc. 'Miscellaneous' includes such items as medical care, education, transport, recreation, gifts and many

others. When more than one price quotation is collected for a single commodity, a simple average is taken. Index number is constructed for each of the five groups using weighted average of the price group; the weights used are proportional to the expenditure on the consumed item by an average family. Next, the overall index (CLI) is computed as an weighted average of group indices, the weights being again the proportional expenditure on different groups (e.g. 50 per cent on food).

Using Laspeyres' formula

$$\text{Cost of living Index: } I = \frac{\sum w\left(\frac{p_n}{p_0}\times 100\right)}{\sum w}$$

where $w = \frac{p_0 q_0}{\sum p_0 q_0}$, is the weight of a group index.

The CLI or consumer price index (CPI) numbers have significant practical implications and extensive public use. Its use as a wage regulator is the most important. The dearness allowance of the employees are primarily determined by this index. When wages or incomes are divided by corresponding CLI, the effect of rise or fall of prices is eliminated. This is known as the process of deflation, which is used to find 'real wages' or 'real income'. As mentioned earlier the reciprocal of CLI measures the purchasing power of money.

Example 10.5: Construction of an Index for food

Item	Prices			Weights	
	p_n	p_0	$P = (p_n \times p_0) \times 100$	w	Pw
Rice	50	40	125.0	30	3750.0
Wheat	45	30	150.0	20	3000.0
Pulses	60	40	150.0	10	1500.0
Sugar	40	20	200.0	5	1000.0
Oil	75	60	125.0	15	1875.0
Potato	60	50	120.0	15	1800.0
Fish	200	150	133.3	5	666.5
Total				**100**	**13591.5**

$$\text{Index (food)} = \frac{\sum w \times (p_n \div p_0)}{\sum w} \times 100$$

$$= \frac{13591.5}{100} = 135.915 = 135.92$$

Example 10.6: Construction of a Final Cost of Living Index Number.

Item	Weight (Percentage Expenditure)	Index	Wt. × Index
Food	45	130	5850
Clothing	15	140	2100
Housing	20	170	3400
Fuel	5	110	550
Misc.	15	125	1875
Total	**100**		**13775**

Cost of Living Index $= \frac{13,775}{100} = 137.75 = 138$

Check Your Progress 3

1) Calculate a number which will indicate the percentage change in volume of traffic from October 1979 to October 1980, when account is taken of the relative values of the different types of traffic.

Type of traffic	Tons ('000)		Receipts (Rs.'000)
	Oct. 1979	Oct. 1980	Oct. 1979
Merchandise	1246	1206	776
Minerals	1125	981	252
Fuel	4794	4229	562

.....

.....

.....

.....

.....

.....

2) Compute Paasche's price index number for 1980 with 1975 as base from the following data:

Commodity	Unit	Price (Rs.) per unit		Quantities sold	
		1970	1980	1970	1980
A	Kg.	4	5	95	120
B	Kg.	60	70	118	130
C	Kg.	35	40	50	70

.....

.....

.....

.....

3) From the following data, compute Laspeyres' price index number for 1980 with 1978 as base:

Item	Price (Rs.)		Total Value (Rs.)
	1978	1980	1978
A	12.50	14.00	112.50
B	10.50	12.00	126.00
C	15.00	14.00	105.00
D	9.40	11.20	47.00

.....

.....

.....

.....

.....

.....

4) Calculate Marshall-Edgeworth index number from the following data:

Commodity	1970		1977	
	Price	Quantity	Price	Quantity
Rice	9.3	100	4.5	90
Wheat	6.4	11	3.7	10
Jowar	5.1	5	2.7	3

.....

.....

.....

.....

.....

.....

10.7 WORKED OUT EXAMPLES

In this section, we shall provide worked out examples so as to further familiarise you with the topic.

Example 10.7: Construction of Price Index

Item	Unit	Price per unit (Rs.)		
		1970 (p_0)	1980 (p_n)	$(p_n \div p_0) \times 100$
Rice	quintal	100.00	220.00	220
Wheat	kg.	1.50	2.40	160
Fish	kg.	15.00	28.00	187

Bread	lb.	0.60	1.35	225
Milk	litre	2.50	4.00	160
Total		**119.60**	**255.75**	**952**

a) **Aggregative method**

Index number for 1980 (base 1970 = 100)

$$\frac{\text{Average price per unit in 1980}}{\text{Average price per unit in 1970}} \times 100$$

$$= \frac{\sum p_n / k}{\sum p_0 / k} \times 100 = \frac{255.75}{119.60} \times 100 = 214$$

b) **Method of price relative**

Index number for 1980 (base 1970 = 100)

$$= \frac{\sum\left(\frac{p_n}{p_0} \times 100\right)}{k}$$

$$= {}^{952}/_{5} = 190.$$

Example 10.8: Calculate price index numbers from the following information, using (a) weighted aggregative formula, and (b) weighted arithmetic mean of price relatives:

Item	Unit	Price (Rs.) per unit		Weight
		Base Year	Current Year	
A	quintal	85	115	19
B	Kg.	15	15	25
C	dozen	45	61	40
D	litre	55	100	20
E	Lb	17	23	21

Calculation for Index numbers

Item	p_0	p_n	w	p_0w	p_nw	$I = (p_n \div p_0) \times 100$	Iw
A	85	115	19	1615	2185	135.3	2570.7
B	15	20	25	375	500	133.3	3332.5
C	45	61	40	1800	2440	135.6	5424.0
D	55	100	20	1100	2000	181.8	3636.0
E	17	23	21	357	483	135.3	2841.3
Total			**125**	**5247**	**7608**		**17804.5**

a) Weighted aggregative index $= \frac{\sum p_n w}{\sum p_0 w} \times 100 = \frac{7608}{5247} \times 100 = 145.0$

b) Weighted arithmetic mean of price relatives

$$= \frac{\sum iw}{\sum w} = \frac{17804.5}{125} = 142.4$$

Example 10.9: Given below are the data on prices of some consumer goods and the weights attached to the various commodities. Calculate price index numbers for the year 1971 (base 1970 = 100), using (a) simple average, and (b) weighted average of price relatives.

Commodities	Unit	Price (Rs.)		Weights
		1970	1971	
Wheat	Kg.	0.50	0.75	2
Milk	Litre	0.60	0.75	5
Egg	Dozen	2.00	2.40	4
Sugar	Kg.	1.80	2.10	8
Shoes	Pair	8.00	10.00	1

Calculations for price relative index.

Commodities	Unit	p_0	p_n	$I = (p_n \div p_0) \times 100$	w	Iw
Wheat	Kg.	0.50	0.75	150	2	300
Milk	Litre	0.60	0.75	125	5	625
Egg	Dozen	2.00	2.40	120	4	480
Sugar	Kg.	1.80	2.10	117	8	936
Shoes	Pair	8.00	10.00	125	1	125
Total	—	—	—	**637**	**20**	**2466**

a) Simple average of price relative index $= \frac{\sum\left(p_n / p_0\right) \times 100}{k} = \frac{637}{5} = 127.4$

b) Weighted average of price relative index $= \frac{\sum Iw}{\sum w} = \frac{2466}{20} = 123.3$

Example 10.10: On the basis of the following data, calculate the wholesale price index number for the five groups combined.

Group	Weight	Index no. for week ending 27.9.69 (Base: 1952-53 = 100)
Food	50	241
Liquor and tobacco	2	221
Fuel, power, light and lubricants	3	204

Industrial raw materials	16	256
Manufactured commodities	29	179

We compute: General index = $\frac{\sum Iw}{\sum w}$

where I = Group index, and w = Group weight

Group	Weight (w)	Group Index (I)	Iw
Food	50	241	12050
Liquor and tobacco	2	221	442
Fuel, power, light and lubricants	3	204	612
Industrial raw materials	16	256	4096
Manufactured commodities	29	179	5191
Total	**100**		**22391**

Index number of wholesale prices = $\frac{22391}{100} = 223.91$

Example 10.11: Annual production (in million tons) of four commodities are given below:

Commodity	Production			Weight
	1950	1954	1955	
A	160	200	216	20
B	24	42	45	30
C	50	72	68	13
D	120	168	156	17

Calculate quantity index numbers for the 2 years 1954 and 1955 with 1950 as base year, using (a) simple arithmetic mean, and (b) weighted arithmetic mean of the relatives.

Quantity relatives for 1954 with base year 1950(= 100)

$I = (q_n / q_0) \times 100 = (q_{54} / q_{50}) \times 100$

Commodity A: $\frac{200}{160} \times 100 = 125$

Commodity B: $\frac{42}{24} \times 100 = 175$

Commodity C: $\frac{72}{50} \times 100 = 144$

Commodity D: $\frac{168}{120} \times 100 = 140$

Quantity relatives for 1955 with 1950 = 100

$$I = (q_{55} / q_{50}) \times 100$$

Commodity A: $\frac{216}{160} \times 100 = 135$

Commodity B: $\frac{45}{24} \times 100 = 187.5$

Commodity C: $\frac{68}{50} \times 100 = 136$

Commodity D: $\frac{156}{120} \times 100 = 130$

Commodity	Quantity relatives (I)		Weight (w)	Iw	
	1954	1955		1954	1954
A	125	135.0	20	2500	2700
B	175	187.5	30	5250	5625
C	144	136.0	13	1872	1768
D	140	130.0	17	2380	2210
Total	**584**	**588.5**	**80**	**12002**	**12303**

a) Simple arithmetic mean of quantity relatives

$$= \frac{\Sigma(q_n / q_0)}{k} \times 100$$

(where k = number of commodities)

Index number for 1954 $= \frac{584}{4} = 146$

Index number for 1955 $= \frac{588.5}{4} = 147$

b) Weighted arithmetic mean of quantity relatives $\frac{\Sigma Iw}{\Sigma w}$

Index number for 1954 $= \frac{12002}{80} = 150$

Index number for 1955 $= \frac{=12303}{80} = 154$

Example 10.12: From the following price (p) and quantity (q) data, compute Fisher's ideal index number.

Commodity	1970 (Base Year)		1978 (Current Year)	
	Price	Quantity	Price	Quantity
A	12	10	17	10
B	14	9	16	11
C	11	12	13	10

Calculations for Fisher's ideal index:

Commodity	p_0	q_0	p_n	q_n	p_0q_0	p_nq_0	p_0q_n	p_nq_n
A	12	10	17	10	120	170	120	170
B	14	9	16	11	126	144	154	176
C	11	12	13	10	132	156	110	130
Total					**378**	**470**	**384**	**476**

$$\text{Laspeyres' price index} = \frac{\sum p_n q_0}{\sum p_n q_0} \times 100 = \frac{470}{378} \times 100 = 124.34 = 124$$

$$\text{Paasche's price index} = \frac{\sum p_n q_n}{\sum p_0 q_n} \times 100 = \frac{476}{384} \times 100 = 123.96 = 124$$

$$\text{Fisher's ideal index} = \sqrt{L \times P} = \sqrt{(124 \times 124)} = 124.$$

10.8 LET US SUM UP

In this Unit you have been introduced to the concepts and methods involved in the construction of Index Numbers. You have been shown how to use the Laspeyres', Paasche's and Fisher's formulae for calculating price as well as quantity indices. You also know now how to measure changes in consumer price or cost of living.

10.9 KEY WORDS

Base Year : Preferably a normal year, in terms of variable concerned, base year index is invariably taken as 100. Current year index is expressed as a percentage of base year index.

Chain Index : Current year's index is expressed as a percentage of previous year's index.

Index Number : A pure number, expressed as a percentage to base year value. Index Number measures the relative changes over time in the variable concerned (price, quantity sales or say, exports) of a group of commodities. This is a special type of weighted average of prices (or any other attribute) of the commodities or items included.

Price Relative : In the construction of a an index number price relative for a commodity is the ratio of the current year price to base year price of that commodity.

Quantity Index Number : The variable considered is the quantity of commodities.

10.10 SOME USEFUL BOOKS

Nagar, A.L. and R.K. Das, 1989: *Basic Statistics*, Oxford University Press, Delhi.

Goon, A.M., M.K. Gupta and B. Dasgupta, 1987: *Basic Statistics*, The World Press Pvt. Ltd., Calcutta.

10.11 ANSWERS AND HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) and (2): Do it yourself.

3) Simple Aggregative Index Number = 117.14

Average of Price Relative Method =122.9

4) 84.2

5) Laspeyres' Index Number = 86.02

Paasche's Index Number = 81.25

Check Your Progress 2

1) 108.33, 135.41, 160.23,165.56

2) Do it yourself.

Check Your Progress 3

1) We find quantity for Oct. 1980 with Oct. 1979 as base. The required index may be obtained as the weighted arithmetic mean of quantity relatives, using the receipts in 1979 as weights.

Type of Traffic	q_0	q_n (w)	Weight	Quantity Relative $(q_n \div q_0) \times 100$	(4) × (5)
(1)	(2)	(3)	(4)	(5)	(6)
Merchandise	1246	1206	776	97	75272
Minerals	1125	981	252	87	21924
Fuel	4794	4229	562	88	49456
Total	—	—	**1590**	—	**146652**

$$\text{Quantity index} = \frac{\sum(q_n/q_0)\times 100 \times w}{\sum w} = \frac{146652}{1590} = 92$$

2) Calculation for Paasche's price index

Commodity	p_0	p_n	q_0	q_n	p_0q_n	p_nq_n
A	4	5	95	120	480	600
B	60	70	118	130	7800	9100
C	35	40	50	70	2450	2800
Total					**10730**	**12500**

$$\text{Paasche's price index} = \frac{\sum p_nq_n}{\sum p_0q_n}\times 100 = \frac{12500}{10730}\times 100 = 116$$

3) We are given the base price (p_0), current price (p_n) and value in the base year (p_0q_0). To find base year quantity (q_0), we can use the relation

$$q_0 = \frac{p_0 q_0}{p_0}$$

Using p_0, p_n and q_0, we can find Laspeyres' index as

$$L = \frac{\sum p_n q_0}{\sum p_0 q_0} \times 100.$$

Calculation for Laspeyres' price index

Item	p_0	p_n	$p_0 q_0$	$q_0 = p_0 q_0 \div p_0$	$p_n q_0$
A	12.50	14.00	112.50	9	126.00
B	10.50	12.00	126.00	12	144.00
C	15.00	14.00	105.00	7	98.00
D	9.40	11.20	47.00	5	56.00
Total	—	—	**390.50**	—	**424.00**

Laspeyres' price index $= \frac{\sum p_n q_0}{p_0 q_0} \times 100 = \frac{42.00}{390.0} \times 100 = 109.$

4) Marshall-Edgeworth index $= \frac{\sum p_n (q_0 + q_n)}{\sum p_0 (q_0 + q_n)} \times 100$

$$= \frac{\sum P_n q_0 + \sum P_n q_n}{\sum P_0 q_0 + \sum P_0 q_n} \times 100$$

Let us take 1970 as base and 1977 as current year.

Commodity	p_0	q_0	p_n	q_n	$p_0 q_0$	$p_0 q_n$	$p_n q_0$	$p_n q_n$
Rice	9.3	100	4.5	90	930.0	837.0	450.0	405.0
Wheat	6.4	11	3.7	10	70.4	64.0	40.7	37.0
Jowar	5.1	5	2.7	3	25.5	15.3	13.5	8.1
Total					**1025.9**	**916.3**	**504.2**	**450.1**

Required index $= \frac{504.2 + 450.1}{1025.9 + 916.3} \times 100 = 49.1.$

UNIT 11 DETERMINISTIC TIME SERIES AND FORECASTING

Structure

11.0 OBJECTIVES

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

- construct a trend line for time series data;
- compute moving averages; and
- calculate various measures of seasonal variations.

11.1 INTRODUCTION

A time series is a set of observations on a variable measured at successive points of time. Usually the variable values are recorded over equal time intervals — yearly, quarterly, monthly, etc. Generally the term 'time series' refers to economic data, but it equally applies to quantitative data collected in other fields also. The time series of National Income, Agricultural Income, Agricultural Production are based on annual observations. Other examples of time series are yield of a crop in different years, population of a country over different points of time, sales of a departmental store during different seasons of the year, quarterly exports of tea, etc. For these types of data one of the variables is time, denoted by t, and the other which is

dependent on time (such as yield, population, sale or export) is represented by y_t. We shall analyse some of these series with the help of the methodology to be developed in this Unit.

11.2 PROBLEMS AND OBJECTS OF STUDY OF TIME SERIES DATA

A study of time series data reveals that in general the observed values of the dependent variable (y_t) change over time. These changes are due to interaction of several forces such as increase in population, change in production techniques, change of tastes and habits of people, variations in climate, etc. Part of these changes are long term while others may be seasonal or cyclical. One of the main objectives of study of time series data is to isolate and measure the effects of various components. This analysis helps us in understanding the past behaviour and predicting the future. Such prediction is of utmost importance to an economist or a producer who can plan his production much ahead of sales.

11.2.1 Components of Time Series

A graphical representation of time series data reveals changes over time. (In rather exceptional cases the series exhibits no change during the period of observation.) However, these changes are not totally haphazard or random and at least a part of it can be explained. Some of the movements are periodic in nature while some others show persistent growth or decline. Along with these some unpredictable movements, random in nature, are also found to be mixed up. Again, not every series shows all the movements. It is assumed that the general series has four important components.

i) Secular or long-term trend (T)

ii) Seasonal variation (S)

iii) Cyclical fluctuation (C)

iv) Irregular or random movement (I)

In the classical approach, it is assumed that the observed value y_t may be represented either as the product of the above components

i.e., $y_t = T \times S \times C \times I$ (multiplicative model)

or as the sum of components

$y_t = T + S + C + I$ (additive model)

Although the additive model facilitates easier calculation, the multiplicative model has been most widely used in analysis of time series.

a) Secular Trend

By secular trend we mean the smooth, regular, long-term changes in the series when observed over a period of time. Some series may exhibit an upward trend, some series a downward trend while some others may remain more or less constant over time. The upward trend of a series may be caused by factors such as increase in population and improvement in techniques of production. For example, the pattern of growth of many industries follows closely that of population growth of

the country. Again the advances in technology may give rise to upward movement of most of the economic time series. But not all time series will exhibit growth. Some may show decline while some others may show fluctuations. The time series of crude death rates of a country is likely to show a declining trend.

b) **Seasonal Variations**

The graphs of most of the time series reveal that a large number of fluctuations are imposed on the trend. By seasonal variation we mean the periodic movement in a time series where the period is not longer than one year. A periodic movement is that which repeats at regular intervals or periods of time. For example, the sales of cold drinks increase during summer and decrease during winter, sales of garments are maximum during some seasons of the year, say during May or festivals, the number of passengers carried by buses has a peak during office hours, the number of books borrowed from a library has a peak during some days of the week, etc. The factors which contribute to this type of fluctuations are the climatic changes of different seasons, customs and habits which people follow at different times.

c) **Cyclical Fluctuations**

By cyclical fluctuations we mean oscillatory movements of a time series, where the period of oscillation, called the cycle, is more than a year. It includes those factors leading to alternating periods of expansion and contraction that characterise most economic and business series. Sometimes these fluctuations are highly irregular with respect to their shape, amplitude and direction. But the phenomena they reflect — the periods of depression, recovery, boom and collapse — have been observed in virtually all time series dealing with business and economic data.

d) **Irregular Movement**

The irregular movement includes component all factors not classifiable elsewhere. Thus factors such as work stoppage, elections, wars, fire may affect a particular time series. This category of movement includes all types of variations not accounted for by secular trend, seasonal or cyclical fluctuations. Unfortunately, factors of these kinds are frequently indistinguishable from cyclical factors and as such in some discussions cyclical and irregular components are combined together.

11.2.2 Construction of Time Series : An Example

As an illustration we prepare a time series according to the multiplicative model. Table 11.1 presents trend, seasonal and cyclical-irregular components of a hypothetical series.

Table 11.1 : Hypothetical Time Series and its Components (Quarterly)

Year	Quarter	Series	Components		
			Trend	Seasonal	Cyclical-Irregular
		(y_t)	(T)	($100S$)	($100CI$)
1	I	79	80	120	82
	II	58	85	80	85
	III	84	90	92	102
	IV	107	95	108	105
2	I	130	100	120	108
	II	93	105	80	132
	III	121	110	92	120
	IV	161	115	108	130

3	I	216	120	120	150
	II	132	125	80	132
	III	150	130	92	125
	IV	163	135	108	112
4	I	176	140	120	105
	II	112	145	80	97
	III	128	150	92	93
	IV	142	155	108	85
5	I	134	160	120	70
	II	86	165	80	65
	III	94	170	92	60
	IV	104	175	108	55

In Table 11.1 the series is represented by a multiplicative model, such that $y_t = T \times S \times C \times I$.

Thus the observation 79 (of I quarter of 1st year) $= 80 \times \frac{120}{100} \times \frac{82}{100}$.

Similarly, 112 (of II quarter of 4th year) $= 145 \times \frac{80}{100} \times \frac{97}{100}$

Thus, each quarterly figure (y_t) is the product of the secular trend (T), the seasonal index (S), cyclical and the irregular component (CI). Such a synthetic composition looks very much like an actual time series and has encouraged use of the model as the basis for the analysis of time series data.

11.3 MEASUREMENT OF TREND

At times we are interested to know the trend movement in a time series. In such circumstances, we have to eliminate the effects of other components (seasonal, cyclical and irregular) from the series.

Two important methods of measuring trend are the Moving Average Method and the Method of Fitting Polynomials. In moving average method, secular trend is obtained by smoothing out fluctuations by the process of averaging. In the latter, a polynomial of suitable degree is chosen either for the original variables or for its transformed variable and its constants are determined by the method of least squares. The choice of the degree of the polynomial can be made by plotting the data on a graph paper where different scales, arithmetic, semi-logarithmic or double-logarithmic scales may be used. Measurement of trend is necessary for studying the behaviour of the time series and for forecasting the future.

11.3.1 Moving Average Method

This is a simple method of smoothing out fluctuations of a series by calculating a number of averages covering overlapping periods of the series. The first step consists in selecting proper period of the moving average. If the period chosen is 3 years, the moving averages are obtained by calculating a series of mean values of three consecutive values covering overlapping periods of the series. Denoting the original series by $y_1, y_2, y_3, \ldots\ldots$, the mean of the first three values, given by $(y_1 + y_2 + y_3)/3$, is placed at the midpoint of the period covering first three years. This is the first moving average value. The second moving average value is obtained

by calculating the mean of the values covering the period from second to fourth year. This is given by $(y_2 + y_3 + y_4)/3$ and is placed at the mid-point of the period covering second to fourth year. This process is repeated. It is clear that some of the values for the years at the beginning as well as at the end cannot be obtained by this method.

Two cases may be distinguished, viz., when the period of moving average is odd and when it is even. If the period is odd (for example, if the period is three years), the first moving average is placed at the second year, the second moving average is placed at the third year and so on. If, however, the period is even (for example four years), the moving average value fall between two consecutive years and 'centering' is necessary for getting trend values for various years.

As an illustration, let us consider a schematic representation for the calculation of centered 4-year moving averages. Here we will present two methods — the direct method (Table 11.2), as well as the short-cut method (Table 11.3).

Table 11.2 : Calculation of centered 4-year moving averages (Direct Method)

Year	y_t	4-year moving total	4-year moving average col 3 ÷ 4	Centered moving total	Centered 4-year moving average
(1)	(2)	(3)	(4)	(5)	(6)
1	y_1			—	—
2	y_2			—	—
		$y_1 + y_2 + y_3 + y_4 = T_1$	$T_1 / 4$		
3	y_3			$(T_1 + T_2)/4$	$(T_1 + T_2)/8$
		$y_2 + y_3 + y_4 + y_5 = T_2$	$T_2 / 4$		
4	y_4			$(T_2 + T_3)/4$	$(T_2 + T_3)/8$
		$y_3 + y_4 + y_5 + y_6 = T_3$	$T_3 / 4$		
5	y_5			$(T_3 + T_4)/4$	$(T_3 + T_4)/8$
		$y_4 + y_5 + y_6 + y_7 = T_4$	$T_4 / 4$		
6	y_6			—	—
7	y_7			—	—

In the above illustration, the period of moving averages is 4 years. Both in the direct and in the short-cut method col. 3 shows 4-year moving totals. The first value (T_1) is placed between the second and the third year, the second moving total (T_2) is placed between the third and the fourth year and so on. The centered 4-year moving averages are placed at the third year, fourth year,.... by taking a further 2 item moving average in the direct method (Table 11.2). In the short cut method (Table 11.3), the calculations of 4-year moving average are omitted (as shown in col. 4 of Table 11.2 in the direct method). Instead, the 2 item moving totals of the 4 year moving averages are obtained (col. 4 and 5).

Note that for a 4-year moving average, the procedure for centering leaves out $4/2 = 2$ years at the beginning and at the end of the series each.

Table 11.3 : Calculation of centered 4-year moving averages (Shortcut Method)

Year	y_t	4-year moving total (M.T.)	2 item moving total of col. 3	Centered 4-year moving average (M.A.), col 4 ÷ 8
(1)	(2)	(3)	(4)	(5)
1	y_1		—	—
2	y_2		—	—
		$y_1 + y_2 + y_3 + y_4 = T_1$		
3	y_3		$(T_1 + T_2)$	$(T_1 + T_2)/8$
		$y_2 + y_3 + y_4 + y_5 = T_2$		
4	y_4		$(T_2 + T_3)$	$(T_2 + T_3)/8$
		$y_3 + y_4 + y_5 + y_6 = T_3$		
5	y_5		$(T_3 + T_4)$	$(T_3 + T_4)/8$
		$y_4 + y_5 + y_6 + y_7 = T_4$		
6	y_6		—	—
7	y_7		—	—

11.3.2 Suitability of Moving Averages

Moving average method is simple to apply but the success of this method depends on the proper choice of the period. Moving average with a period exactly equal to or a multiple of the period of the cycle present in the series will completely eliminate the cyclical component and give an estimate of the trend. This method is flexible but some trend values at the beginning and at the end of the series have to be left out and their number increases with increase in the period of the moving average. Again as moving average assumes no law of change, the method cannot be used for forecasting future trend.

11.3.3 Examples of Moving Averages

Example 11.3.1: Calculate the three and five year moving averages of the following data:

Year	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981
Production ('000 tons)	18	19	20	22	20	19	22	24	25	24	25	26

Steps of Calculation

1) In Table 11.3.1 the figures in col. 3 are obtained as the sum of three consecutive values of col. 2. Thus the first moving total (M.T.) is 57 = 18 + 19 + 20 and is placed against 1971. The second moving total 61 = 19 + 20 + 22 is placed against 1972.

2) The three-year moving average(M.A.) in col. 4 is obtained by dividing the corresponding three-year moving total in col. 3 by 3, the period of the moving average. Thus 57 ÷ 3= 19, 61 ÷ 3 = 20.3, etc.

3) The five-year moving totals in col. 5 are obtained as the sum of five consecutive values in col. 2. Thus the first moving total against the year 1972 is 99 = 18 + 19 + 20 + 22 + 20.

4) The five-year moving average in col. 6 is obtained by dividing the corresponding five-year moving total in col. 5 by 5. Thus the moving average for 1975 is 107 ÷ 5 = 21.4.

Table 11.3.1: Calculation of (I) 3-year moving average (II) 5-year moving average

Year	Production	3-year M.T.	3-year M.A.	5-year M.T.	5-year M.A.
(1)	(2)	(3)	(4)	(5)	(6)
1970	18	—	—	—	—
1971	19	57	19.0	—	—
1972	20	61	20.3	99	19.8
1973	22	62	20.7	100	20.0
1974	20	61	20.3	103	20.6
1975	19	61	20.3	107	21.4
1976	22	65	21.7	110	22.0
1977	24	71	23.7	114	22.8
1978	25	73	24.3	120	24.0
1979	24	74	24.7	124	24.8
1980	25	75	25.0	—	—
1981	26	—	—	—	—

Note that for 3-year centered moving averages $\frac{3-1}{2}=1$ year, and for 5-year centered moving averages $\frac{5-1}{2}=2$ years, respectively, are left out both at the beginning and the end of the series.

Example 11.3.2: Compute trend values for the following time series using 4-yearly moving averages.

Year:	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988
Yield: (qntls.)	52	54	55	57	58	61	63	66	67	70

Solution:

Table 11.3.2(a): Calculation of 4-year moving average (Direct Method)

Year	Yield	4-year M.T.	4-year M.A.	2 item M.T. of col.4 (centered)	Centered 4-year M.A.
(1)	(2)	(3)	(4)	(5)	(6)
1979	52	—	—	—	—
1980	54	—	—	—	—
		218	54.50		
1981	55			110.50	55.250
		224	56.00		
1982	57			113.75	56.875
		231	57.75		
1983	58			117.50	58.750
		239	59.75		
1984	61			121.75	60.875
		248	62.00		
1985	63			126.25	63.125
		257	64.25		
1986	66			130.75	65.375
		266	66.50		
1987	67	—	—	—	—
1988	70	—	—	—	—

Table 11.3.2(b): Calculation of 4-year moving average (Shortcut Method)

Year	Yield	4-year M.T.	2-item M.T.	Centered 4-year M.A.
(1)	(2)	(3)	(4)	(5)
1979	52	—	—	—
1980	54	—	—	—
		218		
1981	55		442	55.250
		224		
1982	57		455	56.875
		231		
1983	58		470	58.750
		239		
1984	61		487	60.875
		248		
1985	63		505	63.125
		257		
1986	66		523	65.375
		266		
1987	67	—	—	—
1988	70	—	—	—

Steps of Calculation (Direct Method)

1) Col. 3 is the sum of four consecutive values in col. 2.
   Thus 52 + 54 + 55 + 57 = 218, 54 + 55 + 57 + 58 = 224.
2) Col. 4 = col. 3 ÷ 4. Thus 218 ÷ 4 = 54.5, 224 ÷ 4 = 56.
3) Col. 5 = Sum of two consecutive values in col. 4.
   Thus 54.5 + 56.0 = 110.5, 56.00 + 57.75 = 113.75.
4) Col. 6 = col. 5 ÷ 2. Thus 110.5 ÷ 2 = 55.25.

Steps of calculation (Shortcut method)

1) Col. 4 is the sum of two consecutive values in col. 3. Thus
   218 + 224 = 442, 224 + 231 = 455.
2) Col.5 = col.4 ÷ 8. Thus 442 ÷ 8 = 55.25.

Example 11.3.3

Find trend values for the following series using a 3-year weighted moving average with weights 1, 2, 1.

Year:	1	2	3	4	5	6
Value:	2	3	5	6	8	11

Solution:

Table 11.3.3: Calculation of 3-year weighted moving average

Year	Value	3-year weighted moving total (M.T.)	3-year weighted moving average (M.A.)
(1)	(2)	(3)	(4)
1	2	—	—
2	3	13	3.25
3	5	19	4.75
4	6	25	6.25
5	8	33	8.25
6	11		

Steps for Calculation

1) Col. 3 figures are the weighted moving totals of col. 2 figures with weights 1, 2, 1.

   Thus $1 \times 2 + 2 \times 3 + 1 \times 5 = 13$

   $1 \times 3 + 2 \times 5 + 1 \times 6 = 19$

2) Col. 4 = col. 3 ÷ (sum of weights, i.e., 4)

   Thus $13 \div 4 = 3.25$, $19 \div 4 = 4.75$

Example 11.3.4: Calculate the 4-quarter moving average for the following time series data.

Quarter	Year			
	1980	1981	1982	1983
1	62	66	72	79
2	58	60	67	74
3	72	74	80	88
4	60	64	69	77

Solution:

Year	Quarter	Value	4-quarter M.T.	Centered M.T.	4-quarter M.A.
(1)	(2)	(3)	(4)	(5)	(6)
1980	1	62	—	—	—
	2	58	—	—	—
			259		
	3	72		508	63.50
			256		
	4	60		514	64.25
			258		
1981	1	66		518	64.75
			260		
	2	60		524	65.50
			264		
	3	74		534	66.75
			270		
	4	64		547	68.38
			277		
1982	1	72		560	70.00
			283		
	2	67		571	71.38
			288		
	3	80		583	72.88
			295		
	4	69		597	74.63
			302		
1983	1	79		612	76.50
			310		
	2	74		628	78.50
			318		
	3	88	—	—	—
	4	77	—	—	—

Check Your Progress 1

1) The data given below give the index of industrial production from 1961 to 1970:

Year:									
1961	1962	1963	1964	1965	1966	1967	1968	1969	1970
Index of Production:									
109.2	119.8	129.7	140.8	153.8	153.2	152.6	163.0	175.3	184.3

Fit the trend line and predict the index of production for the year 1972 by 3-year moving average method.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

2) Depict the following figures of the net national product of India, at 1970-71 prices, in the form of a graph and superimpose the trend by computing the five year moving averages.

Year	NNP (Rs. 00 cr.)	Year	NNP (Rs. 00 cr.)	Year	NNP (Rs. 00 cr.)
1961	259	1969	335	1977	444
1962	269	1970	356	1978	482
1963	275	1971	378	1979	513
1964	292	1972	386	1980	487
1965	315	1973	382	1981	521
1966	301	1974	397	1982	550
1967	299	1975	400		
1968	324	1976	440		

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

11.4 METHOD OF FITTING POLYNOMIALS

Method of fitting polynomials is perhaps the best and the most objective method of determining trend. Here an appropriate type of polynomial is selected for trend and the constants appearing in the trend equation are determined on the basis of the given time series data. The choice of an appropriate polynomial is facilitated by a graphical representation of the data for which, apart from the usual arithmetic scales, semi-logarithmic or doubly-logarithmic scales may be used. If the plotted data show approximately a straight line tendency on an ordinary graph paper, the equation used is $Y = a + bx$ (straight line or first degree polynomial).

If they show a straight line on a semi-logarithmic graph paper, the equation is $\log Y = a + bx$, which is obtained by taking logarithm of $Y = A.B^x$ (exponential function). Note that $a = \log A$ and $b = \log B$.

Some times a second or a third degree polynomial may also be fitted.
$Y = a + bx + cx^2$ (second degree polynomial or parabola)
$Y = a + bx + cx^2 + dx^3$ (third degree polynomial)

The constants appearing in the above equations (such as $a, b, c,....$) are obtained by applying the principle of "least squares", as in regressions (see Unit 9). This states that the values of the constants will be such as to make the sum of squares of the deviations

$\sum(y - Y)^2$ minimum,

where y = observed value,

Y = expected value obtained from the trend equation of the type $Y = a + bx$ or

$Y = a + bx + cx^2$ etc., and the summation is taken over all the observations.

In the case of a straight line fitted by the method of "least squares", the constants a and b are determined from the following normal equations:

$$\sum y = na + b\sum x \text{ and}$$

$$\sum xy = a\sum x + b\sum x^2$$

where n is the number of years covered.

Similarly in the case of parabola or second degree polynomial the constants a, b and c are determined from the three normal equations

$$\sum y = na + b\sum x + c\sum x^2$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$$

$$\sum x^2y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

Rule for writing down the normal equations

To get the first normal equation, multiply each observation by coefficient of a in that equation and take sum over all the n observations.

Thus for straight line $y = a + bx$, as the coefficient of a is 1, the first normal equation is $\sum y = na + b\sum x$.

For the second normal equation, multiply each observation by the coefficient of b in that equation and take sum over all the n observations. In case of straight line, coefficient of b is x. So, the second normal equation is $\sum xy = a\sum x + b\sum x^2$

Now we will consider trend fitting for periods covering odd (Table 11.4) and even (Table 11.5) number of years taking the first degree polynomial.

Case I: Odd number of years ($n = 5$)

Table 11.4

Year	y	x	x^2	xy
(1)	(2)	(3)	(4)	(5)
1	y_1	-2	4	
2	y_2	-1	1	
3	y_3	0	0	
4	y_4	1	1	
5	y_5	2	4	
Total	Sy	0	10	Sxy

The normal equations are:

$$\sum y = 5a + b\sum x = 5a \qquad (\text{Since } \sum x = 0)$$

$$\sum xy = a\sum x + b\sum x^2 = 10b$$

$$\therefore\ a = \frac{\Sigma y}{5},\ b = \frac{\Sigma xy}{10}$$

where the origin ($x = 0$) is taken at the mid point of the interval covered by 5 years, i.e., at the third year and unit of time = 1 year. In real life situations the actual values of y_t's can be recorded. Hence Σx and Σxy will be known.

Case II: Even number of years ($n = 6$)

The constants a and b will be obtained from the following equations

$\Sigma y = 6a$
$\Sigma xy = 70b$
Here the origin ($x = 0$) will be in the middle of 3rd and 4th year and the unit of $x = 6$ months.

Table 11.5

Year	y	x	x^2	xy
(1)	(2)	(3)	(4)	(5)
1	y_1	-5	25	
2	y_2	-3	9	
3	y_3	-1	1	
4	y_4	1	1	
5	y_5	3	9	
6	y_6	5	25	
Total	Σy	0	70	Σxy

11.4.1 Suitability of Least Squares Method

Trend lines are used for description of the growth or decline of the time series and as an aid to the study of the long-term trend of the economy. The method of fitting polynomials completely eliminates personal bias and trend values for all

the given periods can be obtained. This is, however, not possible with moving average method. But the choice of the type of the polynomial curve is arbitrary and one cannot be sure whether a linear or parabolic curve will represent the trend best. The choice of the trend equation may itself lead to a bias. It is, however, possible to get some idea of the pattern of trend from the scatter diagram of the data.

11.4.2 Examples of Least Squares Method

Example: 11.4.1

Fit a straight line trend by the method of least squares to the following data:

Year:	1975	1976	1977	1978	1979	1980	1981
Production:	81	92	100	105	112	120	126

Estimate the production for 1982.

Solution:

Here the number of years is odd ($n = 7$). Let $y = a + bx$ be the equation of the straight line trend with origin ($x = 0$) at 1978 and one unit of $x = 1$ year.
The least squares normal equations are (see unit 9)

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

Thus, substituting the values of Σy, Σxy, Σx, and Σx^2 from the above table in the normal equations, we get
$7a = 736$, so $a = 105.1$
$28b = 203$, so $b = 7.21$

The trend equation is
$Y = 105.1 + 7.21x$, with origin at 1978 and unit of $x = 1$ year.
The value of x for 1982 would be 4.

Table 11.4.1: Fitting Straight Line Trend

Year	Production (y)	x	x^2	xy
(1)	(2)	(3)	(4)	(5)
1975	81	- 3	9	- 243
1976	92	- 2	4	- 184
1977	100	- 1	1	- 100
1978	105	0	0	0
1979	112	1	1	112
1980	120	2	4	240
1981	126	3	9	378
Total	**736**	**0**	**28**	**203**

Hence, using the trend equation the estimate for 1982 is $Y = 105.1 + 4 \times 7.21 = 133.94$.

Example: 11.4.2

Fit a straight line trend to the following time series data:

Year:	1970	1971	1972	1973	1974	1975
Profits: (Rs. lakhs)	3.1	3.3	3.6	3.2	3.7	3.9

Estimate the profit for 1976.

Solution:

Here the number of years is even ($n = 6$). Let $y = a + bx$ be the trend equation with origin
($x = 0$) mid-way between 1972 and 1973 and unit of $x = 6$ months.

The normal equations are

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

Table 11.4.2: Fitting Straight Line Trend

Year	Profit (y) (Rs. lakhs)	x	x^2	xy
(1)	(2)	(3)	(4)	(5)
1970	3.1	– 5	25	– 15.5
1971	3.3	– 3	9	– 9.9
1972	3.6	– 1	1	– 3.6
1973	3.2	1	1	0.0
1974	3.7	3	9	11.1
1975	3.9	5	25	19.5
Total	**20.8**	**0**	**70**	**4.8**

So, substituting the values of Σy, Σxy, Σx, and Σx^2 from the above table in the normal equations, we get

$6a = 20.8$, or $a = 3.47$
$70b = 4.8$, or $b = 0.07$

The trend equation is

$Y = 3.47 + 0.07x$, with origin at the middle of 1972 and 1973 and unit of $x =$ 6 months.

For 1976, x would be 7.

So, estimate for 1976 is

$Y = 3.47 + 0.07 \times 7 = 3.47 + 0.49 = 3.96$

Hence the estimated profit for 1976 is Rs. 3.96 lakhs.

Example: 11.4.3

The sales of a company (in thousands of rupees) for the years 1980 to 1986 are given in the following table. Fit an exponential trend ($Y = A.B^x$) and estimate the sales for 1987.

Year:	1980	1981	1982	1983	1984	1985	1986
Sales:	32	47	65	92	132	190	275

Solution:

Here the number of years is odd ($n = 7$). Taking log of both sides of the given equation, we can write $\log Y = \log A + x \log B$. Let $a = \log A$ and $b = \log B$. Thus, we can write

$\log Y = a + bx$.

Further, we take origin ($x = 0$) at 1983 and one unit of $x = 1$ year.
The least squares normal equations are:

$$\sum \log y = na + b\sum x$$

$$\sum x \log y = a\sum x + b\sum x^2$$

Table 11.4.3: Fitting Straight Line Trend

Year	Sales (y)	x	x^2	log$y$	x.log$y$
1980	32	- 3	9	1.5051	- 4.5153
1981	47	- 2	4	1.6721	- 3.3442
1982	65	- 1	1	1.8129	- 1.8129
1983	92	0	0	1.9638	0
1984	132	1	1	2.1206	2.1206
1985	190	2	4	2.2788	4.5576
1986	275	3	9	2.4398	7.3119
Total	**833**	**0**	**28**	**13.7931**	**4.3237**

So, substituting the values of $\Sigma \log y$, $\Sigma x.\log y$, Σx, and Σx^2 from the above table in the normal equations, we get

$7a = 13.7931$, or $a = 1.97$
$28b = 4.3237$, or $b = 0.154$.

Thus, the fitted function is $\log Y = 1.97 + 0.154x$ or $Y = \text{antilog}\ (1.97 + 0.154x)$.

For 1987, x would be 4

Thus, the estimated value for 1987 is
$Y = \text{antilog}\ (1.97 + 0.154 \times 4) = \text{antilog}\ 2.586 = 385.48$.

A case with even number of years can be attempted as in the fitting of a straight line (see Example: 11.4.2).

Example: 11.4.4

The following table shows the production of cement in India during 1982 to 1988. Fit a second degree polynomial to the data.

Year:	1982	1983	1984	1985	1986	1987	1988

Solution:

Here the number of years is odd ($n = 7$).

Let $y = a + bx + cx^2$ be the trend equation with origin ($x = 0$) at 1985 and unit of $x = 1$ year. The normal equations are:

$$\sum y = na + b\sum x + c\sum x^2$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$$

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

Table 11.4.4: Fitting Second Degree Polynomial

Year	y	x	x^2	x^3	x^4	xy	x^2y
1982	23.7	- 3	9	- 27	81	- 71.1	213.3
1983	27.1	- 2	4	- 8	16	- 54.2	108.4
1984	30.2	- 1	1	- 1	1	- 30.2	30.2
1985	33.1	0	0	0	0	0.0	0.0
1986	36.4	1	1	1	1	36.4	36.4
1987	39.3	2	4	8	16	78.6	157.2
1988	45.0	3	9	27	81	135.0	405.0
Total	**234.8**	**0**	**28**	**0**	**196**	**94.5**	**950.5**

Substituting the values from the table in the normal equations

$$7a + 28c = 234.8$$
$$28b = 94.5$$
$$28a + 196c = 950.5$$

Solving these three equations, simultaneously, we get

$a = 33$
$b = 3.37$
$c = 0.134$

Hence, the second degree polynomial is

$Y = 33 + 3.37x + 0.134x^2$,
with origin ($x = 0$) at 1985 and unit of $x = 1$ year.

Example: 11.4.5

Fit a second degree polynomial to the following data. Estimate the trend value for 1982.

Year	1976	1977	1978	1979	1980	1981
Annual Indian Imports (10^8 Rs.)	507	602	681	914	1255	1361

Solution:

Here the number of years is even ($n = 6$).

Let $y = a + bx + cx^2$ be the trend equation with origin ($x = 0$) mid-way between 1978 and 1979 and unit of $x = 6$ months. The normal equations are

$$\sum y = na + b\sum x + c\sum x^2$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$$

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

Table 11.4.5: Fitting Second Degree Polynomial

Year	y	x	x^2	x^3	x^4	xy	x^2y
1976	507	- 5	25	- 125	625	- 2535	12675
1977	602	- 3	9	- 27	81	- 1806	5418
1978	681	- 1	1	- 1	1	- 681	681
1979	914	1	1	1	1	914	914
1980	1255	3	9	27	81	3765	11295
1981	1361	5	25	125	625	6805	34025
Total	**5320**	**0**	**70**	**0**	**1414**	**6462**	**65008**

Substituting the values from the table in the normal equations

$$\begin{aligned} 6a + 70c &= 5320 \\ 70b &= 6462 \\ 70a + 1414c &= 65008 \end{aligned}$$

Solving these three equations, simultaneously, we get
$a = 829.2$, $b = 92.31$ and $c = 4.924$.

The second degree polynomial is
$Y = 829.2 + 92.31x + 4.924x^2$,
with origin ($x = 0$) mid-way between 1978 and 1979 and unit of $x = 6$ months.

For 1982, x would be 7.
Therefore, estimate for 1982 is
$Y = 829.2 + 92.31 \times 7 + 4.924 \times (7)^2$
$= 829.2 + 646.17 + 241.28 = 1716.65.$

11.5 MONTHLY OR QUARTERLY TREND VALUES FROM ANNUAL DATA

In a time series, annual data may be available in different forms such as (1) monthly or quarterly averages for each year, and (ii) annual totals.

If the trend equation is fitted to the monthly or quarterly data, there will be no difficulty in obtaining monthly or quarterly values. However, it is not advisable to fit a trend line by the method of least square to the monthly or quarterly data. The trend line is usually fitted to the annual data and then the trend values may be obtained for different months (or quarters). The method of obtaining monthly (or quarterly) trend equation is discussed below.

Let $y = a + bx$ be an annual trend equation. If we divide both sides of this equation by 12, we get a monthly average equation. Thus $\frac{y}{12} = \frac{a}{12} + \frac{b}{12}x$ is a monthly average equation. Denoting by $Y = \frac{y}{12}, A = \frac{a}{12}$ *and* $B = \frac{b}{12}$, we can write the

monthly average equation as $Y = A + Bx$. Here, Y is monthly average and B denotes change in monthly average per unit change in x, i.e., 1 year.

To get a monthly equation, we have to determine the corresponding rate of change of Y. Since B is the average monthly change in Y per year, $\frac{B}{12}$ will denote average change per month. Thus, the monthly equation can be written as $Y = A + \frac{B}{12}x$

or $Y = \frac{a}{12} + \frac{b}{144}x,$, where x denotes month rather than year.

Similarly, we can write

$Y = \frac{a}{4} + \frac{b}{4}x$ as the quarterly average equation, where a unit of x denotes one year, and

$Y = \frac{a}{4} + \frac{b}{16}x$ as the quarterly equation, where a unit of x denotes one quarter.

Shifting of Origin

We define a, in the equation $y = a + bx$, as the value of trend in the year of origin. Thus, with the shifting of origin the value of a changes. Let us assume that the year of origin (i.e., $x = 0$) is 1995 and we want to change it to 1998. We note that $x = 3$ for 1998, therefore, trend for 1998 = $a + 3b$. Treating this as constant term in the trend equation, $y = (a + 3b) + bx$ becomes the new trend equation with 1998 as origin.

Example: 11.5.1

The trend equation for certain production data is
$y = 150 + 24x$ (y = annual production in thousand tons and x = time with origin at 1978, unit of x = 1 year).
Estimate the trend value for May 1983.

Solution: The monthly trend equation is

$$Y = \frac{150}{12} + \frac{24}{144}x = 12.5 + 0.167x,$$

where Y = monthly production, unit of x = 1 month and origin at 1978, i.e., 30th June 1978.

To estimate the trend for May 1983, we substitute $x = 58.5$ in the above equation. Thus, we get $Y = 12.5 + 0.167 \times 58.5 = 22.25$ ('000 tons)

Example: 11.5.2

The trend equation fitted to quarterly average sales for 7 years is given by $y = 250 + 20x$ (unit of x = 1 year, origin = 30th June 1970). Estimate the trend value for the first quarter of 1973 (January-March).

Solution: Here the quarterly average refers to average per quarter for each year. The quarterly trend equation is given by

$Y = 250 + \frac{20}{4}x$, where Y = quarterly sales, x = 1 quarter and origin at 30th June

The interval between 30th June 1970 and the 1st quarter of 1973 is 10.5 quarters. Thus, to obtain the trend for 1st quarter of 1973, we substitute $x = 10.5$ in the above equation.

Hence, the required trend is $Y = 250 + 5 \times 10.5 = 302.5$.

Check Your Progress 2

1) Fit a straight line trend to the following data and show how to obtain the monthly trend values from the trend line fitted to the given time series. Obtain two such monthly values.

Year:	1970	1971	1972	1973	1974
Average Monthly Production: (in '000 tons)	38	40	41	45	47

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

2) The trend equation for certain production data is $y = 240 + 48x$ (y = annual production in tons, x = time with origin at 1985, unit of x = 1 year). Estimate the trend for October 1991.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

3) The trend equation fitted to quarterly average sales data is given by $y = 60 + 8x$ (unit of x = 1 year, origin = 30th June, 1988). Estimate the trend value for first quarter (Jan.-Mar.) of 1990.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

11.6 MEASUREMENT OF SEASONAL VARIATIONS

There are a number of methods for measuring seasonal variations in time series depending on how the other components such as cyclical, trend and irregular movements are present in it. For simplicity, we shall consider seasonal variations in monthly or quarterly data only, but the procedure for weekly or daily data will be similar. It may be mentioned here that annual data contains no seasonal variation. The application of the methods, to be discussed below, will give us 4 (or 12) numbers for quarterly (or monthly) data. These will be termed as seasonal indices and are normally expressed as percentages. A figure of a particular quarter (or month) indicates whether that quarter is above or below the normal quarter. For example, a value of 80 for a particular quarter indicates that the business for exports or sales (say) during that quarter is slack and it is below the normal quarter by 20%. We will consider only the multiplicative model for the measurement of seasonal variations. The main methods for the measurement of seasonal variation are given below.

1) Method of Simple Average
2) Ratio to Trend Method
3) Ratio to Moving Average Method

11.6.1 Method of Simple Average

This method assumes that the time series variable y is made up of only two components, viz., seasonal (S) and irregular or random component (I). Thus, we can write

$$y = S.I$$

When we take average of y values for each month or quarter of all the years, the irregular component gets eliminated and we are left with seasonal component. We will illustrate this method in Table 11.6.

Table 11.6: Illustration of the Method of Simple Average

	Quarters			
Year	I	II	III	IV
1	y_1	y_2	y_3	y_4
2	y_5	y_6	y_7	y_8
3	y_9	y_{10}	y_{11}	y_{12}
4	y_{13}	y_{14}	y_{15}	y_{16}
5	y_{17}	y_{18}	y_{19}	y_{20}
Total	T_1	T_2	T_3	T_4
Average	A_1	A_2	A_3	A_4
S. I.	s_1	s_2	s_3	s_4
S. I. (adjusted)	S_1	S_2	S_3	S_4

Explanatory notes:

a) $T_1 = y_1 + y_5 + y_9 + y_{13} + y_{17}$ is the total of y values of first quarter of each year. Similarly, T_2, T_3 and T_4, are the totals of second, third and fourth quarters of each year respectively.

b) A_i is the ith quarter average $= \frac{T_i}{n}$, where i = 1, 2, 3, 4 and n denotes the number of years.

c) G is defined as the grand average $= \frac{\Sigma A_i}{4}$.

d) $s_i = \frac{A_i}{G} \times 100$, i = 1, 2, 3, 4.

e) $s = s_1 + s_2 + s_3 + s_4$

f) S_1, S_2, S_3 and S_4 are the seasonal indices for the first, second, third and the fourth quarters respectively, where $S_i = \frac{s_i}{s} \times 400$, i = 1, 2, 3, 4. Note that the sum of these 4 index numbers must be equal to 400. Further, $S_i = s_i$ if s = 400.

g) For a time series with monthly data, the sum of 12 seasonal indices, one for each month, must be equal to 1200.

Example: 11.6.1

Compute seasonal indices for the following data by the Method of Simple Average.

	Quarters			
Year	I	II	III	IV
1972	72	68	80	70
1973	76	70	82	74
1974	74	66	84	80
1975	76	74	84	78
1976	78	74	86	82

Solution:

Table 11.6.1: Calculation of Seasonal Indices

	Quarters			
Year	I	II	III	IV
1972	72	68	80	70
1973	76	70	82	74
1974	74	66	84	80
1975	76	74	84	78
1976	78	74	86	82
Total	376	352	416	384
Average	75.2	70.4	83.2	76.8
S.I.	43	92.15	108.90	100.52

Notes:

Grand Average $G = \frac{A_1 + A_2 + A_3 + A_4}{4} = \frac{75.2 + 70.4 + 83.2 + 76.8}{4} = 76.4$

Seasonal Index for Quarter I, i.e., S_1 = 98.43

Seasonal Index for Quarter II, i.e., $S_2 = 92.15$
Seasonal Index for Quarter III, i.e., $S_3 = 108.90$
Seasonal Index for Quarter IV, i.e., $S_4 = 100.52$

Since the sum of these indices = 400, no adjustment is needed.

11.6.2 Ratio to Trend Method

If the data contain trend to an appreciable extent, we first find an appropriate trend equation to determine the trend for various quarters or months. Usually the monthly or quarterly trend values are obtained from the quarterly (or monthly) average trend equation. The trend is then eliminated by expressing the original y values as percentages of the corresponding trend values. This method is based upon the assumption that cyclical variations are either not marked or completely absent.

Symbolically, we can write

$$\frac{y}{T}\times 100 = \frac{TSI}{T}\times 100 = SI \times 100$$

From this, the irregular component can be eliminated by the use of Simple Average Method.

Example: 11.6.2

The following table shows the sales (in '000 Rs.) in a departmental store for five different years. Obtain the seasonal indices by Ratio to Trend Method.

	Quarters			
Year	I	II	III	IV
1980	502	1632	605	362
1981	526	1700	680	390
1982	556	1820	780	422
1983	590	1955	888	464
1984	632	2110	1002	515

Solution:

Let us fit a straight line trend to the data on quarterly averages. The trend equation fitted to quarterly averages be $y = a + bx$, where y denotes quarterly average of the year and the unit of $x = 1$ year. The table below has been constructed from the given data by computing the averages of 4 quarters of each year.

Table 11.6.2 (a): Fitting Linear Trend

Year	y	x	x^2	xy
1980	775	- 2	4	- 1550
1981	824	- 1	1	- 824
1982	894	0	0	0
1983	974	1	1	974
1984	1065	2	4	2130
	4532	0	10	730

The normal equations are

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

Substituting values from the above table into the normal equations for linear trend, we get

$5a = 4532$ or $a = 906.4$
$10b = 730$ or $b = 73$.

Thus, the quarterly average trend equation is $T = 906.4 + 73x$
(origin: 1982, unit of x = 1 year)

Note that we are using T instead of Y (used earlier in fitting of trends)
From this, we can write the quarterly trend equation as

$$T = 906.4 + \frac{73}{4}.x = 906.4 + 18.25x$$

(origin: 30th June, 1982, unit of x = 1 quarter)

Shifting the origin to 3rd quarter (mid-point of third quarter) of 1982, the quarterly trend equation becomes

$$T = 906.4 + 18.25\ (x + 0.5) = 906.4 + 9.125 + 18.25x$$
$$= 915.125 + 18.25x.$$

Putting appropriate values of x, we can get the trend values (T) for various quarters. The next step is to express the original values (y) as percentage of trend, i.e., $(y \div T) \times 100$, giving "trend ratios". The trend values along with the trend ratios are shown in Table 11.6.2(b).

Table 11.6.2 (b): Calculation of Trend Ratios

Year	Quarter	x	$T = 915.525 + 18.25x$	y	$(y + T) \times 100$
1980	I	-10	733.0	502	68
	II	-9	751.3	1632	217
	III	-8	769.5	605	79
	IV	-7	787.8	362	46
1981	I	-6	806.0	526	65
	II	-5	824.3	1700	206
	III	-4	842.5	680	81
	IV	-3	860.8	390	45
1982	I	-2	879.0	556	63
	II	-1	897.3	1820	203
	III	0	915.5	780	85
	IV	1	933.8	422	45
1983	I	2	952.0	590	62
	II	3	970.3	1955	201
	III	4	988.5	888	90
	IV	5	1006.8	464	46
1984	I	6	1025.0	632	62
	II	7	1043.3	2110	202
	III	8	1061.5	1002	94
	IV	9	1079.8	515	48

The trend ratios are now arranged by quarters and the seasonal indices are calculated by the method of simple averages.

Table 11.6.2 (c): Calculation of Seasonal Indices

Year	Quarters			
	I	II	III	IV
1980	68	217	79	46
1981	65	206	81	45
1982	63	203	85	45
1983	62	201	90	46
1984	62	202	94	48
Total	320	1029	429	230
Average	64.0	205.8	85.8	46.0
S.I.	63.74	209.98	85.46	45.82

11.6.3 Ratio to Moving Average Method

This method is used when the time series data is composed of all the four components. We know that moving averages, with period equal to the the periodic variations, completely eliminates those variations. Thus, if we take 4 period moving average (M) of quarterly data (or 12 period moving averages of monthly data), the seasonal variations (and some irregular movements) are eliminated from the original (y) values. Further, by expressing y as a percentage to moving average, i.e., $(y \div M) \times 100$, we get values consisting of seasonal and irregular components. The seasonal component is, then, isolated from irregular component by the use of the Method of Simple Average.

Symbolically, we can write

$$\frac{y}{M} \times 100 = \frac{TCSI}{TCI'} = SI''$$

Example: 11.6.3

Use the Ratio to Moving Average Method to calculate seasonal indices for the following data.

Year	Summer	Monsoon	Autumn	Winter
1989	30	81	62	119
1990	33	104	86	171
1991	42	153	99	221
1992	56	172	129	235
1993	67	201	136	302

Solution:

Table 11.6.3: Calculation of seasonal indices by Ratio to Moving Average Method

Year/Qtr.		y	4-pd. M.T.	Centered Totals	4-pd. M.A. (M)	$(y + M) \times 100$
1989	Sum	30	—	—	—	—
	Mon	81	—	—	—	—
			292			
	Aut	62		587	73.38	84.50
			295			
	Win	119		613	76.63	155.30
			318			
1990	Sum	33		660	82.50	40.00
			342			
	Mon	104		736	92.00	113.04
			394			
	Aut	86		797	99.63	86.32
			403			
	Win	171		855	106.88	160.00
			452			
1991	Sum	42		917	114.63	36.64
			465			
	Mon	153		980	122.50	124.90
			515			
	Aut	99		1044	130.50	75.86
			529			
	Win	221		1077	134.63	164.16
			548			
1992	Sum	56		1126	140.75	39.79
			578			
	Mon	172		1170	146.25	117.61
			592			
	Aut	129		1195	149.38	86.36
			603			
	Win	235		1235	154.38	152.23
			632			
1993	Sum	67		1271	158.88	42.17
			639			
	Mon	201		1345	168.13	119.55
			706			
	Aut	136	—	—	—	—
	Win	302	—	—	—	—

The moving ratios are now arranged by quarters and the seasonal indices are calculated by the method of simple averages.

	Quarters			
Year	Summer	Monsoon	Autumn	Winter
1989	—	—	84.50	155.30
1990	40.00	113.04	86.32	160.00
1991	36.64	124.90	75.86	164.16
1992	39.79	117.61	86.36	152.23
1993	42.17	119.55	—	—
Total	158.60	475.10	333.04	631.69
Average	39.65	118.78	83.26	157.92
Seasonal Index	36.69	118.89	83.34	158.08

Check Your Progress 3

1) The following data represent the production of finished steel tins for the years 1972 to 1975:

Production of Finished Steel Tins (000 tons)

Year	Jan.	Feb.	Mar.	Apr.	May.	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
1972	420	414	502	365	368	332	390	396	429	417	422	496
1973	491	466	516	337	342	360	409	402	372	391	394	446
1974	463	465	478	310	325	406	415	437	438	445	430	416
1975	502	487	536	404	418	429	489	492	475	456	476	476

Compute the seasonal indices by the Method of Simple Averages.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

2) The following table gives the production of steel in India from 1972 to 1975 (in'000 tons) over the different quarters.

Year	1st quarter	2nd quarter	3rd quarter	4th quarter
1972	1336	1065	1215	1335
1973	1463	1039	1183	1161
1974	1306	1041	1290	1321
1975	1525	1251	1456	1408

Obtain seasonal indices by the method of Ratio to Trend, assuming a linear trend.

3) Given the following quarterly sales figures in thousands of rupees for the years 1986 to 1989. Find the specific seasonals by the method of moving averages.

	Quarters			
Year	**I**	**II**	**III**	**IV**
1986	290	280	285	310
1987	320	305	310	330
1988	340	321	320	340
1989	370	360	362	380

4) The seasonal indices for the sales of garments of a particular type in a certain shop are given below:

Quarter	*Seasonal Index*
Jan-Mar	97
Apr-Jun	85
Jul-Sep	83
Oct-Dec	135

If the total sales in the first quarter of a year is Rs. 15,000, determine how much worth of garments of this type should be kept in stock by the shop owner to meet the demand for each of the other three quarters of the year?

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

11.7 LET US SUM UP

Time series is a set of observations on a variable recorded over time — usually at equal intervals. The change in the variable concerned can be explained to some extent on the basis of components of time series. These components are trend, seasonal variations, cyclical fluctuations and random movements. The observed value of the variable may be represented either as the product of the aforesaid components (multiplicative model) or as the sum of the components (additive model).

Trend can be measured by the method of moving averages and fitting equations by the method of least squares. Once we estimate the trend, we can predict future values and also estimate the monthly or quarterly values from the annual trend.

Seasonal variation can be estimated by three methods: Method of Simple Averages, Ratio to Trend Method and Ratio to Moving Average Method. The method of simple average is used to average out the irregular component. The ratio to trend method can be used if cyclical variations are supposed to be absent while the ratio to moving average method is recommended when the time series variable is composed of all the four components.

11.8 KEY WORDS

Cyclical Variations : Oscillatory movements of a time series where the period of oscillation, called cycle, is more than a year.

Irregular Movement : The random movement of time series which is not explained by other components. In this sense it is a residual of other components.

Method of Least Squares : When a polynomial function is fitted to the time series, the method of least squares requires that the parameters of the function should be so chosen as to make the sum of squares of the deviations between actual observations and expected values to be the minimum.

Moving Average Method : Taking a suitable period of moving average, (say 3 years), the moving averages are calculated as series of mean values of three (in this case) consecutive years. The average obtained is entered against the middle year.

Seasonal Variation : Periodical movement where the period is not longer than one year.

Secular Trend : The smooth, regular and long-term movement of a time series over a period of time. Trend may be upward or rising, downward or declining or it may remain more or less constant over time.

11.9 SOME USEFUL BOOKS

Nagar, A.L. and R.K. Das, 1989: *Basic Statistics*, Oxford University Press, Delhi.

Goon, A.M., M.K. Gupta and B. Dasgupta, 1987: *Basic Statistics*, The World Press Pvt. Ltd., Calcutta.

11.10 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Read Example 11.3.1 and answer.
2) Read Example 11.3.1 and answer.

Check Your Progress 2

1) $y = 42.2 + 2.3x$, origin: 1972, unit of $x = 1$ year (monthly average equation)
   $y = 42.3 + 0.19x$, origin: July, 1972, unit of $x = 1$ month (monthly equation)
   Estimate for March 1971 ($x = -16$) = 39.23
   Estimate for September 1973 ($x = 14$) = 44.98.
2) 45.17 tons.
3) 73.

Check Your Progress 3

1) 109.57, 107.00, 118.69, 82.71, 84.87, 89.19, 99.47, 100.87, 100.11, 99.82,100.58, 107.12
2) 112.27, 86.62, 100.21, 100.90
3) 104.2, 97.9, 96.5, 101.4
4) Rs. 13144; 12835; 20876.

UNIT 12 VITAL STATISTICS

Structure

12.0 OBJECTIVES

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

- explain the sources of data in vital statistics;
- calculate various vital rates;
- explain the procedure of construction of life table; and
- appreciate the applications and limitations of life tables.

12.1 INTRODUCTION

Vital statistics is mainly concerned with the factors contributing to population growth. Some of these factors are birth rates, death rates, expectancy of life, and migration. As you go through this Unit you will be in a position to appreciate the importance and applications of vital statistics in economics. The main objectives of this Unit are to introduce some of the basic concepts of vital statistics, the data sources, how to measure various ratios, and what are the applications of these ratios in

projecting the population, calculating life expectancy, uses in actuarial science, etc.

12.2 DATA SOURCES

The data for vital statistics are usually collected through the following four methods; viz., Registration, Census, Survey and Sample Registration System. We discuss these methods below:

i) **Registration Method:** This method consists of continuous and permanent recording of births, deaths, marriages, migration, etc. Many countries including India have made registration of births and deaths compulsory under the law. The registration office issues a certificate on registration of a birth or death. Although, the registration method is simple and effective it suffers from the problem that all the births and deaths that are occurring are not registered. This is because the law has not been enforced strictly, particularly, in rural India.

ii) **Census:** Almost all the countries in the world conduct census periodically to enumerate their population. The census provide the vital statistics information such as age, sex, marital status, education level, occupation, religion, etc. However, these information pertain to the census years only (once in ten years). The data for the years other than census years are not available.

iii) **Survey:** Surveys are conducted in areas where the registration method is not effective or not functioning properly. The surveys enable us to have required vital statistics of these regions.

iv) **Sample Registration System (SRS):** This is a large scale demographic survey conducted in India for providing reliable annual estimates of birth rate, death rate and other fertility and mortality indicators at the national and sub-national levels. The field investigation consists of continuous enumeration of births and deaths by a resident part-time enumerator, generally a teacher followed by an independent survey every six months by an official. The data obtained through these operations are matched. The unmatched and partially matched events are re-verified in the field and thereafter an unduplicated count of births and deaths is obtained. The SRS was initiated by the Office of the Registrar General, India on a pilot basis in a few selected states in 1964-65. It became fully operational during 1969-70 covering about 3700 sample units. Thereafter the sample size has been periodically increased. The frame was recently updated based on 1991 Census data.

12.3 USES OF VITAL STATISTICS

Vital statistics are useful in many spheres of human activity. Some important uses of vital statistics are as follows.

1) The vital statistics help us in understanding how the population profile of a country or a region within the country is changing. The profile of the population is in terms of age, sex, religion, births, deaths, migration, marriages, etc. These statistics help us in predicting the future population structure of a country or region.

2) The estimation of population trends and projections help the policy planners and administrators for better planning and evaluation of economic and social development programmes. For example, the transportation infrastructure is influenced directly by the number of people in an area.

3) The mortality statistics help us to improve the health conditions of the communities. For example, statistics on communicable diseases help the health authorities to improve the sanitary conditions of the area affected and improve medical facilities.

4) The actuarial science including life insurance is based on vital statistics. You will learn more about it in section 12.4 of this Unit.

12.4 MEASUREMENT OF POPULATION

The total population of a country is usually expressed at a particular point of time. For example, the latest census in India was conducted in 2001 and the total population of the country was found to be on 31.3.2001. The total population measured at a census is usually considered as accurate. As you may be aware, the census in India are conducted once in 10 years. The inter-censal data are estimated using the following methods.

12.4.1 Linear Interpolation Method

The estimation of total population for a given inter-censal year can be calculated using the following formula:

$$P_t = P_0 + \frac{n}{N}(P_1 - P_0) \qquad \ldots(12.1)$$

Where, P_t is estimated population at a given inter-censal year t
P_0 is population in the previous census
P_1 is population in the succeeding census
n is the number of years between the given year and the previous census year
N is the number of years between the two census years

The above method provides us a good estimate at a constant rate between the inter-censal years.

Example 12.1: The total population of India in 1991 census was 846 million and in 2001 census was 1027 million. Calculate the total population of India in 1996.

Here, $P_0 = 846$, $P_1 = 1027$, N= 10, n= 5

Therefore, $P_{1996} = 846 + \frac{5}{10}(1027 - 846) = 936.5$ million.

The limitation of the above method is that we can estimate the population only for the years between two census years. We cannot have the estimates for the future years.

12.4.2 Using Compound Growth Rate Formula

Normally it was observed that the population growth takes place in a geometrical progression. In case the base year population and the population compound growth rate (between the base census year and succeeding census year) are known, we can estimate the total population for a given year using the following formula.

$$P_t = P_0(1+r)^n$$

where, r is the compound growth rate (between the base census year and succeeding census year)
n is the number of years from the base year (usually previous census year)

P_0 is the base year (usually previous census year)
P_t is the estimated population at a given year t from the base year

Example 12.2: The population of a small town in 1991 was 50500. The compound growth rate of the population of that town between 1991 and 2001 was 0.025. Estimate the population of the town for the year 2005 (assuming that the population growth rate will be the same beyond 2001).

Here, we are given $P_0 = 50500$, $r = 0.025$, and $n = 14$ (since 2005–1991 = 14)

Therefore, $P_{2005} = 50500\ (1 + 0.025)^{14} = 71355$

12.4.3 Natural Increase and Net Migration Method

You know that the census gives us the total population. Similarly, the total number of births, deaths, and migration statistics are obtained from registrars. The population of an area increase by:

i) Natural increase (that is, total number of births – total number of deaths)

ii) Net migration (that is, total number of people immigrated to the area – total number of people emigrated out of the area).

The population for a given period is calculated using the following formula.

$P_t = P_0 + (B–D) + (I–E)$

Where, P_t is the estimated population at a given year *t* from the base year (usually previous census year)

P_0 is the base year (usually previous census year)

B and D are the total number of births and deaths respectively during the base year to the year t.

I and E are the total number of immigrants and emigrants respectively during the base year to the year *t*.

Example 12.3 : The population of a small town in 2001 census was 22000. From 2001 to 2003 the number of births, deaths, immigrants and emigrants are 800, 150, 2500 and 1500 respectively. Find the total population of the town in 2003.

Here, $P_0 = 22000$, $B = 800$, $D = 150$, $I = 2500$, $E = 1500$

Therefore, $P_{2003} = 22000 + (800–150) + (2500–1500)$
$= 23650$

Check Your Progress 1

The following table gives information on mid-year total population of India and annual compound growth rates of population.

Year	Population (Crores)	Period	Compound growth rate (%)
1950	36.99	1950-60	1.9
1960	44.59	1960-70	2.2
1970	55.50	1970-80	2.1
1980	68.70	1980-90	2.0
1990	84.17	1990-2000	1.8
2000	100.27	—	

Source: US Census Bureau: IDB Summary Demographic Data for India

Note that the compound growth rates are in terms of percentage. Divide it by 100 to get the required r. For example, for the period 1950-60 the compound growth rate is 1.9%. Therefore, r = 1.9/100 = 0.019.

On the basis of the above table answer the questions below:

1) Find the mid-year population for the following years using linear interpolation method.

Year	Mid-year population
1954	
1966	
1973	
1985	
1998	

2) Find the mid-year population for the following years using compound growth rate method.

Year	Mid-year population
1954	
1966	
1973	
1985	
1998	

3) Compare the above two methods and draw your conclusions.

.....

.....

.....

.....

4) Briefly explain different data sources for vital statistics.

.....

.....

.....

.....

5) What are the important uses of vital statistics?

.....

.....

.....

.....

12.5 VITAL RATES

Normally, the data on vital statistics are available in the form of number of births, number of deaths, etc. In order to have a meaningful utility of these data we

generally transform this data into some vital rates or ratios. Number of births or deaths in a year per 100 persons is usually low and would result in small fractions. The changes in these ratios would also be not perceptible. In order to avoid this problem, vital rates are expressed on the basis of per thousand persons. In this section you will learn some important vital rates.

12.5.1 Crude Birth Rate

The crude birth rate is defined as the number of births per 1000 population in a specific community or region. To calculate the crude birth rate we use the following formula :

$$\text{Crude birth rate} = \frac{\text{Annual Number births (in a community or region)}}{\text{Annual mid year population (of the community or region)}} \times 1000$$

The crude birth rate tells us at what rate the births are occuring in a region or community.

Example 12.4 : The mid-year population and number of births occurred of a tribal community in Madhya Pradesh in 1995 are 40,000 and 1200 respectively. Find the crude birth rate.

Here, we have 1995 mid-year population = 40000 and the 1995 number of births = 1200

$$\text{Crude birth rate} = \frac{1200}{40000} \times 1000$$

$$= 30 \text{ per } 1000 \text{ persons per annum}$$

12.5.2 Crude Death Rate

The crude death rate is defined as the number of deaths per 1000 population in a specific age group or sex group or community or region. To calculate the crude death rate we use the following formula.

$$\text{Crude Death Rate} = \frac{\text{Annual number of deaths (in a specific age group or sex group or community or region)}}{\text{Annual mid year population (of the specific age group or sex group or community or region)}} \times 1000$$

The crude death rate tells us at what rate the deaths are happening in a age group, sex group or region or community.

Example 12.5: The mid-year population and the number of deaths registered in 2001 for a town in Maharashtra among females are 25000 and 245 respectively. Find the crude death rate.

Here, we have 2001 mid year female population = 25000 and the number of deaths in 2001 =245.

$$\text{Crude death rate (females)} = \frac{245}{25000} \times 1000$$

$$= 9.8 \text{ per } 1000 \text{ persons per annum among females.}$$

12.5.3 Crude Rate of Natural Increase

The crude rate of natural increase is defined as the rate at which a population increases in a given year because of a surplus of births over deaths expressed as per 1000 persons.

The annual natural increase is measured as: annual number of births – annual number of deaths.

The formula for calculating the crude rate of natural increase is:

$$\text{Crude rate of natural increase} = \frac{\text{Annual natural increase}}{\text{Annual mid year population}} \times 1000$$

$$= \text{crude birth rate} - \text{crude death rate}$$

The crude rate of natural increase for a given year tells us at what rate natural increase has added the population over the year.

Example 12.6 : In India the crude birth rate and crude death rates in 1997 are 27.2 and 8.9 respectively. Find the crude rate of natural increase.

Crude rate of natural increase = 27.2 – 8.9
= 18.3 per 1000 per annum

12.5.4 Rate of Net Migration

Migration is defined as the movement of people across a specific boundary of a region for the purpose of establishing a new or semi permanent residence. Immigrants are those who have come into the region and eminigrants are those who have moved out of the region.

The annual net migration is defined as: Annual number of immigrants – annual number of emigrants

The formula for calculating the annual rate of net migration is:

$$\text{Annual rate of net migration} = \frac{\text{Annual net migration}}{\text{Annual mid year population}} \times 1000$$

The annual rate of net migration tells us at what rate the net migration has added to the population over the course of the year.

Example 12.7: In 2002 for a region the annual number of immigrants, emigrants, and mid–year population are given as 6500, 5200 and 66700 respectively. Find the annual rate of net migration.

Here, we have the number of immigrants = 6500
the number of emigrants = 5200
mid–year population = 66700

Annual net migration = 6500 – 5200 = 1300

$$\text{Annual rate of net migration} = \frac{1300}{66700} \times 1000$$

$$= 19.7 \text{ per } 1000 \text{ per annum}$$

12.5.5 Rate of Total Increase

The total increase in population is measured as:

Annual natural increase + annual net migration.

$$\text{Rate of total increase} = \frac{\text{Annual total increase}}{\text{Annual mid year population}} \times 1000$$

= crude rate of natural increase + rate of net migration

The rate of total increase for a given year tells us the rate at which the population has increased over the year.

Example 12.8 : The annual natural increase, annual net migration, and annual mid-year population in 1998 for a region are recorded as 1500, 500 and 50000 respectively. Find the rate of total increase.

Here, we have

Annual natural increase = 1500
Annual net migration = 500
Mid year population = 50000

Annual total increase = 1500 + 500 = 2000

$$\text{Rate of total increase} = \frac{2000}{50000} \times 1000$$

$$= 40 \text{ per } 1000 \text{ per annum.}$$

12.5.6 Infant Mortality Rate

The infant mortality rate is defined as the number of deaths of infants (less than one year old) per 1000 live births in a given year. The formula to calculate the infant mortality rate is given as:

$$\text{Infant mortality rate} = \frac{\text{Annual infant deaths (of males or females or total)}}{\text{Annual live births (of males or females or total)}} \times 1000$$

The infant mortality rate tells us for a given year the chances of a birth failing to survive one year life. The infant mortality rates can be calculated separately for males and females.

Example 12.9 : In 1997 for a small town the total number of live births and infant deaths among females are recorded as 3000 and 25 respectively. Find the infant mortality rate among females.

Here, we have

Annual live female births = 3000
Annual infant deaths = 25

$$\text{Infant mortality rate} = \frac{15}{3000} \times 1000$$

$$= 8.33 \text{ per } 1000 \text{ per annum}$$

Check Your Progress 2

The provisional estimates of crude birth rate, crude death rate, natural growth rate and infant mortality rate in India for the year 1997 are as follows:

Vital Rate	Total	Rural	Urban
Birth rate	27.2	28.9	21.5
Death rate	8.9	9.6	6.5
Natural growth rate	18.3	19.2	15.0
Infant mortality rate	71	77	45

Source: Sample Registration System Bulletin, October 1998

Observe that all the vital rates are higher in rural areas than in urban areas. Write one most significant reason for each of the following:

1) The birth rate in rural areas is high because

.....

.....

.....

.....

2) The death rate in urban areas is low because

.....

.....

.....

.....

3) The infant mortality rate in rural areas is high because

.....

.....

.....

.....

12.6 LIFE TABLES

The life expectancy is defined as the average number of additional years a person could expect to live if the current mortality trends continue for the rest of that person's life. A life table is a tabular display of life expectancy and probability of dying at each age or age group for a given population, according to the age-specific death rates prevailing at that time.

The life table gives us an organised complete picture of a population's mortality.

We can explain it with an example. We start with a group (usually called cohort) of 100,000 female births and estimate the number which will survive to every age or age group, if they are subjected to the existing mortality conditions. We can say, for example, that out of 100,000 initial female births 95,000 will reach the age of 15 years, 92,500 the age of 25 years and so on, and the mean age at which all 100,000 will die is 72 years.

The construction of a life table is a simple process. It involves the following steps that are repeated for each age group.

i) **Age interval (x to x + n):** The period of life between two exact ages. The exact age (x) represents the lower limit of each age interval, beginning with 0 and incrementing to 1, 5, 10, 15 and so on upto 100+ (which is an open interval). The first and second age groups are usually '<1' and '1-4' and the last age group is '100+' whereas the rest of the age groups are of equal size, like '5-9', '10-14', 15-19,..... '95-99'.

ii) **Width of the age interval (n_x)** : This is the number of years in the age interval (x to x+n). Usually, the first value is 1 (interval <1), the second 4 (1-4) and the remaining values are 5 (5-9, 10-14,.....95-99) with the exception of the last value which is again taken as 1 (100+).

iii) **Number of deaths recorded in the age interval (d_x):** This column presents the number of persons dying in that age group during the year corresponding to the life table.

iv) **Number of persons in the age interval (P_x):** This column indicates the number of persons in the age interval during the year corresponding to the life table.

v) **Separation factor ($_na_x$):** This represents the average number of years lived by those who die between ages x and x+n. Although, it is necessary in calculations, this factor is not typically presented as a column of the life table. Each person living in the interval (x to x+n) has lived x completed years plus some fraction of the interval (x to x+n). In a complete life table, a value 0.5 (that is, half of one year) is valid from the age of 5 years. For a simpler calculation, it is assumed that those who die in the 5 year age intervals of a life table live on average 2.5 years. However, remember that the value of the fraction depends on the mortality pattern over the entire interval and not the mortality rate for any single year. In addition, since a large portion of infant deaths occur in the first few weeks of life, this value is much smaller in the <1 and 1-4 age groups.

Similarly, the death rates in the last three groups (namely, 91-94, 95-99, and 100+) are very high. Therefore, the value of the separation factor is small in the age group 91-94 and 95-99. In the last age group (100+) since the death is certain we have taken the separation factor as 1.

Calculation of the separation factor is easy if the date of birth and date of death are available. For the purpose of constructing a life table the separation factor will be given in the table. When they are not, values from model life tables, such as those tabulated by Coale and Demney shown in Table 12.1 can be utilised for and the rest are taken as 0.5 years for every year in the group interval (that is 2.5 in year interval).

Table 12.1 : Separation factors for ages 0 and 1-4

		Separation factor for age <1			Separation factor for ages 1-4		
	Zones	Men	Women	Both sexes	Men	Women	Both sexes
Infant Mortality Rate >0.100	North (1)	0.33	0.35	0.3500	1.558	1.570	1.5700
	East (2)	0.29	0.31	0.3100	1.313	1.324	1.3240
	South (3)	0.33	0.35	0.3500	1.240	1.239	1.2390
	West (4)	0.33	0.35	0.3500	1.352	1.361	1.3610
Infant Mortality Rate <0.100	North (1)	0.0425	0.05	0.0500	1.859	1.733	1.7330
	East (2)	0.0025	0.01	0.0100	1.614	1.487	1.4870
	South (3)	0.0425	0.05	0.0500	1.541	1.402	1.4020
	West (4)	0.0425	0.05	0.0500	1.653	1.524	1.5240

Source: Coale, Ansley J. and Demeny P. (1966) *Regional Model Life Tables and Stable Populations,* Princeton University Press.

Notes: (1) Iceland, Norway and Switzerland; (2) Austria, Czechoslovakia, North-central Italy, Poland and Hungary; (3) South Italy, Portugal and Spain; (4) Rest of the World.

vi) **Central mortality ($_nM_x$) :** This column results from dividing the number of deaths in the age interval x to x+n (column d_x) by the number of people in this age group (column P_x).

$$_nM_x = \frac{d_x}{P_x}$$

vii) **Probability of dying between the ages x and x+n ($_nq_x$):** The probabilities of dying are calculated based on the age-specific mortality rates for each age group. This column is interpreted as the probability of dying between the ages for the person who has survived upto age x. For the last age group of the table, where death is unavoidable, the probability of dying is 1. For other age groups, the calculation is more complicated. The formula for calculation is given below:

$$_nq_x = \frac{n_x \times {}_nM_x}{1 + (n_x - {}_na_x) \times {}_nM_x}$$

viii) **Probability of survival between the ages x and x+n ($_np_x$):** It is interpreted as the probability of a person who reaches age x to reach the exact age x+n alive. The formula for calculation is given below:

$$_np_x = 1 - {}_nq_x$$

Since it is $1 - {}_nq_x$, we normally do not show this as a separate column in the life table.

ix) **Survivors to exact age x ($_nl_x$):** This column indicates the number of persons living in the age group x to x+n out of the initial cohort which is usually taken as 100,000.

x) **Deaths between the exact ages x and x+n ($_nd_x$) :** This is calculated using the following formula.

$$_nd_x = {}_nl_x \times {}_nq_x$$

xi) **Number of years lived by the total of the cohort of 100,000 births in the interval x to x+n ($_nL_x$):** Each member of the cohort who survives the interval x to x+n contributes n years to L, while each member who dies in the interval x and x+n contributes the average number of years lived by those

who die in this period (that is, the separation factor of deaths ${}_na_x$). ${}_nL_x$ is calculated using the following formula.

$${}_nL_x = n_x \times {}_nl_{x+n} + {}_na_x \times {}_nd_x$$

where, ${}_nl_{x+n} = {}_nl_x \times {}_np_x$ or

${}_nl_{x+n} = {}_nl_x - {}_nd_x$

xiii) **Total years lived after exact age x (${}_nT_x$)**: This number is essential for the calculation of life expectancy. It indicates the total number of years lived by the survivor ${}_nl_x$ between the anniversary x and the extinction of the whole generation. The value of the first row of ${}_nT_x$ is the total number of years lived by the cohort until death of its last component.

${}_nT_x$ = Sum of ${}_nL_x$ (from last row of ${}_nL_x$ to the current row of ${}_nL_x$)

xiii) **Life expectancy at age x (${}_ne_x$) :** Among all the indicators provided by the life table, the most widely used is the life expectancy (${}_ne_x$) which represents the average number of years lived by a generation of newborns under given mortality conditions.

Table 12.2 below provides the basic information required for construction of a life table. The data pertains to Indian females in 2000. Let us construct the life table.

Table 12.2 : Basic Information

Age (x)	Number of deaths(d_x)	Number of people(P_x)	n_x	Separation factor (${}_na_x$)
<1	788471	11655599	1	0.1
1-4	430704	44728827	4	1.6
5-9	137870	54725561	5	2.5
10-14	69159	52128201	5	2.5
15-19	100055	48475620	5	2.5
20-24	119360	42745630	5	2.5
25-29	116085	39848328	5	2.5
30-34	109226	35983667	5	2.5
35-39	102540	31934500	5	2.5
40-44	124848	27744053	5	2.5
45-49	150315	23125487	5	2.5
50-54	172910	19212249	5	2.5
55-59	226553	16258203	5	2.5
60-64	288036	13715985	5	2.5
65-69	354148	10813430	5	2.5
70-74	368365	7554310	5	2.5
75-79	335430	4615527	5	2.5
80-84	252665	2332329	5	2.5
85-89	130278	817817	5	2.5
90-94	42440	183658	5	2
95-99	8199	24796	5	2
100+	915	1961	1	+
All ages	4428572	488625738		

Source : World Health Organisation.

Using the formulas given earlier the following life table is constructed.

Table 12.3 : Life Table

Age	n_x	$_na_x$	$_nM_x$	$_nq_x$	$_np_x$	$_nl_x$	$_nd_x$	$_nL_x$	$_nT_x$	$_ne_x$
<1	1	0.1	0.06765	0.06377	0.93623	100000	6376.52	94261.1	6268416	62.6842
1 to 4	4	1.6	0.00963	0.03765	0.96235	93623.5	3524.63	366035	6174155	65.9467
5 to 9	5	2.5	0.00252	0.01252	0.98748	90098.8	1127.83	447675	5808120	64.4639
10 to 14	5	2.5	0.00133	0.00661	0.99339	88971	588.245	443384	5360446	60.2493
15 to 19	5	2.5	0.00206	0.01027	0.98973	88382.8	907.437	439645	4917061	55.6337
20 to 24	5	2.5	0.00279	0.01386	0.98614	87475.3	1212.83	434345	4477416	51.1849
25 to 29	5	2.5	0.00291	0.01446	0.98554	86262.5	1247.4	428194	4043071	46.8694
30 to 34	5	2.5	0.00304	0.01506	0.98494	85015.1	1280.57	421874	3614877	42.5204
35 to 39	5	2.5	0.00321	0.01593	0.98407	83734.5	1333.63	415339	3193003	38.1324
40 to 44	5	2.5	0.0045	0.02225	0.97775	82400.9	1833.39	407421	2777665	33.7092
45 to 49	5	2.5	0.0065	0.03198	0.96802	80567.5	2576.57	396396	2370243	29.4193
50 to 54	5	2.5	0.009	0.04401	0.95599	77990.9	3432.36	381374	1973847	25.3087
55 to 59	5	2.5	0.01393	0.06733	0.93267	74558.6	5019.87	360243	1592474	21.3587
60 to 64	5	2.5	0.021	0.09976	0.90024	69538.7	6937.35	330350	1232230	17.7201
65 to 69	5	2.5	0.03275	0.15136	0.84864	62601.3	9475.39	289318	901880	14.4067
70 to 74	5	2.5	0.04876	0.21732	0.78268	53126	11545.3	236767	612562	11.5304
75 to 79	5	2.5	0.07267	0.3075	0.6925	41580.7	12786.2	175938	375795	9.03774
80 to 84	5	2.5	0.10833	0.42622	0.57378	28794.5	12272.9	113290	199857	6.94081
85 to 89	5	2.5	0.1593	0.56964	0.43036	16521.6	9411.37	59079.6	86567	5.23963
90 to 94	5	2	0.23108	0.68236	0.31764	7110.23	4851.74	20996	27487.4	3.8659
95 to 99	5	2	0.33067	0.82999	0.17001	2258.5	1874.53	5668.9	6491.49	2.87425
100+	1	+	0.46678	1	0	383.969	383.969	822.593	822.593	2.14235

Life expectancy always decreased from the first row of the table to the last row, with the exception of the second row and sometimes the third row (age group/ 5-9), which can be greater than the first row (age group/<1) in countries with high infant mortality. It is generally observed that for a given population, life expectancy is greater in women than in men and overall life expectancy should be approximately between the two. However, in countries where the maternal mortality is high and general living conditions of women are worse, life expectancy among women is lower than men.

12.7 APPLICATIONS OF LIFE TABLES

The life table is widely used in demographic, actuarial, social and health studies. The principal objective of a life table is to calculate life expectancy at birth and at other ages. However, life table provides interesting demographic data which have various applications. In this section you will learn the applications of the life table.

12.7.1 Calculation of Probability of Surviving and Dying

While constructing life table you have learned that $_nq_x$ tells us the *probability of dying* between the two ages (x, x+n) for the person who has survived upto age x. For example, let us consider the row corresponding to age group 30-34 years in Table 12.3. The probability of dying (females) between the ages 30 to 34 year of age who has survived upto 30 years of age is 0.01506 ($_nq_x$). That means out

of every 100,000 Indian females who have survived the age of 30 years, 1506 (= 100,000 × 0.01506) will die between the age 30 and 34 years. Secondly, ${}_nP_x$ tells us the *probability of living* between the two ages (x, 30-34 years/x+n) for the persons who has survived upto age x. For the age group the probability of survival is 1–0.01506 = 0.98494 (${}_nq_x$). That means out of every 100,000 Indian females who have survived the age of 30 years, 98494 will survive in the age group 30-34 years.

Thirdly, we can calculate the *probability at birth* of a person dying between ages 0-4 years. This is given by the number of original births dying (${}_nd_x$) between the ages 0-4 years, divided by the number of original births (usually 100000). In our example, ${}_nd_x$ = 1281 and the probability is 0.01281(= 1281/100000). This probability tells us that on an average out of every 100,000 female births in India (subject to mortality in 2000), 1281 females will die between the ages 0-4 years.

12.7.2 Uses in Actuarial Science

The life tables have significant applications in actuarial science especially in the field of life assurance. Life tables form the basis for determining the rates of premiums necessary to various amount of life assurance. Life tables provide the actuarial science with a sound foundation, converting the insurance business from a mere gambling in the human lives to the ability to offer well calculated safeguard in the event of death.

Actually, the calculations involved in the fixation of premium amounts in life assurance are very complex, but the underlying principles are simple. Let us consider a few examples.

Example 12.10: According to mortality conditions in India for the year 2000, what annual premium would an Indian female have to pay on a whole life policy worth Rs.100,000 if this life was assured at birth, assuming that the assurance office earns no income on its funds?

Let the premium be Rs. x per annum. Since a female on the average can be expected to live 62.7 years, over her life time she will have paid Rs.x × 62.7 in premiums. This will have to be equal to the value of the policy Rs.100000. Therefore, Rs. x × 62.7=100000 and x = 100000/62.7 = Rs.1594.90.

Example 12.11: In the above example if the policy was taken at the age of 25 years, then find the annual premium.

If the policy was taken at age 25 then the total premiums paid will be Rs. x × 46.9 for 46.9 years expectation of life at 25 years age. Then the annual premium must be x = 100000/46.9 = Rs.2132.20.

Example 12.12: In example 12.10 if the policy is an endowment policy, taken at 30 years of age and payable upto 50 years of age or prior deaths. What is the annual premium to be paid?

If the policy is an endowment policy, taken out say at 30 years payable upto 50 years or prior death, we should proceed on a some what different method. From Table 12.3 we know that 850155 (${}_nl_x$) survivors at age 30 live 1600529.5 (415338.6+407421+396396.1+381373.8 (${}_nL_x$) years between ages of 30 and 50. Consequently, on the average a total of Rs. x × (1600529.5/85015) premiums

will be collected and hence the annual premium must be Rs.100000 ÷ 18.83 = Rs. 5310.67.

12.7.3 Other Applications of Life Tables

Apart from its uses in insurance life tables is useful in undertaking comparative analysis of mortality conditions across countries or regions. We discuss some of the applications of life tables below:

i) **Calculation of mortality due to specific causes:** Life tables for different groups of population like sex (male/female), age distribution (different age groups), religion, region are calculated for comparisons. The mortality statistics may prompt us to find the specific causes of deaths in different groups of population.

ii) **Comparison of mortality conditions:** The life expectancy at birth and other ages are the best indices of mortality. These indices considerably vary from place to place and time to time. Over time, in most countries, the life expectancy has increased steadily due to improved health facilities. As you have already learned the female life expectancy is higher than male expectancy except where the female maternal mortality is high. Table 12.4 below explains the life expectancy for males and females in some selected countries.

Table 12.4: Life Expectancy at Birth: Selected Countries - 1999

	Males years	Females years
Australia(a)	76.6	82.0
Canada	75.9	81.4
China	68.3	72.5
France	74.5	82.3
Germany	74.3	80.6
Hong Kong (SAR of China)	76.7	82.2
India	**62.4**	**63.3**
Indonesia	63.9	67.7
Italy	75.2	81.6
Japan	77.3	84.1
Korea, Republic of	70.9	78.4
Netherlands	75.3	80.7
New Zealand	74.8	80.1
Papua New Guinea	55.4	57.3
Singapore	75.2	79.6
United Kingdom	75.0	80.0
United States of America	73.9	79.7

(a) Reference period for Australia is 1998-2000.

Source: Deaths, Australia (3302.0); United Nations Development Programme 2000.

iii) **Population projections:** The life tables have also been used in preparation of population projections by age and sex. That is, in estimating what the size of the population will be at some future date.

12.7.4 Limitations of Life Tables

Life tables are based on demographic data collected from sources such as census and SRS. Therefore, life table estimates have all the disadvantages of any statistical

measure based on population censuses and vital records. Data on ages and mortality registration may be incomplete or biased. Infant mortality weighs heavily on life expectancy, which means that under-reporting of this indicator, a habitual fact in many countries, can have an important effect on the results of the tables. Also, important differences in specific age/sex groups with high mortality may be overlooked, since this would have little effect on the overall life expectancy.

Constructing life tables for small populations, at the local or sub-regional level, is generally not recommended, since migratory movements affect the population structure more than at the regional or national levels. In these cases, a very small number of deaths can be obtained, which may produce imprecise calculations of the table's columns.

Check Your Progress 3

Read the life Table given in Table 12.3 in the text. Now interpret the values in the life table by answering the following questions.

1) What is the probability of a female child in India in 2000 would die before reaching 1 year of age?

.....

.....

2) How many years is a female born in 2000 in India expected to live?

.....

.....

3) What is the probability of dying of a female between 15 and 20 years of age?

.....

.....

4) What is the mortality rate between 15 and 20 years of age?

.....

.....

5) What is the probability that a female reaching 15 years of age reaches 20?

.....

.....

6) How many additional years is a female between 15 and 20 years of age in 2000 in India expected to live?

.....

.....

12.8 LET US SUM UP

Vital statistics is mainly concerned with births and deaths. The reliability of vital rates depends upon the effectiveness of the registration system. Incompleteness of registration of births and deaths, in spite of the laws, has made it difficult to give a correct picture of birth and death rates.

Life tables present the mortality and survival experience of a whole population and

permit evaluation of its affect on specific groups and over different periods. It is a simple instrument that is easily constructed with data collected routinely.

It is important to keep in mind that life tables are constructed based on population data from censuses and mortality registries. Therefore, the quality of the data affects the validity of the life table.

12.9 KEY WORDS

Cohort : A group of people sharing a common demographic experience who are observed through time. For example, the birth cohort of 2003 is the people born in that year.

Rate of Natural increase : The rate at which a population increases in a given year because of surplus of births over deaths expressed as per 1000 of the population. This excludes migration.

Migration : The movement of people across a specified boundary for the purpose of establishing a new or semi permanent residence.

Mid-year population : It is the average of end-year estimates. For example, the mid-year population of 2003 will be the average of the population as on 31st December 2002 and 31st December 2003.

Infant mortality rate : The number of deaths of infants below one year old per 1000 live births in a given year.

Life expectancy : The average number of additional years a person could expect to live if the current mortality trends continue for the rest of that persons' life. Frequently we use life expectancy at birth.

Actuarial Science : Actuarial Science is concerned with the application of mathematical and statistical methods to finance and insurance, particularly where this relates to the assessment of risks in the long term. In actuarial science we compute the insurance risks and premiums.

Natural Increase : The surplus of births over deaths in a population in a given period of time.

12.10 SOME USEFUL BOOKS

Agarwal, B.L. (1988), *Basic Statistics*, Wiley Eastern Limited, New Delhi.

Ansari, M.A., Gupta, O.P. and Chaudhary, S.C. (1980), *Applied Statistics*, Kedarnath Ram Nath & Co., Meerut.

Benjamin, B.(1959), *Elements of Vital Statistics*, George Allen and Unwin, London.

Chang, C.J.(1980) *Life Tables and Mortality Analysis*, Geneva: World Health Organisation

Carmel, P.H. and M. Polasek, (1986), *Applied Statistics for Economists*, Khosla Publishing House, Delhi.

12.11 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1)

Year	Mid-year population
1950	40.03
1966	51.14
1973	59.46
1985	76.44
1998	97.05

2)

Year	Mid-year population
1954	39.88
1966	50.81
1973	59.07
1985	75.85
1998	97.08

3) The estimated mid-year populations using linear interpolation method are slightly less than the method using compound growth rate method.

4) See Section 12.2 and answer.

5) See Section 12.3 and answer..

Check Your Progress 2

1) The birth rate in rural areas is high because of the lack of awareness among the people on the family planning methods and its need.

2) The death rate in urban areas is low because of the improved health facilities in towns and cities.

3) The infant mortality rate in rural areas is high because of the lack of health facilities in rural areas and malnutrition among mothers.

Check Your Progress 3

1) The probability for a female under 1 to die in India in 2000 ($_1q_0$) is 0.06377.

2) The number of years that a female born in 2000 in India expected to live ($_1e_0$) is 62.68 years.

3) The probability of a female dying between 15 and 20 years of age group ($_5q_{15}$) is 0.01027.

4) The mortality rate between 15 and 20 years age group ($_5M_{15}$) is 0.00206.

5) The probability that a female in the 15-19 age group reaches 20-24 years age group ($_5q_{15}$) is 0.98973.

6) The life expectancy of a female in the age group 15-20 years in 2000 in India ($_{15}e_{15}$) is 55.63 years.

UNIT 13 ELEMENTARY PROBABILITY

Structure

13.0 OBJECTIVES

After going through this unit you should be in a position to:

- explain the concept of probability;
- explain the laws of probability including Bayes' Theorem; and
- solve numerical problems in probability and mathematical expectations.

13.1 INTRODUCTION

Often we make statements like

It may rain tomorrow.

There is a fair chance that Team A wins the match.

It is unlikely that Mr. X becomes the President.

Mr. Y probably met Mr. Z.

We can see that all these statements are characterised by an element of uncertainty. For example, in the first statement, we are not sure that it will rain tomorrow. Similarly in the last statement, we exactly don't know whether Mr. Y met Mr. Z or not. Any statement, in which there is an element of uncertainty about the occurrence of some event, is called a *probability statement*. Thus, all the above statements are probability statements.

Suppose in connection with the first probability statement, one asks

How much is the chance of rain tomorrow?

We may have to come out with an answer like

There is a 75% chance of rain tomorrow.

Now, we are not only making a probability statement but also giving a relative measure for the degree of certainty (and implicitly that of uncertainty) associated with the event of rain tomorrow. Thus, the degree of certainty of rain is given a relative value of 75% and at the same time the uncertainty associated with rain is implicitly given a relative value of 25%. Suppose, we have a scale to measure the degree of certainty. In this scale the degree of certainty will vary from 0% to 100%. This scale also implicitly measures the degree of uncertainty. Thus, if one is sure about the non-occurrence of an event, it will have 0% chance or certainty. At the same time, it will have 100% non-chance or uncertainty. Similarly, if one is sure about the occurrence of an event, it will have 100% chance or certainty and 0% non-chance or uncertainty. However, we should note here that both the statements of 100% occurrence (so, 0% non-occurrence) and 0% occurrence (so, 100% non-occurrence) of an event are statements without any element of uncertainty and hence, in a strict sense, are not probability statements.

The relative measure of the degree of certainty with which an event can occur can be termed as the *probability* of the event. Conventionally, this degree is measured relative to 1. Thus, a 30% chance of the occurrence of an event will have a probability of 0.3. Similarly, a 75% chance of rain tomorrow can be stated as:

The probability of rain tomorrow is 0.75.

A relevant question is: How can we obtain the probability of an event? In the next section, we shall deal with this question.

13.2 DEFINITION OF PROBABILITY

It is clear that the problem of obtaining the probability is essentially a problem of measuring the degree of certainty of occurrence or non-occurrence of an event. Mathematicians have had different perceptions about the degree of certainty of an event and accordingly various definitions of probability have been given. These definitions suggest procedures for obtaining the probability of an event. In this Unit, we shall consider three such definitions. They are: (i) the *classical or mathematical definition*, (ii) the *relative frequency definition* or the *statistical definition,* and (iii) the *modern definition* or the *axiomatic approach to probability*.

13.2.1 Classical Definition

The classical definition of probability is based upon certain concepts. Let us first understand these.

a) *Statistical experiment* — An experiment having more than one possible outcome is called a statistical experiment. A statistical experiment is also known as a *trial*. Thus, tossing a coin to see whether it results in a head or a tail is a trial. Certain statements implying more than one possible situation can also be termed as trials. For example, all the four statements given at the beginning of Section 13.1 are trials or statistical experiments.

b) *Event* — A possible outcome of a trial is called an event. Thus, head is an event that may result from tossing a coin. Similarly, the occurrence of five or the occurrence of an odd number is a possible event of the trial of throwing a dice. The latter example indicates that an event may consist of one or more possible outcomes of an experiment. The event of getting an odd number, in fact, consists of three possible outcomes of rolling a dice. We should note

that an event consisting of only one possible outcome is often called an *elementary event.*

c) *Exhaustive events* — A set of events is said to be exhaustive, if it includes all possible outcomes of a trial. For example, the tossing of a coin can result in either a head or a tail and nothing else. Thus, the set {head, tail} is an exhaustive set of events associated with the trial of tossing a coin. Consider another example. We know, a dice has six sides and each side has dots that vary from 1 to 6. When the dice is thrown, it shows up a side with a given number of dots. If the occurrence of a side with a given number of dots is an event, the set {1, 2, 3, 4, 5, 6}, where each number represents the number of dots on a side of a dice, is an exhaustive set of events. The number of elements in an exhaustive set of events is known as the number of *cases* of the trial.

d) *Favourable events* — Such *cases* that support the occurrence of an event are said to be cases favourable to that event. Suppose a dice is thrown to see if it shows up a face with an even number of dots. In this trial, the sides with 2, 4 and 6 dots are all cases that favour the occurrence of the event of a side with an even number of dots.

e) *Equally likely events* — If in a trial, the chance of the occurrence of any possible event is the same, the events are said to be equally likely. Suppose a dice is thrown. If we feel that each of the six sides has an equal chance to show up, the possible six events are then equally likely.

f) *Mutually exclusive events* — If in a trial, the occurrence of an event rules out the simultaneous occurrence of any other possible event, the events are said to be mutually exclusive. We know, the toss of a coin results in either a head or a tail. Thus, the events of a head and a tail in the toss of a coin are mutually exclusive.

Now we are in a position to understand the classical definition of probability. The definition states:

> If a trial can result in n mutually exclusive, equally likely and exhaustive outcomes and out of which m outcomes are favourable to an event A, the probability of A, denoted by $P(A)$, is then $P(A) = \frac{m}{n}$.

It is clear that if A is an impossible event, that is, none of the n possible outcomes favours the occurrence of the event A, we have $m = 0$. The probability of A in that case is $P(A) = \frac{m}{n} = \frac{0}{n} = 0$

On the other hand, if A is a certain event, that is, all of the n possible outcomes favour the occurrence of the event A, we have $m = n$. The probability of A in that case is $P(A) = \frac{m}{n} = \frac{n}{n} = 1$

We shall now consider some applications of the classical definition for the computation of probability.

Example 13.1 What is the probability of getting a head in the toss of a fair coin?

We know that the toss of a coin can result in either a head or a tail and

nothing else. Hence, these two events constitutes a set of exhaustive events. If the toss results in head, it simultaneously cannot result in tail and vice versa. Thus, the two events are mutually exclusive. Finally, if we have nothing to suspect the behaviour of the coin, both head and tail have the same chance of occurrence in the toss. So, the two events are equally likely. In this way, when all the conditions of the classical definition are satisfied, we can proceed with the solution of the problem.

The number of exhaustive outcomes $n = 2$ (head and tail)

The number of outcomes favouring the required event (head) $m = 1$

If $P(H)$ is the probability of the occurrence of head, then $P(H) = \frac{m}{n} = \frac{1}{2}$

From now onwards, we shall proceed straight with the solution. However, you should satisfy yourselves that all the conditions of the classical definition are satisfied.

Example 13.2 A fair dice is thrown. What is the probability that either 1 or 6 will show up?

A dice has six faces with 1, 2, 3, 4, 5 and 6 dots printed on them and any one of these faces will show up when the dice is thrown. Thus the number of exhaustive outcomes $n = 6$. Now, the face with 1 dot favours the required event and the face with 6 dots also satisfies the required event. So, the number of outcomes favouring the required event is $m = 2$. If $P(1 \text{ or } 6)$ is the probability of either 1 or 6 then

$$P(1 \text{ or } 6) = \frac{m}{n} = \frac{2}{6} = \frac{1}{3}$$

Example 13.3 An unbiased coin is tossed twice. What is the probability of getting a head at least once?

If H stands for head and T stands for tail, there are four possible outcomes. They are (H, H) (H, T) (T, H) (T, T)

Thus, $n = 4$ here and the number of outcomes favourable to the required event of at least one head, $m = 3$. Hence

$$P(\textit{at least one head}) = \frac{m}{n} = \frac{3}{4}$$

Limitations of the Classical Definition

The classical definition has some serious drawbacks. They are:

a) The classical definition can be applied only if various outcomes of the trials are equally likely or equally probable. But in practice the outcomes need not be always equally likely. For example, if a coin is biased in favour of head, the classical definition fails to give the probability of a head or a tail.

b) The classical definition is valid for a finite number of outcomes of a trial. It fails when the number of outcomes becomes infinity. In fact, even in the case of a finite number of outcomes, it may not be practically feasible to enumerate all the cases.

c) The classical definition is 'circular' in the sense that while defining probability, the definition uses the term 'equally likely' which pre-supposes the knowledge of the concept of probability.

Check Your Progress 1

1) A box contains 4 white balls and 6 red balls. A ball is drawn without looking into the box. What is the probability that it is a white ball?

.....

.....

2) A six-faced dice is thrown. What is the probability of getting an even number?

.....

.....

3) A coin is tossed twice. What is the probability of getting either two heads or two tails?

.....

.....

4) A card is drawn from a pack of 52 cards. What is the probability of not getting a king?

.....

.....

13.2.2 Relative Frequency or Statistical Definition

Another definition that has often been used is the relative frequency definition of probability. If we repeat a trial and observe the occurrence of an event, we shall see that as the number of trials is progressively increased, the ratio of the number of times a particular event occurs to the total number of trials tends to stabilise at a particular value. Now, the number of times an event occurs is its frequency and when this frequency is divided by the total number of trials, we get the relative frequency of the event. Thus in other words, when the number of trials becomes sufficiently large, the relative frequency of an event tends to a limit. According to the relative frequency definition, this limiting value is the probability of the event under consideration. Suppose, we repeat the experiment of tossing a coin and observe the number of times head occurs. We shall find that as we increase the number of tosses from say, 10 to 100 to 1000 to 10000 and so on, the relative frequency of head will gradually stabilise at $\frac{1}{2}$. Thus, the probability of head in the toss of a fair coin is $\frac{1}{2}$.

Mathematically, if n is the total number of trials out of which, an event A occurs m times, the probability of A

$$P(A) = \lim_{n \to \infty} \frac{m}{n}$$

13.2.3 Axiomatic Approach to Probability

The main limitation of the classical as well as the relative frequency definitions is that these definitions preclude a rigorous mathematical treatment of the subject of probability. This limitation has been taken care of in the modern definition.

Before presenting the modern or axiomatic definition, it is necessary to grasp the following concepts.

a) *Sample Space* — It is the set of all possible (or exhaustive) outcomes of a trial. The sample space of a trial can be denoted by S and is given by $S = \{e_1, e_2, \ldots, e_n\}$, where, $e_1, e_2, \ldots, e_n$ are n elementary events.

If the trial consists of tossing a coin, then the sample space will be $S = \{H, T\}$. Similarly, when a dice is rolled, the sample space is given by $S = \{1, 2, 3, 4, 5, 6\}$.

The elements of a sample space can also be *ordered pairs*. For example, the sample space of the simultaneous toss of two coins is $S = \{ (H, H), (H, T), (T, H), (T, T) \}$. Further, a sample space can be *finite* or *infinite* depending upon whether it consists of finite or infinite number of elements.

b) *Event* — An event is any *subset* of sample space. For example, if A denotes an event that an odd number appears on a dice, then $A = \{1, 3, 5\}$. Again, the event of the occurrence of at least one head when two coins are tossed simultaneously is given by, say, $B = \{ (H, H), (H, T), (T, H) \}$.

c) *Occurrence of an Event* — An event is said to have occurred whenever the outcome of a trial belongs to the relevant event-subset. Thus, when we roll a dice and get 1, we say that event A has occurred. Based upon this, we can say that the sample space of a trial is certain to occur.

According to the modern definition, the probability of an event A, denoted by $P(A)$ is a *real valued set function* that associates a real value $P(A)$ corresponding to any *subset* A of the sample space S.

In order that $P(A)$ is the probability of an event A, it must satisfy the following restrictions. These restrictions are also known as the *axioms of probability theory*.

1) The probability of an event A, in a sample space S, is a non-negative real number less than or equal to unity, i.e., $0 \le P(A) \le 1$.

2) The probability of an event, that is certain to occur, is unity. Since S is certain to occur, this implies that $P(S) = 1$.

3) If A_1, A_2 and A_3 are mutually exclusive events in a sample space S, then

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3).$$

The above relation can be generalised to any number of events.

We should note that in the sample space $S = \{e_1, e_2, \ldots, e_n\}$, the elementary events, $e_1, e_2, \ldots, e_n$ are mutually exclusive. Thus on the basis of the third axiom, we can say that

$$P(S) = \sum_{i=1}^{n} P(e_i)$$

Verbally, the probability of sample space is equal to the sum of the probabilities of its elementary events. In a similar way, we can state that the probability of an event is equal to the sum of its elementary events. Thus, to find the probability of an event, we must know the probability of the occurrence of its elementary events. This can be done in any of the following three ways.

1) In the absence of any information regarding the occurrence of various elementary events, it is reasonable assume them to be equally likely. Thus, we can assign equal probability to each of the elementary events.

Since $P(S) = \sum_{i=1}^{n} P(e_i) = 1$, therefore

$$P(e_i) = \frac{1}{n} \quad (i = 1,2,...,n)$$

If there are m elementary events in the event A, then

$$P(A) = \left(\frac{1}{n} + \frac{1}{n} + ... + \frac{1}{n}\right) m \text{ times}$$

$$= \frac{m}{n} = \frac{\text{number of elements in } A}{\text{number of elements in } S}$$

This is nothing but the classical approach to probability.

2) Another way of assigning probability to various elementary events is to perform an experiment a large number of times. The relative frequencies of various elementary events can be taken as their respective probabilities if n is sufficiently large. This method uses the statistical definition of probability discussed before.

3) The probabilities of various elementary events can also be assigned by the person performing the experiment, on the basis of her experience and expectations. For example, one may ask you to specify the probability of rain today. You may be tempted to specify a higher value, say 0.8, if the day falls in the rainy season and so on. This approach to probability is particularly useful to managers engaged in taking various business decisions.

In practice we encounter many situations which involve a combination of events. In such situations we need to combine the probabilities of these events. In this context, we discuss two important laws of probability below.

13.3 PROBABILITY LAWS

Before considering various probability laws, let us be familiar with certain notations.

a) If A and B are two events, then $P(A \cup B)$ or $P(A+B)$ denotes the probability that either A occurs or B occurs or both occur simultaneously. It can also be interpreted as the probability of the occurrence of at least one of the two events A and B. The symbol $\cup$ above represents 'union' between two events. (Read A$\cup$B as 'A union B').

b) $P(A \cap B)$ or $P(AB)$ denotes the probability of the simultaneous occurrence of both A and B. (Read A$\cap$B as 'A intersection B').

c) $P(A/B)$ denotes the *conditional probability* of the occurrence of A given that B has already occurred.

13.3.1 Addition Law

This law states that the probability of the occurrence of at least one of the two events (i.e., either A or B or both) is equal to the probability of A plus the probability of B minus the probability of both A and B.

Using notations, we can say

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad ...(13.1)$$

We discuss now the modification to (13.1) in certain special cases.

a) Mutually Exclusive Events: Now suppose, A and B are mutually exclusive,

that is, the occurrence of A precludes the occurrence of B and vice versa; then, the two events cannot occur simultaneously and $P(A \cap B) = 0$. Thus, for two mutually exclusive events A and B, probability of the occurrence of either A or B is equal to the probability of A plus the probability of B.

$$P(A \cup B) = P(A) + P(B)$$

b) *Exhaustive Events*: Again, if A and B are the only possible outcomes of a trial (i.e., A and B are exhaustive events), then the occurrence of either A or B is a certainty. We know that the probability of a event that is certain to occur is 1. Hence, in that case

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 1$$

or $P(A \cup B) = P(A) + P(B) = 1$ (when A and B are mutually exclusive)

c) *Complementary events:* Suppose A is a possible outcome of some trial. It is then clear that the trial either results in the occurrence of A or the non-occurrence of A. Thus, A and 'not A' exhaust all the possible outcomes of any trial. So, if $\overline{A}$ (Read as 'A-bar') denotes 'not A', we have

$$P(A) + P(\overline{A}) = 1$$

or $P(\overline{A}) = 1 - P(A)$

Here, $\overline{A}$ is called *complement* to the event A. Thus, the sum of probabilities of any event and its complement is always equal to 1.

13.3.2 Multiplication Law

This law states that the probability of the simultaneous occurrence of the two events A and B is equal to the product of

i) the probability of A and the conditional probability of B given that A has already occurred

or

ii) the probability of B and the conditional probability of A given that B has already occurred.

In symbols,

$$P(A \cap B) = P(A) \cdot P(B/A) = P(B) \cdot P(A/B) \qquad \ldots(13.2)$$

Using the multiplication law, we can find the *conditional probabilities*

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

and

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

In the case of multiplication law, certain modification is required for *independent events*.

Suppose the occurrence of B does not depend upon the occurrence of A and vice versa, then the two events A and B are said to be mutually independent. In this case the two conditional probabilities $P(B/A)$ and $P(A/B)$ are equal to their respective non-conditional simple probabilities. Hence, for independence

$P(B/A) = P(B)$ and

$P(A/B) = P(A)$

Thus, for two independent events A and B, the probability of their simultaneous occurrence is the product of their respective probabilities.

$$P(AB) = P(A) \cdot P(B) = P(B) \cdot P(A) \quad \ldots (13.3)$$

Let us now consider some examples on the application of the probability laws.

13.3.3 Applications of Probability Laws

Let us work out some problems so that you get a fair idea of the application of the above mentioned probability laws.

Example 13.4 A dice is thrown. What is the probability of getting either 1 or 6?

It is clear that in a single throw the two events of 1 and 6 cannot occur together. Hence, the two events are mutually exclusive. Thus

$$P(\textit{either } 1 \textit{ or } 6) = P(1) + P(6) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

Example 13.5 One card is drawn from a pack of 52 cards. The card is *not replaced* in the pack and another card is drawn. What is the probability that both the cards are spade?

Here the first event is drawing a spade and the second event is drawing another spade given that the first card is a spade. Thus the second event is a conditional event. Let $P(A)$ be the probability of the first event and $P(B/A)$ be the probability of the second event given the occurrence of the first event. Now, $P(A) = \frac{13}{52} = \frac{1}{4}$. When the first card is a spade and is not replaced, there are 12 spades left in a pack of 51 cards. So, $P(B/A) = \frac{12}{51}$. Hence, the required probability is

$$P(A \cap B) = P(A) \cdot P(B/A) = \frac{13}{52} \times \frac{12}{51} = \frac{1}{4} \times \frac{12}{51} = \frac{3}{51}$$

We should note that when the card is not replaced after the first drawing, the two events are not independent as the probability of the occurrence of the second event depends upon the probability of the occurrence of the first event.

Example 13.6 One card is drawn from a pack of 52 cards. The card is *replaced* in the pack and another card is drawn. What is the probability that both the cards are spade?

In this example, the card is replaced after the first drawing. Thus, when the second card is drawn, there are 13 spades in a pack of 52 cards. As a result, the probability of the second card being a spade does not depend upon whether the first card drawn is a spade or not. Hence, the two events are independent here. If $P(B)$ is the probability of the second card being a spade, in this case

$$P(B/A) = P(B) = \frac{13}{52} = \frac{1}{4}.$$

Thus, the required probability is

$$P(A \cap B) = P(A) \cdot P(B/A) = P(A) \cdot P(B) = \frac{13}{52} \times \frac{13}{52} = \frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$$

Example 13.7 A dice is thrown. What is the probability of getting a number less than 5 or an odd number?

Let A be the event of a number less than 5 and B be the event of an odd number. We should note here that the two events are not mutually exclusive as a number can be both less than 5 and an odd number. So the required probability is obtained by applying the formula

$$P(A \cap B) = P(A) + P(B) - P(A \cap B)$$

Now in a dice, out of 6 numbers, there are 4 numbers (1, 2, 3, and 4) less than 5. Therefore,

$$P(A) = \frac{4}{6} = \frac{2}{3}.$$

Again, there are 3 odd numbers (1,3 and 5) out of the possible six numbers. So

$$P(B) = \frac{3}{6} = \frac{1}{2}$$

Suppose $P(B/A)$ is the probability of an odd number given that it is less than five. Then

$$P(B/A) = \frac{2}{4} = \frac{1}{2}$$

Now

$$P(AB) = P(A) \cdot P(B/A) = \frac{2}{3} \times \frac{1}{2} = \frac{1}{3}$$

Thus, the probability of getting a number less than 5 or an odd number is

$$\frac{2}{3} + \frac{1}{2} - \frac{1}{3} = \frac{5}{6}$$

Example 13.8 If A and B are two events such that $P(A) = \frac{2}{3}, P(\overline{A} \cap B) = \frac{1}{6}$ and $P(A \cap B) = \frac{1}{3}$. Find $P(B), P(A \cup B), P(A/B), P(B/A)$ $P(\overline{A} \cup B)$ and $P(\overline{A} \cap \overline{B})$. Also examine whether A and B are

a) Equally likely
b) Exhaustive
c) Mutually exclusive
d) Independent.

We can write

$$P(B) = P(\overline{A} \cap B) + P(A \cap B) = \frac{1}{6} + \frac{1}{3} = \frac{3}{6} = \frac{1}{2}$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{2}{3} + \frac{1}{2} - \frac{1}{3} = \frac{5}{6}$$

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{1}{3} \cdot 2 = \frac{2}{3}$$

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{1}{3} \cdot \frac{3}{2} = \frac{1}{2}$$

$$P(\overline{A} \cup B) = P(\overline{A}) + P(B) - P(\overline{A} \cap B) = \frac{1}{3} + \frac{1}{2} - \frac{1}{6} = \frac{2}{3} \quad [\because P(\overline{A}) = 1 - P(A)]$$

$$P(\overline{A} \cap \overline{B}) = 1 - P(A \cup B) = 1 - \frac{5}{6} = \frac{1}{6}$$ (Using the concept of complementary event)

a) Since $P(A) \neq P(B)$, A and B are not equally likely.

b) Since $P(A \cup B) \neq 1$, A and B are not exhaustive.

c) Since $P(A \cap B) \neq 0$, A and B are not mutually exclusive.

d) Since $P(A) \cdot P(B) = P(A \cap B)$, A and B are independent.

Check Your Progress 2

1) A student takes Mathematics and English tests. His independent chances of passing the two tests are $\frac{2}{3}$ and $\frac{3}{4}$ respectively. What is the probability that

   a) he passes at least one test?

   b) he fails in both the tests?

.....

.....

.....

.....

2) Two cards are drawn from a pack of 52 cards. What is the probability

   a) that both the cards are kings when the first card is *replaced* before the second card is drawn?

   b) that both the cards are spades when the first card is *not replaced* before the second card is drawn?

.....

.....

.....

.....

3) There are two urns. The first urn contains 7 white balls and 3 red balls. The second urn contains 4 white balls and 6 red balls. An urn is selected at random

and a ball is drawn. What is the probability that the first urn is selected and a red ball is drawn from it?

.....

.....

.....

.....

4) The probabilities that A and B speak the truth independently are $\frac{1}{2}$ and $\frac{1}{3}$ respectively. If they make the same statement, what is the probability that the statement made by them is a true one?

.....

.....

.....

.....

13.4 BAYES' THEOREM

Let A_1, A_2, and A_3 be three mutually exclusive and exhaustive events and there be an event D which can occur in conjunction with any of them. If D actually happens, then the conditional probability of the occurrence of A_i ($i = 1, 2, 3$) given D, is given by

$$P(A_i / D) = \frac{P(A_i \cap D)}{P(D)} = \frac{P(A_i) \cdot P(D / A_i)}{P(D)}$$

where

$$P(D) = \sum_{i=1}^{3} P(A_i \cap D) = \sum_{i=1}^{3} P(A_i) \cdot P(D / A_i)$$

We should note that the above result can be generalised to any number of mutually exclusive and exhaustive events.

Let us consider some practical applications of Bayes' Theorem.

Example 13.9 In a factory that produces bolts there are three machines *A, B* and *C.* They manufacture 25%, 35% and 40% of total output respectively. However, 5%, 4% and 2% of their respective output are defective. A bolt is drawn at random from a day's output and is found to be defective. What is the probability that it was produced by (i) machine *A,* (ii) machine *B,* and (iii) machine *C* ?

Since a day's output consists of the bolts produced by all the three machines, the probability that a bolt selected at random is produced by machine *A* is given by $P(A) = 0.25$. Similarly we have, $P(B) = 0.35$ and $P(C) = 0.40$. Further, let D be the event that the bolt is defective. Since machine *A* produces 5% defective bolts, we have $P(D / A) = 0.05$. Similarly, $P(D / B) = 0.04$ and $P(D / C) = 0.02$.

Thus

$$P(D) = P(A) \cdot P(D / A) + P(B) \cdot P(D / B) + P(C) \cdot P(D / C)$$

$$= 0.25 \times 0.05 + 0.35 \times 0.04 + 0.40 \times 0.02 = 0.0345$$

The probability that the bolt is manufactured by machine A given that it is defective

$$P(A/D) = \frac{P(A) \cdot P(D/A)}{P(D)} = \frac{0.25 \times 0.05}{0.0345} = \frac{0.0125}{0.0345} = 0.362$$

Similarly,

$$P(B/D) = \frac{P(B) \cdot P(D/B)}{P(D)} = \frac{0.35 \times 0.04}{0.0345} = \frac{0.0140}{0.0345} = 0.406$$

and

$$P(C/D) = \frac{P(C) \cdot P(D/C)}{P(D)} = \frac{0.40 \times 0.02}{0.0345} = \frac{0.0080}{0.0345} = 0.232$$

There is an alternative method you can pursue. You can determine the above probabilities by making the following table. The three events A, B and C have been renamed as A_1, A_2, and A_3 respectively.

A_i	A_1	A_2	A_3	Total
$P(A_i)$	0.25	0.35	0.45	1.00
$P(D/A_i)$	0.05	0.04	0.02	
$P(D \cap A_i)$	0.0125	0.014	0.008	
$P(A_i/D) = \frac{P(D \cap A_i)}{P(D)}$	0.362	0.406	0.232	1.00

Note that the probabilities $P(A_1)$, $P(A_2)$ and $P(A_3)$, which are known before conducting the trial, are known as *prior* probabilities. The conditional probabilities of A_1, A_2, and A_3 i.e., $P(A_1/D)$, $P(A_2/D)$ and $P(A_3/D)$, after the result of the trial is known; are termed as *posterior* probabilities.

To begin with the analysis of a problem, the manager of a firm assigns probabilities to certain events on subjective basis, i.e., based on his experience and expectation. These probabilities are the prior probabilities. Then the trial is conducted. Subsequent to the trial, the prior probabilities are revised on the basis of the occurrence of certain event like D to obtain posterior probabilities. Again in the next round, these posterior probabilities can be taken as prior probabilities and the whole procedure may be repeated to revise the posterior probabilities. Such a revision can be repeated a number of times. Generally, after a certain number of revisions the posterior probabilities tend to stabilise and the subjective probabilities tend to become objective probabilities. Thus, Bayes' Theorem proves to be very useful for the analysis of business phenomena.

Example 13.10 The probability that the product of a company will be successful given that the result of the survey is favourable is 0.6 and the probability of its being successful with unfavourable survey is 0.3. If the probability that the survey shows a favourable result is 0.7, find the probability that (i) the product is successful, (ii) the result of the survey is favourable given that the product is successful, and (iii) the result of the survey is unfavourable given that the product is successful.

Let S denote the event that the product of the company is successful and F denote the event that the survey result is favourable. Let $\overline{S}$ and $\overline{F}$ be the events denoting the negation of the respective events.

In terms of notations we are given

$P(S/F)=0.6,\ P(S/\overline{F})=0.3$ and $P(F)=0.7$

Thus we have

$P(\overline{F})=1-0.7=0.3$

i) The probability that the product is successful is given by

$$P(S)=P(S\cap F)+P(S\cap\overline{F})$$
$$=P(F)\cdot P(S/F)+P(\overline{F})\cdot P(S/\overline{F})$$
$$=0.7\times0.6+0.3\times0.3=0.51$$

ii) $P(F/S)=\dfrac{P(F\cap S)}{P(S)}=\dfrac{0.42}{0.51}=0.824$

iii) $P(\overline{F}/S)=\dfrac{P(\overline{F}\cap S)}{P(S)}=\dfrac{0.09}{0.51}=0.176$

Note that $P(\overline{F}/S)=1-P(F/S)$

Check Your Progress 3

1) A talcum powder manufacturing company had launched a new type of advertisement. The company estimated that a person who comes across their advertisement will buy their product with a probability of 0.7 and those who do not see the advertisement will buy the product with a probability of 0.3. If in an area of 1000 people, 70% had come across the advertisement, find the probability that a person who buys the product

   a) has not come across the advertisement.

   b) has come across the advertisement.

.....

.....

.....

.....

2) An insurance company insured 2000 scooter drivers, 4000 car drivers and 6000 truck drivers. The probability of an accident is 0.01, 0.03 and 0.15 in the respective categories. One of the insured drivers meets with an accident. What is the probability that the person is a scooter driver?

.....

.....

.....

.....

.....

.....

13.5 LET US SUM UP

In ordinary parlance, probability refers to chance. In statistics, however, we go deeper than this. Here, we not only consider the uncertainty involved in the occurrence of an event but also try to quantify it. Thus, probability is a quantitative measure of chance. In this Unit, we have considered three different approaches to probability, namely, the classical approach , the relative frequency approach and the axiomatic approach. The probabilities of compound events are essentially governed by two laws. They are the addition law and the multiplication law. Finally, Bayes' Theorem provides a framework for revising the probabilities on the basis of the occurrence and non-occurrence of certain events. This revision is very useful for business decisions.

13.6 KEY WORDS

Probability : It is a relative measure for the degree of certainty (and implicitly that of non-chance) associated with an event. For an event A, the probability varies between 0 and 1, that is, $0 \leq P(A) \leq 1$

Conditional Probability : If A and B are not mutually exclusive events then the probability of B given that A has already occurred is known as the conditional probability of B given A. It is denoted by $P(B/A)$.

Independent Events : Two events A and B are said to be mutually independent if the occurrence of B does not depend upon the occurrence of A and vice versa.

Complementary Event : If A is an event, then the non-occurrence of the event A, denoted by $\overline{A}$ is called complement to the event A. The sum of probabilities of any event and its complement is always equal to 1.

Prior Probabilities : The probabilities assigned to various events on the basis of the classical definition or statistical definition or in a subjective manner are prior probabilities.

Posterior Probabilities : The revised probabilities of various events are known as posterior probabilities. This revision is made on the basis of the occurrence or non-occurrence of certain events by using Bayes' Theorem.

13.7 SOME USEFUL BOOKS

Newbold, P., 1991. *Statistics for Business and Economics* (Third Edition): Prentice Hall, New Jersey, Chapter 3.

Nagar, A. L. and Das, R. K., 1989, *Basic Statistics:* Oxford University Press, Delhi, Chapter 8.

Anderson, D. R., Sweeney, D., J. and Williams, T., A., 1993. *Statistics for Business and Economics* (Fifth Edition): West Publishing Company, Minneapolis/ St. Paul, Chapter 4.

13.8 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) $\frac{2}{5}$

2) $\frac{1}{2}$

3) $\frac{1}{2}$

4) $\frac{12}{13}$

Check Your Progress 2

1) a) $\frac{11}{12}$, b) $\frac{1}{12}$

2) a) $\frac{1}{169}$, b) $\frac{1}{17}$

3) $\frac{3}{20}$

4) $\frac{1}{3}$

Check Your Progress 3

1) $\frac{9}{58}$, $\frac{49}{58}$

2) $\frac{1}{52}$

UNIT 14 PROBABILITY DISTRIBUTIONS-I

Structure

14.0 OBJECTIVES

After going through this unit you should be able to:

- explain the meaning of a random variable;
- explain the concept of a probability distribution;
- distinguish between a discrete probability distribution and a continuous probability distribution; and
- explain the binomial and the Poisson distributions.

14.1 INTRODUCTION

In Unit 13, we discussed about probability of the occurrence of an event. In that unit an event was defined as the set of one or more possible outcomes of a chance experiment. The outcomes of such a chance experiment can be related to the concept of a random variable. In the present unit, we shall consider probability in the context of a random variable and understand the notion of a probability distribution.

Any probability distribution is based upon the behaviour of some random variable. In this unit, we shall define a random variable and distinguish between a discrete random variable and a continuous random variable. Then, in the context of discrete random variables, we shall discuss two important discrete probability distributions. They are the binomial distribution and the Poisson distribution.

14.2 RANDOM VARIABLE

Before presenting the formal definition of a random variable, let us intuitively try to understand the concept of a random variable. As mentioned in the introduction, a random variable is related to the outcomes of a chance experiment. Such a chance experiment is also known as a random experiment. Let us consider an example.

Suppose a coin is tossed. There are two possible outcomes: a head (H) and a tail (T). In the previous unit, we have discussed the concept of a sample space. The sample space of this experiment consists of the outcomes head and tail. If S denotes the sample space, then

$$S = (H, T)$$

In this experiment, we are not sure whether a head will result or a tail. This is an example of a chance experiment or a random experiment. Now suppose, we assign a number 0 to the occurrence of tail (T = 0) and a number 1 to the occurrence head (H = 1). Let us define a variable X that refers to the occurrence of an outcome. Then the variable and its possible values can be written as

$$X = (0, 1)$$

However, there is an important difference between this variable and our common notion of a variable. Here, the value that the variable takes depends upon the outcome of the chance or random experiment that we are considering. In other words, we are not sure, whether as a result of the experiment, the variable will take the value 0 or 1. We can only attach some probabilities to these values. These probabilities depend upon the chances of the occurrence of the different outcomes of the experiment. If in our example, for instance, the coin is unbiased, the probability of the occurrence of tail is $\frac{1}{2}$ and that of head is also $\frac{1}{2}$; because, both the outcomes have an equal chance to occur. Accordingly, we attach a probability of $\frac{1}{2}$ to both 0 and 1. In the case of the conventional variable, on the other hand, no such probability is attached to a value taken by the variable.

From the above discussion we can say that a random variable is *a variable that takes different values with some probabilities*. Thus, the variable X referring to the possible outcomes of tossing a coin, is an example of a random variable. We will use the following notations: Let X be a random variable and it assumes values $x_1, x_2, x_3, \ldots, x_n$.

The corresponding probabilities are $p_1, p_2, p_3, \ldots p_n$. Thus $p\ (X = x_1) = P_1$

Example 14.1

Let us consider an experiment of simultaneous tossing of two coins. The sample space of the experiment is

$$S = \{(H, H), (H, T), (T, H), (T, T)\}$$

If we define a random variable X as the number of heads obtained, then $X = 2$ corresponds to the outcome (H, H); $X = 1$ corresponds to the outcomes (H, T)

and (T, H); and finally, $X = 0$ corresponds to the outcome (T, T). Thus X can take three possible values, i.e., 0, 1 and 2.

$$X = (0, 1, 2)$$

Example 14.2

As another example, we consider the roll of a dice. The sample space is $S = \{1, 2, 3, 4, 5, 6\}$. A random variable X can be defined such that it takes a value equal to 0 when an odd number appears on the dice and 1 when an even number appears. Thus

$$X = (0, 1)$$

In the tossing of two coins experiment, we can also define a random variable in monetary terms. For instance, we may decide to pay a player Rs. 10 if two heads are obtained, Rs. 5 if one head is obtained and ask the player to pay Rs. 8 (i.e., we pay Rs. –8) if no head is obtained. Here X is a random variable denoting the amount of payment that can be made to a player. Thus

$$X = (10, 5, -8)$$

A random variable can be either discrete or continuous.

i) *Discrete Random Variable*: When the sample space of an experiment is discrete, the corresponding random variable will also be discrete, i.e., it will take certain isolated values. The random variables discussed above are examples of discrete random variable.

ii) *Continuous Random Variable*: We know that a continuous variable can take any value in an interval. Accordingly, a continuous random variable is defined when the accompanying sample space is also continuous. In the next unit, we shall discuss the concept of the normal variable which is an example of a continuous random variable.

14.3 PROBABILITY DISTRIBUTION

Let us begin with a definition of probability distribution. It is defined as a statement about the possible values of a random variable along with their respective probabilities.

Let us take a concrete example of probability distribution. In the earlier example of tossing two coins, we defined a random variable X as the number of heads. Further, X took three values, viz., 0, 1 and 2. Assuming that the two coins are unbiased, we can write

$$p(X = 0) = \frac{1}{4},$$

$$p(X = 1) = \frac{1}{2}$$ [i.e., the probability of the occurrence of (H, T) or (T, H)]

$$p(X = 2) = \frac{1}{4}.$$

These probabilities along with the corresponding values of the random variable written in a tabular form constitute *the probability distribution of the random variable* X where X is the number of heads. It is shown in Table 14.1.

Table 14.1: Probability Distribution of the Number of Heads Obtained in Tossing of Two Unbiased Coins

Number of Heads (x)	Probability $p(x)$
0	$\frac{1}{4}$
1	$\frac{1}{2}$
2	$\frac{1}{4}$

In the above example, the events of getting no head ($X = 0$), one head ($X = 1$), and two heads ($X = 2$) *exhaust* all the possibilities (this means, there is no other possible outcome than the above three). Thus, the probability distribution resulting from the above experiment has enumerated all the possible values of the random variable X and assigned specific probabilities to them. We can see that the sum of these probabilities equals 1.

Probability distribution can be of two types: Discrete Probability Distribution and Continuous Probability Distribution.

14.3.1 Discrete Probability Distribution

We have already seen that the probability distribution for a random variable describes how the probabilities are distributed over the values of the random variable. Now, for a discrete random variable, the probability distribution is defined by a function called *probability mass function*, denoted by $p(x)$. This probability mass function provides the probability for each value of the discrete random variable. In fact, the probability distribution of the number of heads in tossing two coins that we have presented in Table 14.1 is an example of a discrete probability distribution. We can consider another example of a discrete probability distribution. Suppose we observe the number of children per household in a locality. Here, we can consider the number of children as a discrete random variable. A discrete probability distribution for the number of children per household can be constructed by computing the relative frequencies for the possible values of this random variable. Such a probability distribution is shown in Table 14.2.

Table 14.2: Probability Distribution of the Number of Children per Household

Number of Children (x)	$p(x)$
0	0.10
1	0.15
2	0.23
3	0.25
4	0.14
5	0.13

Thus, the set of ordered pairs $[x, p(x)]$ is called the probability distribution of a discrete random variable X or the discrete probability distribution.

Since the values $p(x)$ are all probabilities and a value x of the random variable will always occur, the probability mass function should satisfy the following two conditions

1) Probability of an event cannot be negative, i.e., for any value of X

$p(x) \geq 0$

2) Probabilities of all possible outcomes sum to unity, i.e.,

$\sum_{all\,x} p(x) = 1$

Let us work out some problems on discrete probability distribution.

Example 14.3

Is the following a valid probability mass function? $p(x) = \frac{x^3}{2}$, $x = -1, 0, 1$

Let us find out the probability of x, when x assumes the specified values (—1, 0 and 1).

When $x = -1$

$$p(x) = p(-1) = -\frac{1^3}{2} = -\frac{1}{2} < 0.$$

But we know that probability of an event cannot be negative. Thus, the first condition of a probability mass function is violated. Hence, the given function is not a valid probability mass function.

Example 14.4

Given a function $p(x) = \frac{k}{x}$, $x = 3, 4, 5$ and k is a constant. Find k such that the given function is a valid probability mass function.

From the given function, we have

$$p(3) = \frac{k}{3}$$

$$p(4) = \frac{k}{4}$$

$$p(5) = \frac{k}{5}$$

For the satisfaction of the second condition of a probability mass function, we have

$$\Sigma p(x) = \frac{k}{3} + \frac{k}{4} + \frac{k}{5} = 1$$

$$or\ k\left(\frac{1}{3} + \frac{1}{4} + \frac{1}{5}\right) = 1$$

$$or\ k \cdot \frac{47}{60} = 1$$

$$or\ k = \frac{60}{47}$$

When $k = \frac{60}{47}$

$$p(3) = \frac{1}{3} \cdot \frac{60}{47} = \frac{60}{141} \geq 0$$

$$p(4) = \frac{1}{4} \cdot \frac{60}{47} = \frac{60}{188} \geq 0$$

$$p(5) = \frac{1}{5} \cdot \frac{60}{47} = \frac{60}{235} \geq 0$$

Thus, for $k = \frac{60}{47}$, the first condition for a probability mass function is also satisfied.

14.3.2 Continuous Probability Distribution

A continuous random variable X has a zero probability of assuming exactly any of its values. Apparently, this seems to be a surprising statement. Let us try to understand this by considering a random variable say, weight. Obviously weight is a continuous random variable since it can vary continuously. Suppose, we do not know the weight of a person exactly but have a rough idea that her weight falls between 60 kg and 61 kg. Now, there are an infinite number of possible weights between these two limits. As a result, by its definition, the probability of the person's assuming a particular weight say, 60.3 kg will be negligibly small, almost equal to zero. But we can definitely attach some probability to the person's weight being between 60 kg and 61 kg. Thus, for a continuous random variable X, one assigns a probability to an interval and not to a particular value. Here, we look for a function $p(x)$, called the *probability density function*, such that with the help of this function we can compute the probability

$P(a < x < b)$, a and b are the limits of an interval (a, b) where, $a < b$

A probability density function is defined in such a manner that the area under its curve bounded by x-axis is equal to one when computed over the domain of X for which $p(x)$ is defined. The probability density function for a continuous random variable X defined over the entire set of real numbers R should satisfy the following conditions.

1) $p(x) \geq 0$ for all $x \in R$

2) $\int_{-\infty}^{+\infty} p(x)dx = 1$

3) $p(a < X < b) = \int_{a}^{b} p(x)dx$

Although the probability distribution of a continuous random variable cannot be presented in the form of a table like that of a discrete random variable, it can nevertheless be expressed by a specific form of the probability density function $p(x)$. We shall study some of these forms in the next unit on the theoretical distributions for continuous random variables.

14.3.3 Theoretical Distributions

We should note that a probability distribution is based upon the empirical observations associated with a probability experiment. For obtaining the relevant

probability distribution, the experiment has to be repeated a very large number of times under an identical condition which may sometimes prove to be an extremely difficult task. Alternatively, by means of a formula, we can specify a probability mass function or a probability density function, as the case may be, theoretically by satisfying the conditions of the experiment. Such kind of a probability distribution is known as a *theoretical distribution*. An advantage with the theoretical distribution is that a few theoretical distributions describe many real life random phenomena. Consequently, with the help of a handful of important theoretical distributions we may get an insight into several such real life random phenomena without actually experimenting with them. A theoretical distribution can either be a discrete or a continuous one. Later in this unit, we discuss two important *discrete theoretical distributions* that have often been employed in the statistical analysis. In the next unit, we shall study some continuous theoretical distributions.

14.4 MEAN AND VARIANCE OF A RANDOM VARIABLE

The mean of a random variable, also known as its *mathematical expectation* or *expected value* is defined as the sum of the products of the values of the random variable and the corresponding probabilities.

Thus, if X is a discrete random variable that can assume the values x_1, x_2, x_3, ..., x_n with specific probabilities $p_1, p_2, p_3, ..., p_n$ respectively, then the mathematical expectation of x is

$$E(X) = x_1p_1 + x_2p_2 + x_3p_3 + ... + x_np_n = \sum_{i=1}^{n} x_ip_i$$

It is interesting to note that the mathematical expectation of a random variable corresponds to the arithmetic mean of an ordinary variable. This can be easily shown for a discrete random variable. We have seen from the *relative frequency definition of probability* that the probability of an event can be interpreted as the limit of relative frequency of the occurrence of that event when the number of trials tends to infinity, i.e.,

$$p_i = \frac{f_i}{N}$$

where f_i is frequency of x_i and $N = \sum_{i=1}^{n} f_i$ is the total frequency.

Thus $E(X) = \sum_{i=1}^{N} p_ix_i = \sum_{i=1}^{N} \frac{f_i}{N} x_i$

$$= \frac{1}{N}\sum_{i=1}^{N} f_ix_i = \overline{X}$$, the arithmetic mean of X.

Example 14.5

A fair coin is tossed. If it is 'head', you win Rs. 20. If it is 'tail', you lose Rs. 10. What is the amount that you are expected to win or lose per toss?

Since the coin is given to be unbiased, the probability of getting a 'head' or a 'tail' is $\frac{1}{2}$. Let X be a random variable which takes the values equal

to the amounts of gain and loss. So, $X = 20$ with probability $\frac{1}{2}$ and $X = -10$ (loss can be considered to be negative gain) again with probability $\frac{1}{2}$.

Therefore, the expected amount to win or lose per toss is

$$Rs.\left[20 \cdot \frac{1}{2} + (-10) \cdot \frac{1}{2}\right] = Rs.5$$

A game with positive expected gain is said to be biased in favour of the player. If the expected gain is zero, the game is said to be fair. The above game can be made fair if we charge Rs. 5 (equal to the expected value) as entry fee. The possible values of the random variable X now become 15 and –15 and the expected value $E(X) = 0$.

For a continuous random variable, the mathematical expectation takes the form of a definite integral. Thus,

$$E(X) = \int_a^b x\, p(x)dx$$

where, X is a continuous random variable with domain from a to b and $p(x)$ is its probability density.

14.4.1 Theorems on Mathematical Expectation

i) The mathematical expectation of a constant is the constant itself. If c is a constant, then

$$E(c) = c$$

ii) The mathematical expectation of the product of a constant and a random variable is the product of the constant and the mathematical expectation of the random variable. If c is a constant and X is a random variable, then

$$E(cX) = cE(X)$$

iii) The mathematical expectation of any function of a random variable is the sum of the products of the values of the function and the corresponding probabilities of the values of the random variable. Thus if $f(X)$ is a function of a random variable X that takes the values $x_1, x_2, x_3, \ldots, x_n$ with specific probabilities the $p_1, p_2, p_3, \ldots, p_n$ mathematical expectation of $f(X)$ is

$$E[f(X)] = \sum_{i=1}^{n} (x_i) p_i$$

We may note here that the above summation, strictly speaking, applies to a discrete random variable. However, without any loss of generality, the theorem is valid for a continuous random variable also. There, instead of a summation over some finite values, an integration over the domain of the random variable has to be performed.

iv) The mathematical expectation of the sum of a given number of random variables is the sum of their expectations. If X and Y are two random variables, the mathematical expectation of $X + Y$ is

$$E(X + Y) = E(X) + E(Y)$$

v) The mathematical expectation of the product of a given number of *independent random variables* is the product of their expectations. If X and Y are two independent random variables, the mathematical expectation of XY is

$$E(XY) = E(X) \cdot E(Y)$$

The *variance* of a random variable X is given by

$$V(X) = E[X - E(X)]^2 = E(X^2) - [E(X)]^2$$

The variance of the random variable in Example 14.5 (tossing of a coin) can be computed in the following manner.

First we compute

$$E(X^2) = 20^2 \cdot \frac{1}{2} + 10^2 \cdot \frac{1}{2} = 200 + 50 = 250$$

$$[E(X)]^2 = 5^2 = 25$$

Now

$$V(X) = \sigma_x^2 = 250 - 25 = 225$$

Also the standard deviation of X

$$\sigma_x = \sqrt{225} = \text{ Rs. } 15$$

14.4.2 Theorem on Variance

i) The variance of a constant is zero. If c is a constant, then

$V(c) = 0$

ii) The variance of the product of a constant and a random variable is the product of the square of the constant and the variance of the random variable. If c is a constant and X is a random variable, then

$V(cX) = c^2 V(X)$

iii) The variance of the sum of a given number of random variables is the sum of their variances. If X and Y are two random variables, the variance of $X + Y$ is

$$V(X+Y) = V(X) + V(Y) + 2Cov(X,Y)$$

Here, $Cov(X, Y)$ is called the covariance between X and Y. We should note that covariance is a measure of simultaneous variability of the two variables.

Covariance can be shown as

$$Cov(X,Y) = E[\{X - E(X)\}\{Y - E(Y)\}] = E(XY) - E(X)E(Y)$$

But, if X and Y are independent, then a variation in one variable does not cause a variation in the other variable. Consequently, $Cov(X,Y) = 0$ and

$$V(X+Y) = V(X) + V(Y)$$

It may be noted here that all the theorems on mathematical expectation and variance discussed above are valid for both discrete and continuous random variables.

We can now state and prove an important result.

14.4.3 Standard Normal Variate

For any variable (random or otherwise) with a given mean and standard deviation, whenever the mean is subtracted from it and the result is divided by the standard deviation; the resultant variable has a mean equal to zero and a standard deviation equal to one.

Let us prove the above statement.

Let X be a random variable with mean (expectation) μ and standard deviation σ. Suppose

$$z = \frac{X-\mu}{\sigma}$$

$$E(z) = E\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}E(X-\mu) = \frac{1}{\sigma}[E(X) - E(\mu)] = \frac{1}{\sigma}(\mu-\mu) = 0$$

Now

$$V(z) = V\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2}V(X-\mu) = \frac{1}{\sigma^2}[V(X) + (-1)V(\mu)]$$

The covariance term $Cov(X,\mu)$ above vanishes because X and μ are independent. Thus

$$V(z) = \frac{1}{\sigma^2}[V(X)] = \frac{1}{\sigma^2}\sigma^2 = 1 \quad [\because V(\mu) = 0]$$

Now

$$Std.dev(z) = \sqrt{V(z)} = \sqrt{1} = 1$$

In this way we can see

$$E(z) = E\left(\frac{X-\mu}{\sigma}\right) = 0$$

and

$$Std.dev(z) = Std.dev\left(\frac{X-\mu}{\sigma}\right) = 1$$

The variable z defined in the above manner is called *standard normal variate.*

In the next unit, we shall consider how this result is used in the context of the normal distribution.

Check Your Progress 1

1) Check whether the following is a valid probability mass function or not.

$$p(x) = \frac{x^2 - x}{16}, \quad x = -2, -1, 0, 1, 2, 3.$$

.....

.....

.....

.....

2) Find k such that the following is a valid probability mass function.

$$p(x)=\frac{k}{x^2},\ x=1,2$$

.....

.....

.....

.....

3) A and B throw a dice for an amount of Rs. 99 to be won by one who throws a six first. If A has the first throw, what are their respective expected gain?

.....

.....

.....

.....

4) A contractor spends Rs. 3,000 to prepare for a bid on a construction project which, after deducting the manufacturing and the cost of bidding, will yield a profit of Rs. 25,000 if the bid is won. If the chance of winning the bid is ten percent, compute the contractor's expected profit and state the likely decision on whether to bid or not.

.....

.....

.....

.....

5) Prove the following results

a) $E(cX) = cE(X)$, where c is a constant.

b) $V(c) = 0$, where c is a constant.

c) $V(cX) = c^2V(X)$, where c is a constant.

.....

.....

.....

.....

14.5 BINOMIAL DISTRIBUTION

The binomial distribution is an example of a discrete probability distribution. James Bernoulli presented it in the year 1700. The word 'binomial' suggests 'two'. It signifies two possible outcomes of an experiment, the occurrence of an event or the non-occurrence of the event. A probability experiment can be termed as a Bernoulli experiment, if it satisfies the following conditions.

1) The experiment consists of a sequence of n repeated trials.

2) Each trial results in an outcome that may be classified either as a *success* or a *failure*.

3) The probability of a success, denoted by p, is known and remains the same in each trial. Consequently, the probability of a failure, denoted by $q = (1 - p)$ is also known and remains the same in each trial.

We may get some idea about a Bernoulli experiment by considering the experiment of tossing a coin a certain number of times and counting the number of heads appearing. Suppose, the coin is unbiased and we toss it 5 times. It is clear that the experiment consists of a sequence of 5 identical trials. There are two possible outcomes of each toss, a head (success) and a tail (failure). The probability of a head (success) is $\frac{1}{2}$ and this does not vary from one toss to another. The probability of a tail (failure) is again $\frac{1}{2}$ and this also does not vary from one toss to another. Finally, the tosses are independent in the sense that the outcome of one toss in no way depends upon the outcome of another toss. Thus, we find that this experiment of tossing a coin a certain number of times and observing the number of heads appearing satisfies all the conditions of a Bernoulli experiment.

In a Bernoulli experiment, we are interested in deriving the probability of a given number of successes, say, x occurring in n trials. (For example, in the previous example we may be interested in finding out the probability of getting 3 heads in 5 tosses.) It is clear that the random variable X can assume values 0, 1, 2, 3, ..., n. Suppose, we denote a success by S and a failure by F, then x successes and $(n-x)$ failures may occur in a number of different sequences. One possible sequence is that the first x trials are all successes and the remaining $(n-x)$ trials are all failures. Symbolically the sequence is shown by

$$\underbrace{SS\ldots S}_{x \text{ times}}\ \underbrace{FF\ldots F}_{(n-x)\text{ times}}$$

The probability of the above sequence of x successes and $(n-x)$ failures can be obtained by applying the multiplication theorem of probability. The probability is

$$\underbrace{pp\cdots p}_{x \text{ times}} \times \underbrace{(1-p)(1-p)\cdots(1-p)}_{(n-x)\ times} = p^x(1-p)^{n-x}$$

But, as mentioned earlier, x successes and $(n-x)$ failures can occur in other sequences also. However, each of the sequences in which x successes and $(n-x)$ failures occur will have a probability of $p^x(1-p)^{n-x}$. Thus, the probability of x successes in n trials is the probability of the occurrence of the x successes and $(n-x)$ failures in any of the possible sequences. This probability can be obtained by applying the addition theorem of probability over the possible sequences. But, as the probability of x successes and $(n-x)$ failures is the same for each of the possible sequences, the required probability of x successes in n trials is the product of the total number of possible sequences and the probability of the occurrence of a sequence. The total number of sequences in which x successes (and $n-x$ failures) can occur in n trials is basically a problem of obtaining the number of combinations of n things taken x at a time and is denoted by ${}^nC_x$. From the mathematics of *permutation* and *combination*, we have

$${}^nC_x = \frac{n!}{x!(n-x)!}$$

where

$$n! = n(n-1)(n-2)\ \ldots\ 2.1$$

$x! = x(x - 1)(x- 2) \ldots 2.1$

and

$0! = 1.$

We should note that the notation '!' is called *factorial*. For example, 4! = 4×3×2×1 = 24. The symbol C in ${}^nC_x$ represents combintion.

For example, ${}^5C_3 = \frac{5!}{3!(5-3)!} = \frac{5!}{3!2!}$

$$= \frac{5\times4\times3\times2\times1}{3\times2\times1\times2\times1}$$

$$= 10$$

Thus, the probability of x successes in n trials is given by

$$p(x) = {}^nC_x p^x (1-p)^{n-x}$$

$x = 0, 1, 2, \ldots\ldots, n.$

The above expression is the probability mass function for the Binomial distribution. This function has been used for presenting the Binomial distribution of $x = 0, 1, 2, \ldots n$ successes in n trials in Table 14.3. We note that a binomial distribution has two parameters n and p. It means that the distribution is completely specified if the values of n and p are known.

Table 14.3: Binomial Distribution

Number of Successes x	**Probability** $p(x)$
0	$(1-p)^n$
1	$np(1-p)^{n-1}$
2	$\frac{n(n-1)}{2.1}p^2(1-p)^{n-2}$
⋮	⋮
N	p^n
Total	**1**

Now let us find out the mean of the binomial distribution.

Let there be n number of trials in a Bernoulli experiment with p as the probability of a success in a trial. This implies the probability of failure is υ.

$$E(X) = \sum_{x=0}^{n} x\ {}^nC_x p^x (1-p)^{n-x}$$

By simplifying the above we find that the mean of binomial distribution is given by np. We do not present the proof of the above as it is quite cumbersome.

Similarly the variance of binomial distribution is given by

$$V(X) = E[X - E(X)]^2 = E(X^2) - [E(X)]^2$$

By simplification of the above it can be shawn that the variance of binomial distribution is given by npq [which is equal to $np\,(1-P)$].

Thus we observe that the mean and variance of binomial distribution are given by its two parameters n and p. We give some examples on the applications of binomial distribution below.

Example 14.6

A machine is generally known to be producing 20 per cent defective items. A quality control inspector selects 5 items at random. Find the probability of getting (i) exactly 1 defective item, (ii) at least 3 defective items.

This is an example of binomial distribution with $p = 0.20$ and $n = 5$. Let us now solve the question.

(i) We know that the probability of x defective items (i.e., $n - x$ non-defective items) in n items is ${}^{n}C_{x}p^{x}(1-p)^{n-x}$. Here $n = 5$ and $x = 1$. Thus the probability of exactly 1 defective item is

$$p(1) = {}^{5}C_{1}(0.20).(0.80)^{4} = 0.4096.$$

(ii) At least 3 defective items means that there can be 3 or 4 or 5 defective items. Thus the probability of at least 3 defective items is the probability of 3 defective items plus the probability of 4 defective items plus the probability of 5 defective items.

Now, the probability of 3 defective items is

$$p(3) = {}^{5}C_{3}(0.20)^{3}(0.80)^{2} = 0.0512$$

The probability of 4 defective items is

$$p(4) = {}^{5}C_{4}(0.20)^{4}\cdot 0.8 = 0.0064$$

The probability of 5 defective items is

$$p(5) = {}^{5}C_{5}(0.20)^{5} = 0.0003$$

Therefore, the probability of at least 3 defective items = 0.0512 + 0.0064 + 0.0003 = 0.0579.

Example 14.7

If a fair dice is thrown 36 times, what is the expected number of times of getting a 6? What is the variance?

If p is the probability of getting a 6, then $p = \frac{1}{6}$ and $(1-p) = \frac{5}{6}$. Now, $n = 36$.

The mathematical expectation

$$np = 36 \times \frac{1}{6} = 6$$

Thus, one can expect to get a 6, six times when a dice is thrown 36 times.

The variance

$$np(1-p) = 36 \times \frac{1}{6} \times \frac{5}{6} = 5$$

14.6 POISSON DISTRIBUTION

The Poisson distribution is another discrete probability distribution. It is named after a French mathematician Simeon Poisson who derived this distribution in 1837. The distribution is in fact a special (limiting) case of binomial distribution. When the probability of success, p, in a binomial distribution is very small and the number of trials, n, is so large that the expectation, $\mu = np$, is a finite quantity; the binomial distribution tends to Poisson distribution. This distribution is particularly useful while dealing with the number of occurrences of some thing over a specified interval of time or space. For example, the random variable under consideration may be the number of telephone calls arriving at a telephone switch-board in 1 hour, the number of leaks in 100 kilometres of pipeline, or the number of bus accidents reported in a particular day in Delhi.

To obtain the probability mass function of Poisson distribution, we can consider the example of the number of telephone call, x, in an hour and assume that the expected number of telephone calls per hour (i.e., the mathematical expectation) is λ. For the applicability of the binomial distribution, we divide the interval of one hour into sub-intervals that are so small that the probability p of having a telephone call in a sub-interval is very low and that of getting more than one call is approximately zero. Thus, each sub-interval can be treated as a Bernoulli trial having only two possible outcomes; either there will be a telephone call (a success) or no telephone call (a failure). The number of sub-intervals is taken to be equal to n, the total number of trials. We note that the expected number of telephone calls, λ, remains the same and is equal to np (as we have seen from the binomial distribution). Therefore, the probability of a telephone call in a sub-interval is $\frac{\lambda}{n}$.

Thus, the probability of x telephone calls in one hour amounts to finding the probability of x successes in n trials when n tends to infinity (as argued above, n trials correspond to n sub-periods that constitute one hour and n tends to infinity as a result of making each sub-trial extremely small so that the total number of trials tends to be infinitely large). This probability is given as the following limit of a binomial distribution.

$$\lim_{n\to\infty} {}^{n}C_{x}\left(\frac{\lambda}{n}\right)^{x}\left(1-\frac{\lambda}{n}\right)^{n-x}$$

Let us try to find the above-mentioned limit.

The Poisson probability mass function is derived from the above and is given by

$$p(x) = \frac{\lambda^{x}\, e^{-\lambda}}{x!}$$

$$x = 0,1,2\cdots$$

where, X is a random variable denoting the number of successes in a specified time interval or length interval.

λ = expected value or average number of occurrences in an interval of time or length etc.

e = a constant (base of the natural logarithm) whose value is $e = 2.7182\cdots$

Table 14.4: Poisson Distribution

The value of the Poisson random variable x	Probability
0	$e^{-\lambda}$
1	$\frac{\lambda e^{-\lambda}}{1!}$
2	$\frac{\lambda^2 e^{-\lambda}}{2!}$
⋮	⋮
Total	**1**

In Poisson distribution, there is no upper limit on the random variable X, the number of occurrences. It is a discrete random variable that can assume an infinite sequence of values ($x = 0, 1, 2,3, \ldots$). The distribution has only one parameter λ. Table 14.4 presents the Poisson distribution generated by the Poisson.probability mass function.

Expectation and Variance

The expectation of the Poisson distribution is given by the constant λ. It can be shown that the variance of the Poisson distribution is also given by λ.

Example 14.8

An analysis shows that on an average 10 cars arrive at a petrol pump in a 15-minute period of time.

i) Find the probability of the arrival of 5 cars in 15 minutes.

ii) What is the probability of the arrival of 1 car in 3 minutes?

Here, $\lambda = 10$ (since average is given to be 10) and $x = 5$. Thus, the required probability is $p(5) = \frac{10^5 e^{-10}}{5!} = 0.0378$.

Since the expected number of arrivals in a 15-minute period is 10, the expected number of arrivals in a 3-minute period is $\frac{10}{15} \times 3 = 2$. Thus for the second part of the question, $\lambda = 2$ and $x = 1$. Hence, the probability of one arrival in a 3-minute period is

$$p(1) = \frac{2e^{-2}}{1!} = 0.2707.$$

Check Your Progress 2

1) In a hospital, there are 3 ambulances for the transportation of patients. The probability of the availability of an ambulance is 0.75. If an ambulance is needed what is the probability that

a) no ambulance will be available?

b) at least one ambulance will be available?

.....

.....

.....

.....

2) Can for a binomial distribution the mean and the variance be 3 and 5 respectively?

.....

.....

.....

.....

3) It is known from the past experience that in a certain plant, there are on an average 4 industrial accidents per month. Find the probability that in a given month, there will be less than 4 accidents. Assume Poisson distribution.

.....

.....

.....

.....

14.7 LET US SUM UP

In this unit, we have used the concept of probability to understand the meaning of a probability distribution. We have understood the concepts of the mathematical expectation and the variance of a probability distribution. We have distinguished between a discrete probability distribution and a continuous probability distribution. In this context, we have introduced the notions of a probability mass function and a probability density function. We have studied two specific discrete probability distributions, namely, the binomial distribution and the Poisson distribution. We have learnt the characteristics of these distributions, particularly, the expressions for their mean and variance. We have tried to understand the use of these two distributions in various situations.

14.8 KEY WORDS

Binomial Distribution : It is a discrete probability distribution that satisfies the following conditions:

1) It involves repetition of identical trials.
2) Each trial results in an outcome that may be classified either as a *success* or a *failure*.
3) The probability of a success, denoted by p, is known and remains the same in each trial. Consequently, the probability of a failure, denoted by $q = (1 - p)$ is also known and remains the same in each trial.
4) The trials are independent.
5) The binomial random variable x is the total number of successes in n trials; $0 \leq x \leq n$.

Continuous Probability Distribution : It is the probability distribution for a continuous random variable.

Continuous Random Variable : It is a random variable that can take all values in a certain interval.

Discrete Probability Distribution : It is the probability distribution for a discrete random variable.

Discrete Random Variable : It is a random variable that either assumes a finite number of values or an infinite sequence (like 1, 2, 3, ...).

Mathematical Expectation : The mathematical expectation or the expected value of a random variable is the sum of the products of the values of the random variable and the corresponding probabilities.

Poisson Distribution : It is a discrete probability distribution and is a limiting form of the binomial distribution when the probability of success, p, in a binomial distribution is very small and the number of trials, n, is so large that the expectation, $\mu = np$ is finite. The mean and variance for the Poisson distribution is the same.

Probability Density Function : It is a function of a continuous random variable. However, like the probability mass function, it cannot directly give the probability for a specified value of the random variable. Here, we can only find the probability of the random variable lying in an interval.

Probability Distribution : It is a statement about the possible values of a random variable along with their probabilities.

Probability Mass Function : It is a function that gives the probability for a specified value of a discrete random variable.

Theoretical Distribution : It is a probability distribution that is generated by specifying the conditions of a random experiment. Some examples of probability distributions are the binomial distribution, the Poisson distribution, the normal distribution, etc.

Variance : If $E(x)$ is the mathematical expectation of a random variable x, the variance of x is defined as $\mathrm{E}\,[\,x - \mathrm{E}(x)]^2$.

14.9 SOME USEFUL BOOKS

IGNOU Course Material EEC-03, (1992) *Probability and Probability Distributions:* Block 7, Unit 14.

Newbold, P. (1991) *Statistics for Business and Economics* (Third Edition)*:* Prentice Hall, New Jersey, Chapters 4 and 5.

Walpole, R. E., Myers, R. H., Myers, S. L. and Ye, K. (2002) *Probability and Statistics for Engineers and Scientists* (Seventh Edition): Pearson Education, India, Chapters 3, 4, and 5.

Anderson, D. R., Sweeney, D., J. and Williams, T., A., (1993) *Statistics for Business and Economics* (Fifth Edition): West Publishing Company, Minneapolis/ St. Paul, Chapters 5 and 6.

Bhardwaj, R. S. (1999) *Business Statistics* (First Edition): Excel Books, New Delhi, Chapters 18 and 19.

14.10 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) $p(-2)=\frac{3}{8}, p(-1)=\frac{1}{8},$

$p(0)=0$

$p(1)=0$

$p(2)=\frac{1}{8}$

$p(3)=\frac{1}{8}$

Since $p(x) \geq 0$ for all values of x, the first condition is satisfied.

$\sum p(x)=\frac{3}{8}+\frac{1}{8}+0+0+\frac{1}{8}+\frac{3}{8}=1.$

Therefore, second conditon is also satisfied.

Hence, the given function is a valid probability mass function.

2) $k=\frac{4}{5}$

3) *A* can win if *A* throws a six in the first throw *or A* cannot throw a six in the first throw *and* B cannot throw a six in the second throw *and* A throws a six in the third throw, and so on. Therefore, the probability that *A* throws a six first

$$=\frac{1}{6}+\frac{5}{6}\cdot\frac{5}{6}\cdot\frac{1}{6}+\frac{5}{6}\cdot\frac{5}{6}\cdot\frac{5}{6}\cdot\frac{5}{6}\cdot\frac{1}{6}+\cdots=\frac{6}{11}$$

Thus, the expected gain of A is $99\times\frac{6}{11}$ = Rs. 54

The probability that B throws a six first

= 1 – P (A throws a six first) = $1-\frac{6}{11}=\frac{5}{11}$

So, the expected gain of B is $99\times\frac{5}{11}$ = Rs. 45

4) Rs. (-)200

5) a) $E(cx)=E(c)E(x)=cE(x)$ (because, expectation of a constant is the constant itself)

b) $V(c)=E(c^2)-[E(c)]^2=c^2-c^2=0$

c) $V(cX)=E(c^2X^2)-[E(cX)]^2=c^2E(X^2)-c^2[E(X)]^2$

$=c^2\left[E(X^2)-\{E(X)\}^2\right]=c^2V(X)$

Check Your Progress 2

1) a) $\frac{1}{64}$ b) $\frac{63}{64}$

2) Mean np = 3, variance $np(1-p)$ = 5. Now,

$\frac{np(1-p)}{np}=\frac{5}{3}$

or $1-p=\frac{5}{3}$

or $p=1-\frac{5}{3}=-\frac{2}{3}<0$, which is not possible.

3) 0.4332.

UNIT 15 PROBABILITY DISTRIBUTIONS-II

Structure

15.0 OBJECTIVES

After going through this unit you should be able to:

- explain and use the normal distribution;
- explain the concept of the degrees of freedom; and
- form some elementary ideas about the chi-square distribution, the student' s-*t* distribution and the *F* distribution.

15.1 INTRODUCTION

In the previous Unit, we made a distinction between a discrete random variable and a continuous random variable. In that unit, we introduced the concept of a probability distribution and found that it is essentially a statement regarding the values taken by a random variable with their associated probabilities. We studied two important discrete probability distributions namely, the binomial distribution and the Poisson distribution. In the present Unit we will continue with the topic and study a very important continuous probability distribution called the normal distribution. It may be mentioned that the normal distribution plays an important role in the statistical inferences and tests of hypotheses and these are going to be our subject matter of Block-7. In fact, we will consider in Unit 16 the topic of sampling distribution that forms the foundation of statistical inference and tests of hypotheses. However, sampling distribution can be properly appreciated only if we have some rudimentary ideas about three other continuous probability distributions besides the normal distribution, viz., the chi-square distribution, the student's-*t* distribution and the *F* distribution. We will discuss about these probability distributions below.

15.2 NORMAL DISTRIBUTION

Normal distribution is perhaps the most widely used distribution in Statistics and related subjects. It has found applications in inquiries concerning heights and weights of people, IQ scores, errors in measurement, rainfall studies and so on. Abraham de Moivre gave the mathematical equation for the normal distribution in 1733. Karl Friedrich Gauss also independently derived its equation from a study of errors in repeated measurements of the same quantity. Accordingly, sometimes it is also referred to as the Gaussian distribution. The distribution has provided the foundation for much of the subsequent development of mathematical statistics.

We have seen in the previous Unit that for a continuous random variable, the counterpart of a probability mass function is the *probability density function.* We shall denote the probability density function also by $p(x)$. The probability density function of a continuous random variable that follows the normal distribution is given by

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where $-\infty < x < \infty$, and
$\pi = 3.17141$ (approximately)
$e = 2.71828$ (approximately).

It is clear that the normal density function is completely determined by the parameters μ and σ. It means that given the values of μ and σ, we can trace out the normal curve by obtaining the values of $p(x)$ for different values of x. In fact, it can be shown that μ and σ are respectively the mean and the standard deviation of the normal distribution. When a random variable X follows normal distribution with mean μ and standard deviation σ we write it in symbols as $X \sim N(\mu, \sigma)$ and read as 'X follows normal distribution with mean μ and standard deviation σ.' The normal curve is a symmetrical bell-shaped curve as shown in Fig. 15.1.

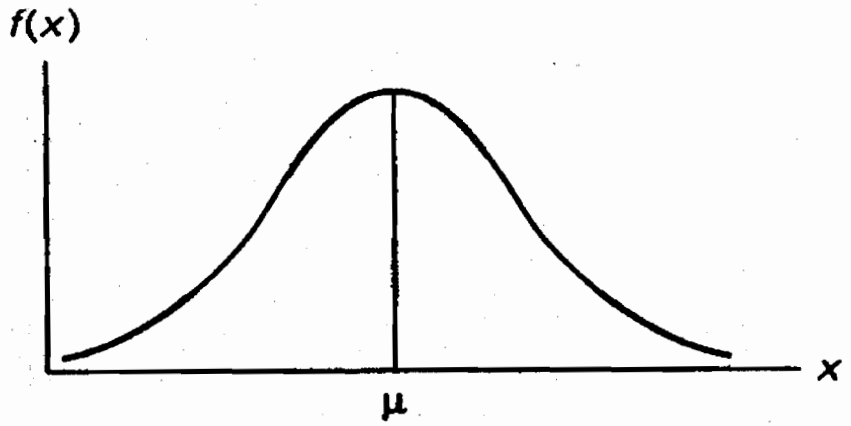


Fig 15.1: Normal Curve

The important features of the normal distribution are as follows:

1) The normal curve stretches from $-\infty$ to $+\infty$. This means that a normal random variable (X) assumes values between $-\infty$ to $+\infty$.

2) The curve is symmetric about its mean, i.e., $= \mu$. This means that corresponding to $x = \mu + a$ and $x = \mu - a$, the values of $p(x)$ are the same (for any arbitrarily chosen 'a').

3) The median and the mode of the distribution coincide with the mean. Thus mean = median = mode = μ.

4) The maximum of the normal curve occurs at $x = \mu$. Thus $p(x)$ is maximum when $x = \mu$.

5) The points of inflexion of the normal curve occurs at $x = \mu + \sigma$ and $x = \mu - \sigma$. At the points of inflexion, the normal curve changes its curvature.

The following area-properties hold for a normal distribution. In Fig.15.2 below we plot a normal curve with mean $\mu = 50$ and standard deviation $\sigma = 4$.

a) 68.8 % of the area under the normal curve lies between the ordinates at $\mu - \sigma$ and $\mu + \sigma$. Thus in Fig.15.2, 68.8% area is covered when x ranges between 46 and 54.

b) 95.5% of the area under the normal curve lies between the ordinates at $\mu - 2\sigma$ and $\mu + 2\sigma$. In Fig. 15.2, 95.5% area is covered when $42 \le X \le 58$.

c) 99.7% of the area (i.e., almost the whole of the distribution) under the normal curve lies between the ordinates at $\mu - 3\sigma$ and $\mu + 3\sigma$. In Fig. 15.2 we find that 99.7% area is covered when $38 \le X \le 62$.

If we have different values of μ and σ, the range of x mentioned in Fig. 15.2 will change.

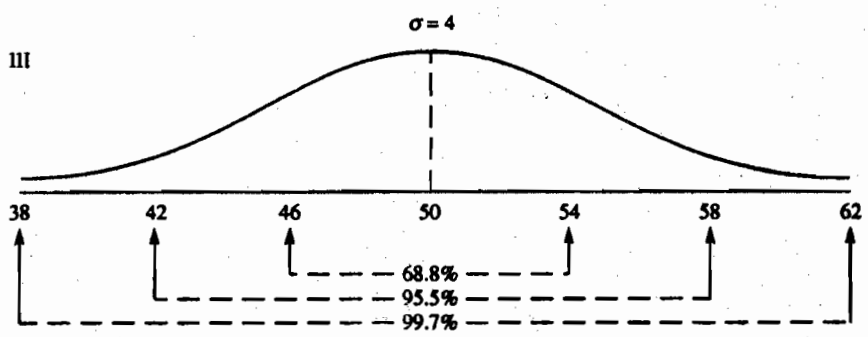


Fig. 15.2: Area Under Normal Curve

15.2.1 Standard Normal Curve

We have seen in the previous unit that the curve for any continuous probability distribution or probability density function is so traced out that the area under the curve bounded by the two ordinates corresponding to $x = x_1$ and $x = x_2$ gives the probability that the random variable assumes a value between $x = x_1$ and $x = x_2$. Thus, for a normal curve

$$P(x_1 \le X \le x_2) = \int_{x_1}^{x_2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{x^1}^{x^2} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Obviously, this probability depends upon the values of the two parameters μ and σ. However, it is very difficult to solve the above-mentioned integral of the normal distribution. This has necessitated the tabulation of normal curve areas for quickly obtaining the probabilities of the normal variable assuming values in different intervals. But, it is really meaningless to attempt to construct a separate table for every conceivable combination of values for μ and σ. Fortunately, the solution to an apparently hopeless task has been achieved by the application of a standard result in statistics that we have seen and proved in the last unit. Let us recapitulate the result. We have seen that

For any variable with a given mean and standard deviation, whenever the mean is subtracted from the variable and the result is divided by the standard deviation; the resultant variable has a mean equal to zero and a standard deviation equal to one.

Thus if X is a variable with mean (expectation) μ and standard deviation σ then

$z = \frac{X - \mu}{\sigma}$ has a mean equal to zero and standard deviation equal to one.

It means that normal variables with different combinations of μ and σ can all be transformed into a unique normal variable with mean 0 and standard deviation 1.

Thus if X is a normal variable with mean (expectation) μ and standard deviation σ, then $z = \frac{X - \mu}{\sigma}$ for any combination of μ and σ, is always a normal variable with mean 0 and standard deviation 1.

Symbolically,

If $X \sim N(\mu, \sigma)$

then

$$z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

for any μ and σ.

Such a transformed normal variable is called a *standard normal variate*. The probability density function of the standard normal variate z is given by

$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \quad -\infty < z < \infty$$

Once we obtain a standard normal variate, our seemingly hopeless task of obtaining probability areas for different combinations of μ and σ becomes elegantly simple. Let us see how. We should note that a standard normal variate has a unique mean of 0 and a unique standard deviation of 1. It means, if we can construct a table for probability areas of such a unique standard normal variate, it can be used for obtaining probability for any normal variable with any combination of mean and standard deviation. The only thing is that the given normal variable is to be transformed into the standard normal variate. In fact, such a table for areas (or probability) has been compiled for a standard normal variate (see Table 15.1: Areas under the Standard Normal Curve) and is very much in use in statistics. Thus, for the computation of the required probability for any normal variable with some mean and standard deviation, the upper and the lower limits say, $x = x_1$ and $x = x_2$ of the given interval are converted into the corresponding z-values say, $z = z_1$ and $z = z_1$ and the relevant area is obtained from Table 15.1 given at the end of this Unit.

Remember that the standard normal curve is symmetrical and it covers an area of 1.0 Since the value of z ranges between $-\infty$ and ∞, we find that the area between $0 < z < \infty$ is 0.5 (half the area under standard normal curve). Similarly, the area between $-\infty < z < 0$ is 0.5. Since the standard normal curve is symmetric we have one advantage; the area under the curve is the same on both sides. In Table 15.1 the area for different positive values of z are given.

If we look into column 1 of Table 15.1 we find that values assumed by z ranges

from 0.0 to 3.0. Corresponding to each value there are 10 columns marked .00, .01, ..., .09. These columns represent the second digit after decimal. For example, if $z = 0.45$, then we look for the row corresponding to 0.4. On this row we move to the right and look for the column representing .05. In Table 15.1 we find that when $z = 0.45$ the area covered is 0.1736. Note that when z = –0.45 the area under standard normal curve again is 0.1736. As another example, the area under the standard normal curve when $z = 1.31$ is 0.4049. Theoretically z can assume any value between $-\infty$ and ∞. However, when $z = 3.09$ the area covered is 0.4990. Therefore, in Table 15.1 areas for $z > 3.09$ are not given.

Let us now consider some examples to see the applications of the normal area table.

Example 15.1

Find the area under the standard normal curve in each of the following cases by using Table 15.1 for areas under the standard normal curve.

a) Between $z = 0$ and $z = 1.8$.

b) Between $z = -0.25$ and $z = 0$.

c) Between $z = -0.52$ and $z = 2.25$.

a) In Table 15.1, let us move downward under the column marked z until we reach the entry 1.8. Now, let us turn right to the column marked 0. We find here an entry equal to 0.4641. This is the required area.

b) Since the standard normal curve is symmetric about the mean, the required probability between $z = -0.25$ and $z = 0$ can be obtained by finding the area between $z = 0$ and $z = 0.25$ from the table. So, let us move downward under the column marked z until we reach the entry 0.2. Then we turn right to the column marked .05. We find here an entry equal to 0.0987. This is the required area.

c) It is clear that the required area is

(area between $z = -0.52$ and $z = 0$) + (area between $z = 0$ and $z = 2.25$)

= (area between $z = 0$ and $z = 0.52$) (by symmetry) + (area between $z = 0$ and $z = 2.25$)

= 0.1985 + 0.4878

= 0.6863.

Example 15.2

In a sample of 120 workers in a factory the mean and standard deviation of daily wages are Rs. 11.35 and Rs. 3.03 respectively. Find the percentage of workers getting wages between Rs. 9 and Rs. 17 in the whole factory assuming that the wages are normally distributed.

Let x be a random variable denoting wages. Then, x is a normal variable with mean $\mu = 11.35$ and the standard deviation $\sigma = 3.03$. The corresponding standard normal variate

$$z = \frac{x - 11.35}{3.03}$$

For $x = 9$, $z = \frac{9 - 11.35}{3.03} = \frac{-2.35}{3.03} = -0.78$

For $x = 17$, $z = \frac{17 - 11.35}{3.03} = \frac{5.63}{3.03} = 1.86$

$$\therefore P(9 \le x \le 17) = P(-0.78 \le z \le 1.86)$$
$$= P(-0.78 \le z \le 0) + P(0 \le z \le 1.86)$$
$$= P(0 \le z \le 0.78) + P(0 \le z \le 1.86)$$
$$= 0.2823 + 0.4686$$
$$= 0.7509 = 75.09\%$$

Thus, 75.09% of workers get wages between Rs. 9 and Rs. 17.

15.2.2 Normal Approximation to the Binomial Distribution

Sometimes in statistics, one distribution is obtained as the limiting form of another distribution. For example, in Unit 14, we have learnt that at times the probability of success, p, in a binomial distribution is very small and the number of trials, n, is so large that the expectation, $\mu = np$, is a finite quantity. In such cases the binomial distribution tends to Poisson distribution. You can recall that both the binomial distribution and the Poisson distribution are discrete distributions. However, the limiting form of the discrete binomial distribution need not be uniquely the discrete Poisson distribution. In fact, when n is very large and p is not extremely close to 0 or 1; the binomial distribution approaches the continuous normal distribution. As a result, in real world situations, the normal distribution is often used for approximating the binomial distribution. We have already seen that the standard normal table comes very handy in obtaining probabilities for a normal variable falling within a specified interval. We may now state a result that helps us to use areas under the standard normal curve to approximate the binomial properties of a random variable:

If X is a binomial variable with mean $\mu = np$ and variance $\sigma^2 = npq$ then

$$z = \frac{X - np}{\sqrt{npq}}$$

tends to a normal distribution with mean zero and standard deviation one, as n tends to infinity.

In symbols,

$$\lim_{n \to \infty} \frac{X - np}{\sqrt{npq}} \sim N(0,1)$$

It has been observed that the normal distribution provides a good approximation for a binomial distribution even when n is not very large and p is reasonably close to 0.5.

Let us consider an example.

Example 15.3

Suppose the probability of a particular kind of machine being defective is 0.4. A quality control inspector examines a lot of 15 such machines to identify the defective machines.

Find the probability that the inspector will find 4 machines to be defective.

We know from our discussion on binomial distribution in Unit 14 that the distribution of defective machines in a given lot is a binomial variable. Here

$n = 15$, $p = 0.4$ and $q = 1 - p = 0.6$

The probability of 4 defective machines is given by

$$^{15}c_4(0.4)^4(0.6)^{11} = \frac{15!}{4!\,1!}(0.4)^4(0.6)^{11} = 1365 \times 0.0256 \times 0.0036279 = 0.1268$$

Now suppose, we want to find the required probability by the normal approximation. Then

$np = 15 \times 0.4 = 6$, $npq = np\,(1-p) = 15 \times 0.4 \times 0.6 = 3.6$ and $\sqrt{npq} = \sqrt{3.6}$ = 1.897.

We have

$$z = \frac{X - np}{\sqrt{npq}}$$

which is approximately a standard normal variate. But the standard normal variate is a continuous random variable. We know that for such a random variable, the probability of its taking a particular value cannot be determined. It is only the probability of the random variable lying within an interval that can be obtained. Thus the probability of the binomial variable taking a value 4 has to be translated into the probability of the corresponding normal variable falling in an interval around the value 4 for the required normal approximation. Since, a binomial variable assumes zero and positive integer values; the value immediately preceding 4 and the value immediately succeeding 4 that the binomial variable under question can take are 3 and 5 respectively. As a result, the probability of the binomial variable taking a value equal to 4 can be reasonably approximated by the probability of the corresponding normal variable falling within an interval (3.5, 4.5). The required probability is thus approximately equal to the area under the normal curve between the ordinates = 3.5 and = 4.5. Converting to z-values, we have

$$z_1 = \frac{x_1 - np}{\sqrt{npq}} = \frac{3.5-6}{1.897} = -1.32 \text{ and } z_2 = \frac{x_2 - np}{\sqrt{npq}} = \frac{4.5-6}{1.897} = -0.79$$

In X is the binomial variable and z is the corresponding standard normal variate, then $P(X = 4) = P(-1.32 \le z \le -0.79) = 0.1214$ (From the area under standard normal curve given in Table 15.1).

We can see that the normal approximation of 0.1214 agrees quite closely with the actual probability of 0.1268 for 4 defective machines obtained from the binomial distribution.

Check Your Progress 1

1) Find the area under the normal curve in the following cases.
   a) Between $z = 1.55$ and $z = 2.55$.
   b) To the left of $z = -1.5$.
   c) To the right of $z = 2.5$.

.....

.....

.....

.....

2) The mean height of 1000 persons is 68 inches and the standard deviation is 5 inches. If the heights are normally distributed, find how many persons have heights between 67 inches and 69 inches.

.....

.....

.....

.....

3) A company, that sells 5000 batteries in a year, guarantees them for a life of 24 months. The life of the batteries is estimated to be approximately normal with mean equal to 34 months and standard deviation equal to 5 months. Find the number of batteries that will have to be replaced under the guarantee.

.....

.....

.....

.....

15.3 SOME OTHER CONTINUOUS DISTRIBUTIONS

There are some other continuous probability distributions that play important roles in various branches of statistics. In Block 7, we are going to study statistical inference. In that Block, in addition to normal distribution, we shall often use concepts of three more continuous distributions, *Chi-Square* (Pronounced as *kai-square* and denoted by the symbol χ^2) *distribution*, the *Student's-t distribution* and the *F-distribution.* In this section, we are going to discuss these distributions in brief. We begin with a general concept, the degrees of freedom, which finds applications in all these distributions.

15.3.1 Degrees of Freedom

In connection with these distributions, we shall often come across a term: degrees of freedom. Let us get some idea about the term now. In a general sense, *the degrees of freedom refers to the number of pieces of independent information* that are required to compute some characteristic (say, variance) of a given set of observations. We consider an example here.

Suppose there are 5 observations: 4, 7, 12, 3 and 15. Hence, the arithmetic mean $\bar{X}$ is 8. The computation of variance $\frac{1}{n}\Sigma(x_i - \bar{X})^2$ involves obtaining the squares of the deviations of the values of the observations from their arithmetic mean and adding them as shown below:

$(4 - 8)^2 + (7 - 8)^2 + (12 - 8)^2 + (3 - 8)^2 + (14 - 8)^2$

$= (-4)^2 + (-1)^2 + (4)^2 + (-5)^2 + (6)^2$

From the properties of arithmetic mean ($\bar{X}$), we know that the summation of the values inside the brackets must be equal to zero, i.e., in general $\sum_{i=1}^{n}(x_i - \bar{X}) = 0$.

It means that in the computation of the variance, if the first four terms inside the brackets happen to be –4, –1, 4 and –5 respectively, the fifth term cannot be

any other term but 6. Thus in this example, we do not have 5 independent pieces of information inside the brackets. The fifth piece of information inside the bracket, i.e., '6' depends upon the first four pieces of information inside their respective brackets.

We can fix the idea better if we think of a person who does not have any idea about the individual observations. She is only told that there are 5 observations and the first 4 deviations of the values from the mean of the five observations are –4, –1, 4 and –5 respectively. She is then asked to calculate the variance of the 5 observations. If she knows the law that the sum of the deviations of the values of a variable from their arithmetic mean is always equal to zero, she will at once be able to figure out that the deviation of the 5th value from its arithmetic mean is 6. She will then proceed to take the squares of these deviations and add them together to arrive at the last step for the computation of the required variance. The last step involves a division by the number of observations to get an idea about the average dispersion of the values about their arithmetic mean (we take variance as the required measure here). What is the appropriate value that she should take for the number of observations? Is it 5? Let us probe a little into it. We have seen that the fifth deviation is determined by the first four deviations, which means that there are 4 independent pieces of information that produce the dispersions of the given 5 values about their arithmetic mean. Therefore, for measuring the average dispersion, i.e., the variance, the sum of the squares of the five deviations should be divided by 4 and not by 5. Thus, in this example, the degrees of freedom are 4. Generalising the above example, for obtaining the variance of n observations, there are $n-1$ degrees of freedom because of the restriction $\sum_{i=1}^{N}(x_i-\bar{X})=0$.

As a result, for variance, $\sum_{i=1}^{N}(x_i-\bar{X})^2$ is divided by $n-1$, i.e., $S^2=\frac{1}{n-1}\sum_{i=1}^{n-1}(x_i-\bar{X})^2$ where, S^2 is the variance. From the above discussion, we can say that the degrees of freedom that we have for the computation of any characteristic is equal to the number of observations minus the number of restrictions put on the computation of the required characteristic. In symbols, $d.f. = n-r$ where, $d.f.$ is the degrees of freedom, n is the number of observations and r is the number of restrictions.

We should note here that when n is quite large, for calculating the variance or its positive square root, i.e., the standard deviation, we often divide $\sum_{i=1}^{N}(x_i-\bar{X})^2$ by n and not by $n-1$. However, strictly speaking, we should divide $\sum_{i=1}^{N}(x_i-\bar{X})^2$ by $n-1$, particularly when n is small.

15.3.2 The χ^2 (Chi-Square) Distribution

Suppose, X is a normal variable with mean (expectation) μ and σ, then $z=\frac{X-\mu}{\sigma}$ is a standard normal variate, i.e., $z \sim N(0,1)$. If we take the square

of z, i.e., $z^2 = \left(\frac{X-\mu}{\sigma}\right)^2$, then z^2 is said to be distributed as a χ^2 variable with one degree of freedom and is expressed as χ_1^2.

- It is clear that χ_1^2 is a squared term; for z lying between $-\infty$ and $+\infty$, χ_1^2 will lie between 0 and $+\infty$ (because a squared term cannot take negative values). Again since, z has a mean equal to zero, most of the values taken by z will be close to zero. As a result, the probability density of a χ^2 variable will be maximum near zero.

Generalising the result mentioned above, if $z_1, z_2, ..., z_k$ are independent standard normal variates (i.e., normal variables with zero mean and unit variance), then the variable $z = \sum_{i=1}^{k} z_i^2$ is said to be a χ^2 variable with k degrees of freedom and is denoted by χ_1^2.

Fig. 15.3 given below shows the probability curves for χ^2 variables with different degrees of freedom.

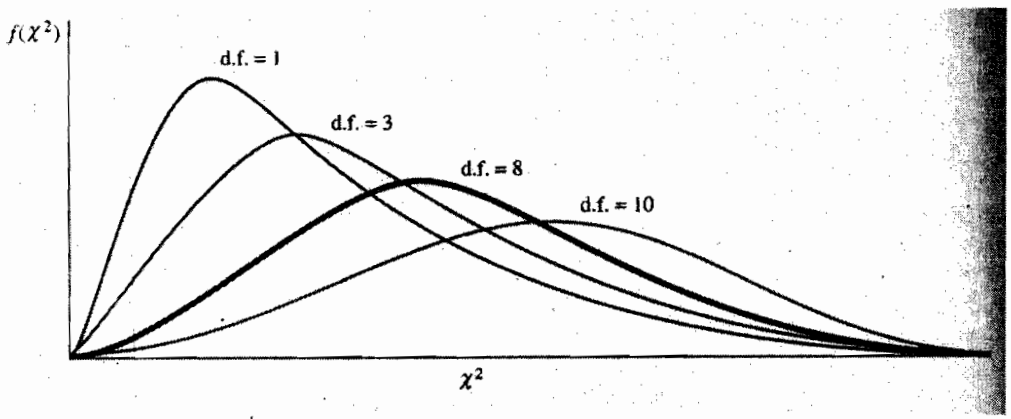


Fig 15.3: Chi-square Probability Curves

We should note the following features of the χ^2 distribution.

1) As Fig. 15.3 shows, the χ^2 is a positively skewed distribution. Its degree of skewness depends on its degrees of freedom. For lower degrees of freedom, the distribution is highly skewed. As the number of degrees of freedom increases, the distribution becomes increasingly symmetric. In fact, for degrees of freedom more than 100, the variable $\sqrt{2\chi^2} - \sqrt{(2k-1)}$ can be treated as a standard normal variate, where k is the degrees of freedom.
2) The mean of the chi-square distribution is k, and its variance is $2k$, where k is the degrees of freedom.
3) If Z_1 and Z_2 are two independent chi-square variables with k_1 and k_2 degrees of freedom respectively, then $Z_1 + Z_2$ is also a chi-square variable with degrees of freedom equal to $k_1 + k_2$.

As in the case of the normal distribution, a similar table has been prepared for

consult this table to obtain the required probability of the χ^2 variable for different degrees of freedom.

In Table 15.2 *df* represent degrees of freedom. While the columns $\chi^2_{0.05}$ and $\chi^2_{0.01}$ denotes χ^2 values for 5% and 1% *level of significance* respectively. We will explain the concept of level of significance in Unit 18.

The following example illustrates the use of the chi-square table.

Example 15.4

What is the probability of obtaining a χ^2 value of 34 or greater, given the degrees of freedom 25?

We can see from Table 15.2 that if we move down the degrees of freedom column to reach the figure of 25, the nearest figure to 34 that we find across the corresponding row is 34.08747. The probability for 34.08704, as we can see from the table, is 0.10. Hence the required probability is 0.10.

15.3.3 The Student's-*t* Distribution

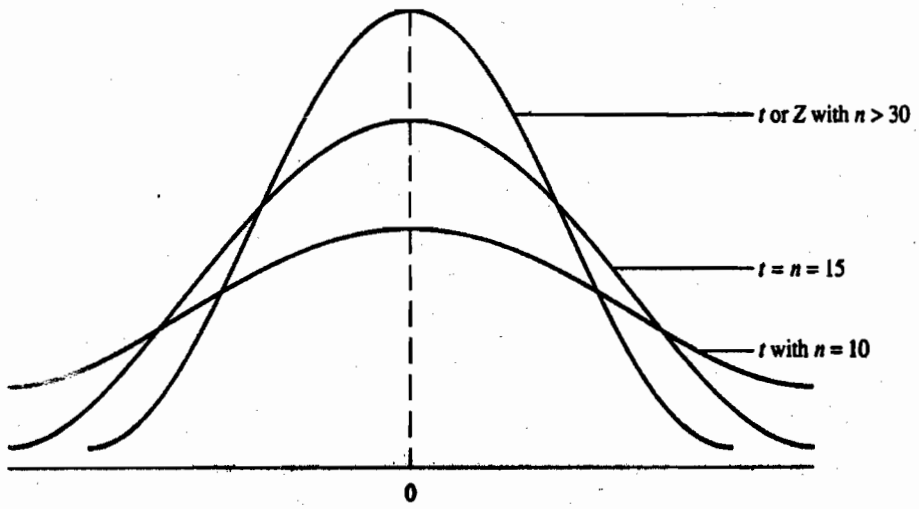
W.S. Gosset presented the *t*-distribution. The interesting story is that Gosset was employed in a brewery in Ireland. The rules of the company did not permit any employee to publish any research finding independently. So, Gosset adopted the pen-name 'student' and published his findings about this distribution anonymously. Since then, the distribution has come to be known as the student's-*t* distribution or simply, the *t* distribution.

If z_1 is a standard normal variate, i.e., $z_1 \sim N(0,1)$ and z_2 is another independent variable that follows the chi-square distribution with *k* degrees of freedom, i.e., $z_2 \sim \chi^2_1$, then the variable

$$t = \frac{z_1}{\sqrt{z_2/k)}} = \frac{z_1\sqrt{k}}{\sqrt{z_2}}$$

is said to follow student's-*t* distribution with *k* degrees of freedom.

The probability curves for the student's-*t* distribution for different degrees of freedom are presented in Fig. 15.4



We may note the important characteristics of this distribution.

1) As we can see in Fig. 15.4, like the normal distribution, the student's-*t* distribution is also symmetric and its range of variation is also from $-\infty$ to $+\infty$; however, it is flatter than the normal distribution. We should also note here that as the degrees of freedom increase, the student's-*t* distribution approaches the normal distribution.

2) The mean of the student's-*t* distribution is zero and its variance is $\frac{k}{(k-2)}$, where, k is the degrees of freedom.

Like the normal distribution, the student's-*t* distribution is often used in statistical inferences and tests of hypotheses to be discussed in Block-7. The task involves the integration of its density function; which may prove to be tedious. As a result, in this case also, like the normal distribution, a table has been constructed for ready-reference purposes (see Table 15.3).

We shall consider an example to see the use of Table 15.3.

Example 15.5

Given the degrees of freedom equal to 10, what is the probability of obtaining a *t* value of (i) 2.7638 or greater, (ii) –2.7638 or lower?

i) In Table 15.3 for student's-*t* distribution, first, we move down the degrees of freedom column and reach the figure of 10 and then look across the corresponding row and locate the figure of 2.7638. The corresponding lower probability figure of 0.01 is the required probability.

ii) Since the student's-*t* distribution is symmetric, the probability of obtaining a *t* value of –2.7638 or lower is also 0.01.

15.3.4 The *F* Distribution

Another continuous probability distribution that we are going to discuss in this Unit is the *F* distribution.

If z_1 and z_2, are two chi-square variables that are independently distributed with k_1 and k_2 degrees of freedom respectively, the variable

$$F = \frac{z_1 / k_1}{z_2 / k_2}$$

follows *F* distribution with k_1 and k_2 degrees of freedom respectively. The variable is denoted by F_{k_1,k_2}, where, the subscripts k_1 and k_2 are the degrees of freedom associated with the chi-square variables.

We may note here that k_1 is called the numerator degrees of freedom and in the same way, k_2, is called the denominator degrees of freedom.

Some important properties of the *F* distribution are mentioned below.

1) The *F* distribution, like the chi-square distribution, is also skewed to the right. But, as k_1 and k_2 increase, the *F* distribution approaches the normal distribution.

2) The mean of the *F* distribution is $k_1/(k_2 - 2)$, which is defined for $k_2 > 2$,

and its variance is $\frac{2k_2^2(k_1+k_2-2)}{k_1(k_2-2)^2\ (k_2-4)}$ which is defined for $k_2 > 4$.

3) An F distribution with 1 and k as the numerator and denominator degrees of freedom respectively is the square of a student's-t distribution with k degrees of freedom. Symbolically,

$F_{1,k} = t_k^2$

4) For fairly large denominator degrees of freedom k_2, the product of the numerator degrees of freedom k_1 and the F value is approximately equal to the chi-square value with degrees freedom k_1, i.e., $k_1 F = \chi^2_{k_1}$.

As we have mentioned earlier with reference to other continuous probability distributions, F distribution is also extensively used in statistical inference and testing of hypotheses. Again, such uses also require obtaining areas under the F probability curve and consequently integrating the F density function. However, in this case also our task is facilitated by the provision of the F Table.

15.3.5 Distributions Related to the Normal Distribution

We have already seen from the features of the chi-square, student's-t and the F distributions that for large degrees of freedom, these distributions approach the normal distribution. Consequently, these distributions are also known as the distributions related to the normal distribution. This relationship between the chi-square distribution, the student's-t distribution and the F distribution on one hand and the normal distribution on the other has tremendous practical implications. When the degrees of freedom happen to be fairly large, instead of using the chi-square distribution or the student's-t distribution or the F distribution separately as the situation may demand; we can uniformly apply the normal distribution. As a result, our task gets considerably simplified.

Check Your Progress 2

1) What is the probability of obtaining a χ^2 value of 8 or greater, given the degrees of freedom 20?

.....

.....

.....

2) Given the degrees of freedom equal to 25, what is the probability of obtaining a t value of 1.708 or greater?

.....

.....

.....

3) Given $k_1 = 10$ and $k_2 = 8$, what is the probability of obtaining an F value of 5.8 or greater?

.....

.....

.....

15.4 LET US SUM UP

In this unit, we studied some continuous probability distributions. Among these distributions, the normal distribution is considered to be the most important one. We have learnt its characteristics and seen its practical applications. We have considered the important concept of a standard normal variate. We have also learnt the technique of using the table for the areas under the standard normal curve for solving problems relating to the normal distribution. Besides the normal distribution, we have considered three other continuous probability distributions. These are: the chi-square distribution, the student's-t distribution and the F distribution. These three distributions use the notion of the degrees of freedom. So, we have tried to explain the concept of the degrees of freedom. We have learnt the characteristics of these distributions and seen how these distributions can be applied to various situations by using tables relating to these distributions. Finally, we have seen that for fairly large degrees of freedom, these three distributions approach the normal distribution.

15.5 KEY WORDS

Chi-square Distribution : It is an asymmetric distribution where the range of variation for the random variable is from zero to infinity. For fairly large degrees of freedom, it approaches the normal distribution.

Chi-square Variable : A random variable that follows the chi-square distribution.

Continuous Probability Distribution : It is the probability distribution for a continuous random variable.

Continuous Random Variable : It is a random variable that can take all the values in a given interval.

Degrees of Freedom : It refers to the number of pieces of independent information that are required to compute some characteristic of a given set of observations.

Discrete Probability Distribution : It is the probability distribution for a discrete random variable.

Discrete Random Variable : It is a random variable that either assumes a finite number of values or an infinite sequence (like 1, 2, 3, ...).

***F* Distribution** : It is an asymmetric distribution that is skewed to the right. For fairly large degrees of freedom, it approaches the normal distribution.

***F* Variable** : A random variable that follows the F distribution.

Gaussian Distribution : It is the other name for the normal distribution.

Normal Distribution : The best known of all the theoretical

probability distributions. It traces out a bell-shaped symmetric probability curve.

Normal Variable : A random variable that follows the normal distribution.

Probability Distribution : It is a statement about the possible values of a random variable along with their probabilities.

Standard Normal Variate : A normal variable with mean 0 and standard deviation equal to 1.

Student's-*t* Distribution : It is a symmetric distribution about the value zero. The range of variation for the student's-*t* random variable is $-\infty$ to $+\infty$. It is, however, flatter than the normal distribution curve. For fairly large degrees of freedom, it approaches the normal distribution.

Student's-*t* variable : A random variable that follows the student's-*t* distribution.

15.6 SOME USEFUL BOOKS

Anderson, D. R., Sweeney, D.J. and Williams, T.A., 1993. *Statistics for Business and Economics* (Fifth Edition), West Publishing Company, Minneapolis/St. Paul, Chapter 6.

Bhardwaj, R.S.1999. *Business Statistics* (First Edition), Excel Books, New Delhi, Chapters 19 and 20.

IGNOU Course Material EEC-03, 1992. *Probability and Probability Distributions,* Block 7, Unit 15.

Newbold, P. 1991. *Statistics for Business and Economics* (Third Edition), Prentice Hall, New Jersey, Chapter 5.

Walpole, R. E., Myers, R. H., Myers, S. L. and Ye, K. 2002. *Probability and Statistics for Engineers and Scientists* (Seventh Edition), Pearson Education, India, Chapter 6.

15.7 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) a) 0.0552
   b) 0.0668
   c) 0.0062

2) The proportion of persons having heights between 67 inches and 69 inches = 0.1586

   Therefore, the number of persons having heights between 67 inches and 69 inches = $1000 \times 0.1586 = 159$ (approximately).

3) $z = \frac{24-34}{5} = -2$

From the standard normal table, the area between $z = 0$ and $z = 2$ is 0.4772. Therefore, the area between $z = 0$ and $z = -2$ is 0.4772 (because the standard normal distribution is symmetric). For finding the number of batteries that will have to be replaced, we have to consider the area to the left of $z = -2$, which is equal to the area to the right of $z = 2$. Now, the area to the right of $z = 2$ is $0.5 - 0.4772 = 0.0228$. Therefore, the probability that a battery is defective is 0.0228. Thus out of 5000 batteries, the number of batteries that will have to be replaced = $0.0228 \times 5000 = 114$.

Check Your Progress 2

1) 0.99

2) 0.05

3) 0.01

Table 15.1: Normal Area Table

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Table 15.2: Critical Values of Chi-square Distribution

df\area	0.1	0.05	0.025	0.01	0.005
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.860
5	9.236	11.071	12.833	15.086	16.750
6	10.645	12.592	14.449	16.812	18.548
7	12.017	14.067	16.013	18.475	20.278
8	13.362	15.507	17.535	20.090	21.955
9	14.684	16.919	19.023	21.666	23.589
10	15.987	18.307	20.483	23.209	25.188
11	17.275	19.675	21.920	24.725	26.757
12	18.549	21.026	23.337	26.217	28.300
13	19.812	22.362	24.736	27.688	29.819
14	21.064	23.685	26.119	29.141	31.319
15	22.307	24.996	27.488	30.578	32.801
16	23.542	26.296	28.845	32.000	34.267
17	24.769	27.587	30.191	33.409	35.718
18	25.989	28.869	31.526	34.805	37.156
19	27.204	30.144	32.852	36.191	38.582
20	28.412	31.410	34.170	37.566	39.997
21	29.615	32.671	35.479	38.932	41.401
22	30.813	33.924	36.781	40.289	42.796
23	32.007	35.172	38.076	41.638	44.181
24	33.196	36.415	39.364	42.980	45.559
25	34.382	37.652	40.646	44.314	46.928
26	35.563	38.885	41.923	45.642	48.290
27	36.741	40.113	43.195	46.963	49.645
28	37.916	41.337	44.461	48.278	50.993
29	39.087	42.557	45.722	49.588	52.336
30	40.256	43.773	46.979	50.892	53.672

Table 15.3: Critical Values of *t* Distribution

dfp	0.25	0.10	0.05	0.025	0.01	0.005
1	1.0000	3.0777	6.3138	12.7062	31.8205	63.6567
2	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.7649	1.6377	2.3534	3.1825	4.5407	5.8409
4	0.7407	1.5332	2.1318	2.7765	3.7470	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0321
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	2.3646	2.9980	3.4995
8	0.7064	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.6955	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.6938	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.6924	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.6912	1.3406	1.7531	2.1315	2.6025	2.9467
16	0.6901	1.3368	1.7459	2.1199	2.5835	2.9208
17	0.6892	1.3334	1.7396	2.1098	2.5669	2.8982
18	0.6884	1.3304	1.7341	2.1009	2.5524	2.8784
19	0.6876	1.3277	1.7291	2.0930	2.5395	2.8609
20	0.6870	1.3253	1.7247	2.0860	2.5280	2.8453
21	0.6864	1.3232	1.7207	2.0796	2.5177	2.8314
22	0.6858	1.3212	1.7171	2.0739	2.5083	2.8188
23	0.6853	1.3195	1.7139	2.0687	2.4999	2.8073
24	0.6849	1.3178	1.7109	2.0639	2.4922	2.7969
25	0.6844	1.3163	1.7081	2.0595	2.4851	2.7874
26	0.6840	1.3150	1.7056	2.0555	2.4786	2.7787
27	0.6837	1.3137	1.7033	2.0518	2.4727	2.7707
28	0.6834	1.3125	1.7011	2.0484	2.4671	2.7633
29	0.6830	1.3114	1.6991	2.0452	2.4620	2.7564
30	0.6828	1.3104	1.6973	2.0423	2.4573	2.7500
inf	0.6745	1.2816	1.6449	1.9600	2.3264	2.5758

Table 15.4: Critical Values of *F* Distribution

(5% level of significance)

df2/df1	1	2	3	4	5	6	7	8	9	10
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883	240.543	241.882
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396
3	10.128	9.552	9.277	9.117	9.014	8.941	8.887	8.845	8.812	8.786
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
5	6.608	5.786	5.410	5.192	5.050	4.950	4.876	4.818	4.773	4.735
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
8	5.318	4.459	4.066	3.838	3.688	3.581	3.501	3.438	3.388	3.347
9	5.117	4.257	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
10	4.965	4.103	3.708	3.478	3.326	3.217	3.136	3.072	3.020	2.978
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
15	4.543	3.682	3.287	3.056	2.901	2.791	2.707	2.641	2.588	2.544
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
21	4.325	3.467	3.073	2.840	2.685	2.573	2.488	2.421	2.366	2.321
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.237
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.266	2.220
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165
40	4.085	3.232	2.839	2.606	2.450	2.336	2.249	2.180	2.124	2.077
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993
120	3.920	3.072	2.680	2.447	2.290	2.175	2.087	2.016	1.959	1.911
inf	3.842	2.996	2.605	2.372	2.214	2.099	2.010	1.938	1.880	1.831

Table 15.4: Critical Values of *F* Distribution (Contd.)

(5% level of significance)

df2/df1	12	15	20	24	30	40	60	120	INF
1	243.906	245.950	248.013	249.052	250.095	251.143	252.196	253.253	254.314
2	19.413	19.429	19.446	19.454	19.462	19.471	19.479	19.487	19.496
3	8.745	8.703	8.660	8.639	8.617	8.594	8.572	8.549	8.526
4	5.912	5.858	5.803	5.774	5.746	5.717	5.688	5.658	5.628
5	4.678	4.619	4.558	4.527	4.496	4.464	4.431	4.399	4.365
6	4.000	3.938	3.874	3.842	3.808	3.774	3.740	3.705	3.669
7	3.575	3.511	3.445	3.411	3.376	3.340	3.304	3.267	3.230
8	3.284	3.218	3.150	3.115	3.079	3.043	3.005	2.967	2.928
9	3.073	3.006	2.937	2.901	2.864	2.826	2.787	2.748	2.707
10	2.913	2.845	2.774	2.737	2.700	2.661	2.621	2.580	2.538
11	2.788	2.719	2.646	2.609	2.571	2.531	2.490	2.448	2.405
12	2.687	2.617	2.544	2.506	2.466	2.426	2.384	2.341	2.296
13	2.604	2.533	2.459	2.420	2.380	2.339	2.297	2.252	2.206
14	2.534	2.463	2.388	2.349	2.308	2.266	2.223	2.178	2.131
15	2.475	2.403	2.328	2.288	2.247	2.204	2.160	2.114	2.066
16	2.425	2.352	2.276	2.235	2.194	2.151	2.106	2.059	2.010
17	2.381	2.308	2.230	2.190	2.148	2.104	2.058	2.011	1.960
18	2.342	2.269	2.191	2.150	2.107	2.063	2.017	1.968	1.917
19	2.308	2.234	2.156	2.114	2.071	2.026	1.980	1.930	1.878
20	2.278	2.203	2.124	2.083	2.039	1.994	1.946	1.896	1.843
21	2.250	2.176	2.096	2.054	2.010	1.965	1.917	1.866	1.812
22	2.226	2.151	2.071	2.028	1.984	1.938	1.889	1.838	1.783
23	2.204	2.128	2.048	2.005	1.961	1.914	1.865	1.813	1.757
24	2.183	2.108	2.027	1.984	1.939	1.892	1.842	1.790	1.733
25	2.165	2.089	2.008	1.964	1.919	1.872	1.822	1.768	1.711
26	2.148	2.072	1.990	1.946	1.901	1.853	1.803	1.749	1.691
27	2.132	2.056	1.974	1.930	1.884	1.836	1.785	1.731	1.672
28	2.118	2.041	1.959	1.915	1.869	1.820	1.769	1.714	1.654
29	2.105	2.028	1.945	1.901	1.854	1.806	1.754	1.698	1.638
30	2.092	2.015	1.932	1.887	1.841	1.792	1.740	1.684	1.622
40	2.004	1.925	1.839	1.793	1.744	1.693	1.637	1.577	1.509
60	1.917	1.836	1.748	1.700	1.649	1.594	1.534	1.467	1.389
120	1.834	1.751	1.659	1.608	1.554	1.495	1.429	1.352	1.254
inf	1.752	1.666	1.571	1.517	1.459	1.394	1.318	1.221	1.000

Table 15.4: Critical Values of F Distribution (contd.)
(1% level of significance)

df2/df1	1	2	3	4	5	6	7	8	9	10
1	4052.181	4999.500	5403.352	5624.583	5763.650	5858.986	5928.356	5981.070	6022.473	6055.847
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388	99.399
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345	27.229
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051
6	13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874
7	12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620
8	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814
9	10.561	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257
10	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691
17	8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682	3.593
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523	3.434
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368
21	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398	3.310
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346	3.258
23	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299	3.211
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168
25	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129
26	7.721	5.526	4.637	4.140	3.818	3.591	3.421	3.288	3.182	3.094
27	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256	3.149	3.062
28	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226	3.120	3.032
29	7.598	5.420	4.538	4.045	3.725	3.499	3.330	3.198	3.092	3.005
30	7.562	5.390	4.510	4.018	3.699	3.473	3.304	3.173	3.067	2.979
40	7.314	5.179	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801
60	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632
120	6.851	4.787	3.949	3.480	3.174	2.956	2.792	2.663	2.559	2.472
inf	6.635	4.605	3.782	3.319	3.017	2.802	2.639	2.511	2.407	2.321

Table 15.4: Critical Values of *F* Distribution (contd.)
(1% level of significance)

df2/df1	12	15	20	24	30	40	60	120	INF
1	6106.321	6157.285	6208.730	6234.631	6260.649	6286.782	6313.030	6339.391	6365.864
2	99.416	99.433	99.449	99.458	99.466	99.474	99.482	99.491	99.499
3	27.052	26.872	26.690	26.598	26.505	26.411	26.316	26.221	26.125
4	14.374	14.198	14.020	13.929	13.838	13.745	13.652	13.558	13.463
5	9.888	9.722	9.553	9.466	9.379	9.291	9.202	9.112	9.020
6	7.718	7.559	7.396	7.313	7.229	7.143	7.057	6.969	6.880
7	6.469	6.314	6.155	6.074	5.992	5.908	5.824	5.737	5.650
8	5.667	5.515	5.359	5.279	5.198	5.116	5.032	4.946	4.859
9	5.111	4.962	4.808	4.729	4.649	4.567	4.483	4.398	4.311
10	4.706	4.558	4.405	4.327	4.247	4.165	4.082	3.996	3.909
11	4.397	4.251	4.099	4.021	3.941	3.860	3.776	3.690	3.602
12	4.155	4.010	3.858	3.780	3.701	3.619	3.535	3.449	3.361
13	3.960	3.815	3.665	3.587	3.507	3.425	3.341	3.255	3.165
14	3.800	3.656	3.505	3.427	3.348	3.266	3.181	3.094	3.004
15	3.666	3.522	3.372	3.294	3.214	3.132	3.047	2.959	2.868
16	3.553	3.409	3.259	3.181	3.101	3.018	2.933	2.845	2.753
17	3.455	3.312	3.162	3.084	3.003	2.920	2.835	2.746	2.653
18	3.371	3.227	3.077	2.999	2.919	2.835	2.749	2.660	2.566
19	3.297	3.153	3.003	2.925	2.844	2.761	2.674	2.584	2.489
20	3.231	3.088	2.938	2.859	2.778	2.695	2.608	2.517	2.421
21	3.173	3.030	2.880	2.801	2.720	2.636	2.548	2.457	2.360
22	3.121	2.978	2.827	2.749	2.667	2.583	2.495	2.403	2.305
23	3.074	2.931	2.781	2.702	2.620	2.535	2.447	2.354	2.256
24	3.032	2.889	2.738	2.659	2.577	2.492	2.403	2.310	2.211
25	2.993	2.850	2.699	2.620	2.538	2.453	2.364	2.270	2.169
26	2.958	2.815	2.664	2.585	2.503	2.417	2.327	2.233	2.131
27	2.926	2.783	2.632	2.552	2.470	2.384	2.294	2.198	2.097
28	2.896	2.753	2.602	2.522	2.440	2.354	2.263	2.167	2.064
29	2.868	2.726	2.574	2.495	2.412	2.325	2.234	2.138	2.034
30	2.843	2.700	2.549	2.469	2.386	2.299	2.208	2.111	2.006
40	2.665	2.522	2.369	2.288	2.203	2.114	2.019	1.917	1.805
60	2.496	2.352	2.198	2.115	2.028	1.936	1.836	1.726	1.601
120	2.336	2.192	2.035	1.950	1.860	1.763	1.656	1.533	1.381
inf	2.185	2.039	1.878	1.791	1.696	1.592	1.473	1.325	1.000

Source: Tables 15.1 to 15.4 are adapted from the website http://www.statsoft.com/textbook/sttable.html (accessed on 23.09.2004)

UNIT 16 BASIC CONCEPTS OF SAMPLING

Structure

16.0 OBJECTIVES

After going through this unit you should be able to:

- explain the concepts of population, sample, parameter, statistic, estimator and estimate;
- distinguish between a census and a sample survey;
- explain the advantages of a sample survey;
- distinguish between sampling error and non-sampling error;
- explain the concept of sampling distribution; and
- explain the concept of standard error.

16.1 INTRODUCTION

We need data for the construction of national income accounts, input-output tables, various production indices, price indices and a host of other quantitative indicators. It is very clear that without the relevant data, we will not be able to formulate policy objectives for a complex economy like ours. In a sense, modern society is increasingly becoming an information society. In this society, various economic and social processes are represented by certain quantitative characteristics that require various kinds of information in the form of data.

The task of collecting data is getting increasingly complex and difficult. The total number of units to be consulted and investigated for the required information may be too large and our resources in terms of money, time or personnel may be limited. Moreover, obtaining error-free information from such a large-scale investigation makes the job even more daunting. As a result, very often we try to obtain the required information from a smaller group that is easier to handle and control. Here, however, it is important to ensure that this smaller group is truly representative of the entire collection of relevant units. The subject matter of sampling provides a mathematical theory for obtaining such kind of a representative group.

16.2 CENSUS AND SAMPLE SURVEY

In this Section, we will distinguish between the census and sampling methods of collecting data. We will try to explain the meaning and coverage of census survey and sample survey.

16.2.1 Population and Census

We have a collection of units relevant for a particular enquiry. A unit, in this connection, is an entity on which we can make observations according to a well-defined procedure. The entire collection of such units is called a *population* or *universe*. Thus, we may have a population of human beings, cattle, trees, prices, production, etc.

You can make out that a population can be finite or infinite. If the number of units is finite, it is a finite population and if the number of units is infinite, it is an example of an infinite population. Usually in practice, we are concerned with a finite population.

When an inquiry is based upon obtaining information from all the units of a population, the procedure is known as the *complete enumeration method* or the *census method*.

16.2.2 Sample and Sample Survey

When we have a collection of a part or section of the population, it is called a *sample*. A census, as we have seen earlier, is based upon obtaining information from every member of the population. However, in order to obtain information about certain characteristic of the population, we need not always resort to a census. In practice, we get quite satisfactory results by studying an appropriate sample from the population. The procedure of obtaining a sample is known as *sample survey*. In the case of a census, we examine the entire population; on the other hand, when we take a sample, we consider a representative fraction of the population and use the sample information to infer about the entire population.

16.3 SOME CONCEPTS

We explain below some of the concepts frequently used in sampling theory.

16.3.1 Parameter

In a statistical inquiry, our interest lies in one or more characteristics of the population. A measure of such a characteristic is called a *parameter*. For example, we may be interested in the mean income of the people of some region for a particular year. We may also like to know the standard deviation of these incomes of the people. Here, both mean and standard deviation are parameters.

Parameters are conventionally denoted by Greek alphabets. For example, the population mean can be denoted by μ and population standard deviation can be denoted by σ.

It is important to note that the value of a parameter is computed from all the population observations. Thus, the parameter 'mean income' is calculated from all the income figures of different individuals that constitute the population. Similarly, for the calculation of the parameter 'correlation coefficient of heights and weights', we require the values of all the pairs of heights and weights in a population. Thus, we can define a *parameter as a function of the population values.* If θ is a parameter that we want to obtain from the population values $X_1, X_2, \cdots X_N$, then

$\theta = f(X_1, X_2, \cdots X_N)$ •

16.3.2 Statistic

While discussing the census and the sample survey, we have seen that due to various constraints, sometimes it is difficult to obtain information about the whole population. In other words, it may not be always possible to compute a population parameter. In such situations, we try to get some idea about the parameter from the information obtained from a sample drawn from the population. This sample information is summarised in the form of a *statistic*. For example, sample mean or sample median or sample mode is called a statistic. Thus, a statistic is calculated from the values of the units that are included in the sample. So, a *statistic can be defined as a function of the sample values*. Conventionally, a statistic is denoted by an English alphabet. For example, the sample mean may be denoted by $\bar{x}$ and the sample standard deviation may be denoted by s. If T is a statistic that we want to obtain from the sample values $x_1, x_2, \cdots x_n$, then

$T = f(x_1, x_2, \cdots x_n)$

16.3.3 Estimator and Estimate

The basic purpose of a statistic is to estimate some population parameter. The procedure followed or the formula used to compute a statistic is called an *estimator* and the value of a statistic so computed is known as an *estimate*.

If we use the formula $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ for calculating a statistic, then this formula is an estimator. Next, if we use this formula and get $\bar{x} = 10$, then this '10' is an estimate.

16.4 NON-SAMPLING AND SAMPLING ERRORS

As mentioned above the basic purpose of sampling is to draw inferences about the population on the basis of the sample. For example, we have to find out the per capita income of a village. Due to shortage of time, money and personnel we do not undertake a complete census and opt for a sample survey. In this case it is very likely that the per capita income obtained from the sample is not equal to the actual per capita income of the village. This discrepancy could arise because of two reasons:

i) Since we are collecting data from only a part of the population (i.e., the sample selected by us), sample mean (per capita income in this case) is not equal to population mean. If at all both are equal, it is a rare coincidence! If we take sample mean as population mean we are committing an error called sampling error.

ii) A second source of error could arise because of wrong reporting or recording or tabulation or processing of data. This type of error is termed non-sampling error. Remember that non-sampling error, as its name suggests, has nothing to do with our sampling process. Wrong reporting or recording or processing of data can take place in a sample survey also.

We explain the sources of these errors below.

16.4.1 Non-Sampling Error

Various sources of non-sampling error are given below:

1) **Error due to measurement**

It is a well-known fact that precise measurement of any magnitude is not possible. If some individuals, for example, are asked to measure the length of a particular piece of cloth independently up to, say, two decimal points; we can be quite sure that their answers will not be the same. In fact, the measuring instrument itself may not have the same degree of accuracy.

In the context of sampling the respondents of an inquiry, for example, may not be able to provide the accurate data about their incomes. This may not be a problem with individuals earning fixed incomes in the form of wages and salaries. However, self-employed persons may not be able to do so.

2) **Error due to non-response**

Sometimes the required data are collected by mailing questionnaires to the respondents. Many of such respondents may return the questionnaires with incomplete answers or may not return them at all. This kind of an attitude may be due to:

a) the respondents are too casual to fill up the answers to the questions asked

b) they are not in a position to understand the questions, or

c) they may not like to disclose the information that has been sought.

We should note that the error due to non-response may also arise because of the possibility of the questionnaire being lost in transit.

If the data are collected through personal interviews, some of the reasons for the error due to non-response pointed out above may not arise. However, in that case this error may arise because some of the individuals:

Table 7.2
Bivariate Frequency Distribution of Occupation and Number of Children

Number of Children	*Occupation (t)*					*Total*
	Unemployed	*Unskilled Labour*	*Skilled Labour*	*Self-Employed*	*Professional*	
(X)	(1)	(2)	(3)	(4)	(5)	
0	10	15	10	12	5	52
1	35	25	17	18	25	120
2	22	33	45	40	43	183
3	11	40	48	58	30	187
4	8	22	12	11	8	61
5	3	11	18	8	1	41
Total	**89**	**146**	**150**	**147**	**112**	**644**

7.4.1 One Variable is Numerical Discrete

In Table 7.2 we present data where Occupation of Father (t) is a nominal variable and Number of Children (x) is a numerical variable. The variable Occupation has 5 categories. The variable Number of Children takes value from 0 to 5. Here the construction of the bivariate frequency table is quite similar to that discussed in the previous section.

Now let us look into the computation of marginal and conditional distributions. In this example, we may be interested in studying the effect of the 'independent variable' (Occupation) on the 'dependent variable' (Number of Children). As before we can compute the conditional and marginal distributions of the Number of Children for the various Occupation levels. But unlike in the examples above, Number of Children is a numerical and hence it is possible to work out summary measures like mean or mode from the conditional and marginal distributions. Thus, besides comparing the conditional distributions, we may compare the arithmetic means or modes of these conditional distributions. Recall the formulae we have studied for computing the *arithmetic mean* from a grouped frequency distribution (see Unit 4). Let f_{xt} be the frequency of xt^{th} cell where $X = 0,1,2,3,4,5$; $t = 1,2,3,4,5$; that is x represents the Number of children and t represents the Occupation class. Let f_t be the marginal frequency of t^{th} Occupation group that is $f_t = \sum_{x=0}^{5} f_{xt}$. These marginal frequencies are numbers in the last row of the table.

Then the formula for the *conditional arithmetic mean* of x in the t^{th} Occupation group is:

$$\bar{X}_t = \frac{1}{f_t}\sum_{x=0}^{5} f_{xt}X \qquad t = 1,2,3,4,5$$

For example, the arithmetic mean of conditional distribution for unemployed is

$$\frac{10\times 0+35\times 1+22\times 2+11\times 3+8\times 4+3\times 5}{89}=1.79$$

and for unskilled labour is

$$\frac{15\times 0+25\times 1+33\times 2+40\times 3+22\times 4+11\times 5}{146}=2.42.$$

Similarly you can check that the arithmetic means of conditional distributions for other occupation groups are: Skilled Labour: 2.59; Self-employed : 2.42; Professional : 2.12.

The arithmetic mean of the marginal distribution, i.e., for all occupations together, is $\frac{52\times0+120\times1+183\times2+187\times3+61\times4+41\times5}{644}=2.32$.

Let us work out the mode from this frequency distribution. As you know, mode is the value that occurs most frequently (see Unit 4). Using this definition the modes of the conditional distribution are:

Unemployed: 1; Unskilled Labour: 3; Skilled Labour: 3; Self-employed: 3; Professional: 2.

The mode from the marginal distribution, i.e., for all occupations together, is 3.

Note that, although the Number of Children is a variable which takes only whole numbers as values, it does not make sense to present means and such averages in decimals. In this case, if we round off the means they will all come to 2 (except for skilled labour) and we shall not be able to discern the difference between various occupation groups.

7.4.2 One Variable is Numerical Continuous

When one of the variables is continuous and if the number of observations is not too many, often data can be presented and computations required for statistical analysis be done using the data in the form in which it is collected. For instance, if we are investigating the annual salaries of Male and Female economists in a bank and if you have 15 observations, they can be presented in a simple form as in Table 7.3.

Table 7.3
Data on Annual Salaries (Rs.) of 15 Economists in a Bank by Gender

Male	Female
45120	80505
72580	75012
80912	60045
120100	40010
30042	35010
80045	—
81250	—
105505	—
111005	—
60123	—

However, if you have a large number of observations, it is inconvenient to present data in the form of simple tables. Suppose we have data on annual salaries of 709 officers employed in public and private sectors and we want to make a study of their salaries. Here we have a nominal variable 'Employees' with two categories depending upon whether they belong to private sector or public sector, and a numerical continuous variable 'Salaries' which can be divided into 14 class intervals. A bivariate frequency table as in Table 7.4 can be constructed.

Table 7.4
Frequency Distribution of Annual Salaries (Rs.) of Officers in Public and Private Sectors

Annual Income (000 Rs.)	*Mid-value (xj)*	*Number in Sector*		*Percent of Total*	
		Private	*Public*	*Private*	*Public*
45 - 50	47.5	84	0	13.6	0.0
50 - 55	52.5	31	11	5.0	12.2
55 - 60	57.5	135	12	21.8	13.3
60 - 65	62.5	115	12	18.6	13.3
65 - 70	67.5	73	15	11.8	16.7
70 - 75	72.5	77	8	12.4	8.9
75 - 80	77.5	31	5	5.0	5.6
80 - 85	82.5	13	5	2.1	5.6
85 - 90	87.5	18	7	2.9	7.8
90 - 95	92.5	32	3	5.2	3.3
95 - 100	97.5	4	8	0.6	8.9
100 - 105	102.5	2	3	0.3	3.3
105 - 110	107.5	1	1	0.2	1.1
110 - 115	112.5	3	0	0.5	0.0
	$f_t \rightarrow$	619	90	100.0	100.0

The last two columns in Table 7.4 give the frequency distributions (in percentage form) which help in making a comparison of the two sectors. As in the case of Table 7.2, a comparison of arithmetic means of the two salaries in the two sectors may also be useful. Let f_{jt} be the frequency of jt^{th} cell, where $j = 1, 2, ..., k$; $t = 1, 2, ..., l$. Let X_j be the mid-value of the j^{th} class of the x variable. In Table 7.4, $k = 14$; $l = 2$; the mid-value of the x variable (income) are given in second column of the table. Let $f_t = \sum_{j=1}^{k} f_{jt}$. These f_t's are marginal frequencies and appear in the last row of the table. The formula for the conditional arithmetic mean of x in the t^{th} sector is:

$$\bar{X}_t = \frac{1}{f_t} \sum_{j=1}^{k} f_{jt} X_j$$

The arithmetic means computed using this formula are:

Private Sector: 64.83 ('000 Rs.); Public Sector: 72.17 ('000 Rs.).

For certain purposes of statistical analysis, it is also useful to have the variances computed for each group. Recall from Unit 5 that the formula for variance from a grouped frequency distribution is:

$$\sigma_{xt}^2 = \frac{\sum_{j=1}^{k} f_{jt} X_j^2}{f_t} - \bar{X}_t^2$$

The variances computed by this formula are: Private Sector: 163.37
Public Sector: 232.74

7.4.3 Both Variables are Numerical Continuous

Now let us consider the case where both variables are continuous. If the number of observations are large, we may represent the bivariate frequency distribution

using class intervals for both the variables. We consider an example of this type in Table 7.5, where data have been obtained from a sample of 99 families and where y denotes family expenditure (Rs.) on entertainment in a year and x denotes total annual income (Rs.) of the family.

Table 7.5
Bivariate Frequency Distribution of Annual Family Income and Annual Family Expenditure on Entertainment

Expenditure on Entertainment (Xj)			*Annual Income ('00 Rs.)* (y_i)									
			25-80	80-135	135-190	190-245	245-300	300-355	355-410	410-465	465-520	520-575
			Mid-value									
('00 Rs.)	M.V. ↓	x_j' ↓	52.5	107.5	162.5	217.5	272.5	327.5	382.5	437.5	492.5	547.5
	$y_i' \rightarrow$		- 6	- 5	- 4	- 3	- 2	- 1	0	1	2	3
45 - 50	47.5	5	-	-	-	-	-	-	-	-	-	1
40 - 45	42.5	4	-	-	-	-	-	-	1	-	1	-
35 - 40	37.5	3	-	-	-	-	1	-	-	1	2	1
30 - 35	32.5	2	-	-	-	-	-	-	-	4	3	2
25 - 30	27.5	1	-	-	-	-	3	4	4	5	6	1
20 - 25	22.5	0	-	-	-	-	-	5	7	12	1	1
15 - 20	17.5	- 1	-	-	-	1	4	8	1	1	-	-
10 - 15	12.5	- 2	-	1	3	1	1	-	-	-	-	-
5 - 10	7.5	- 3	-	4	4	-	-	-	-	-	-	-
0 - 5	2.5	- 4	4	-	-	-	-	-	-	-	-	-

In a situation of this sort one may like to examine the conditional distributions of the dependent variable Y (expenditure on entertainment) in each class of the independent variable X (income). However, in this table, since the overall sample size is small and consequently many cell frequencies are zero, such distributions are not useful. One may, however, calculate arithmetic means and other summary measures from the conditional distribution of expenditure on entertainment for each income class, besides computing these measures from the marginal distribution of expenditure on entertainment. The means of the conditional distributions computed using the formula described above, are given in Table 7.6.

Table 7.6
Means of Annual Family Expenditure on Entertainment for each Class of Annual Family Income

Income (in '00 Rs.)	*Average Expenditure on Entertainment (in '00 Rs.)*
25 - 80	2.50
80 - 135	8.50
135 - 190	9.65
190 - 245	15.00
245 - 300	22.50
300 - 355	21.30
355 - 410	25.19
410 - 465	25.76
465 - 520	30.96
520 - 575	33.33

Check Your Progress 2

1) Suppose you have data on the country of citizenship and the amount of money spent by a foreign tourist in India, obtained on the basis of a survey of 2000 tourist at the time of their exit. Indicate how you may present the result in the form of a table.

.....

.....

.....

.....

.....

2) The following table gives the bivariate frequency distribution of the monthly family expenditure on food in samples of three localities of a city. Find the conditional distribution of the expenditure in each of the localities and the conditional mean and variance therefrom.

Frequency Distribution of Monthly Family Expenditure on Food in Three Localities

Expenditure on Food (Rs.)	*Locality*		
	A	*B*	*C*
< 250	122	2	0
251 - 500	100	5	1
501 - 750	75	11	3
751 - 1000	59	25	18
1001 - 1250	34	39	27
1251 - 1500	23	56	56
1501 - 2000	12	67	89
> 2000	0	45	114

.....

.....

.....

.....

.....

.....

.....

7.5 LET US SUM UP

There are three types of variables: nominal, ordinal and numerical. In the case of nominal variables we can categorise them. On the other hand, in the case of ordinal variables we can order them. Numerical variables can be discrete or continuous and these take quantitative values.

In this unit we presented the above types of variables in the form of bivariate frequency distributions. Also we calculated marginal and conditional distributions of these variables.

7.6 KEY WORDS

Bivariate Data : Data in which there are two measurements for each items. For example, the income and years of education of each person studied.

Marginal Distributions : It refers to the distribution of row totals or column totals in two-way or multi-way tables.

7.7 SOME USEFUL BOOKS

Nagar, A.L. and R.K. Das, 1989, *Basic Statistics*, Oxford University Press, Delhi.

Goon, A.M., M.K. Gupta and B. Dasgupta, 1987, *Basic Statistics*, The World Press Pvt. Ltd., Calcutta.

7.8 ANSWERS/ HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) i) Nominal, ii) ordinal, iii) numerical, iv) numerical, v) nominal.
2) Go through Section 7.3 and answer.

Check Your Progress 2

1) Amount of money spent is a numerical continuous variable while country of citizenship is a nominal variable. Thus you should form class intervals for amount of money spent and categories for country of citizenship. Accordingly you can construct a bivariate frequency distribution where the number of tourists falling under each cell are reported.
2) Refer to Section 7.3.

UNIT 17 SAMPLING PROCEDURE

Structure

17.0 OBJECTIVES

On the completion of this Unit, you should be able to:

- explain different methods of drawing a sample;
- use random number tables to draw a sample; and
- determine the sample size.

17.1 INTRODUCTION

In Unit 2 of this course, you have learned about different types of data, namely, primary data and secondary data. In that Unit, we also discussed the use of different survey techniques like face-to-face interview, telephone survey, postal survey, internet survey, etc. in collecting primary data. In Unit 15, you also have learned the meaning of sampling, advantages of sampling, and sampling error.

In statistics, we often rely on a sample (that is, a small subset of a larger set of data) to draw inferences about the population (that is, the larger set of data). For example, you are interested to know the voting behaviour of Delhi people in the next election. Who will you ask? Naturally, it is not possible for you to ask every single Delhi voter how he or she is likely to vote. Instead, you querry a relatively small number of Delhi voters and draw inferences about entire Delhi from their responses. In this case total voters of Delhi constitute the population and the voters actually querried constitute your sample.

Ideally, the characteristics of a sample should reflect the characteristics of the population from which it is drawn. In such cases, the inferences drawn from a sample are probably applicable to the entire population.

In this unit you will learn how to draw a sample under different population characteristics and how to determine the sample size.

17.2 SAMPLING PROCESS

In conducting a sample survey, the sampling process determines which sampling units will be included in the survey. Sampling makes data process more manageable and affordable. It enables the population characteristics to be inferred with minimal errors on the basis of the sample. The sampling process includes defining the target population from which we draw the sample, identification of the sampling frame, selection of the sampling method, and selection of the sampling units.

1) **Survey Objectives :** The sample survey begins with the specification of the objectives. We should have a clear and un-ambiguous idea of the objectives of the survey. Because all other steps— target population, sampling frame, sampling procedure, etc.— are designed according to survey objectives.
2) **Questionnaire Design:** Keeping the objectives of the survey in view we are required to design a questionnaire. We have already learnt the major steps involved in designing of a questionnaire in Unit 2 of this course. In addition to the questionnaire we need to develop training documents for the investigators, particularly when the sample survey is conducted at a larger scale involving a number of investigators.
3) **Defining the Target Population:** To draw the samples from a population we must know the target population about which conclusions are to be drawn. The target population is also referred to as the universe. The target population is the group about which we wish to generalize or make inferences from the sample. For example, you want to conduct a sample survey on the family planning methods used by eligible couples in Delhi. Here, all those couples in Delhi in the reproductive age group form the target population.
4) **Identifying Sampling Frame**: The sampling frame is a list of cases from the target population. The sampling frame is the actual operational definition of

the target population. In our earlier example of eligible couples in Delhi using family planning methods, all the people in the reproductive age group form the sampling frame. Many times we may not be able to list all the cases in the target population for some reason or other. For example, we want to list the people for a survey based on telephone directory. In this case, certainly those people who do not have telephone numbers in the directory will be excluded from the listing. This type of error is called sampling frame error.

5) **Selecting Sampling Procedure**: Once the sampling frame is identified, we select appropriate sampling procedure to select the sample for the survey. We will discuss various sampling procedures in detail in the next section of this Unit.

6) **Selecting the Sampling Units**: Sampling units are those cases from the sampling frame which are included in the sample by using appropriate sampling procedure. Essentially, a sampling unit is the case on which data is collected. For example, you may decide to take 1000 sampling units from the sampling frame (consisting of all the reproductive age group people in Delhi) for your sample survey.

7) **Survey data Processing:** After selection of sampling units the next step is data collection and processing. We need to check the incomplete questionnaires and edit or cross-check the responses wherever there is a doubt. Data entry and tabulation follows.

8) **Analysis of Data:** The next step in the sequence is analysis of data. Keeping in view our requirements we analyse the data by using various statistical tools.

9) **Publication and Dissemination of Results:** On the basis of data analysis we prepare technical and research reports. Finally, the socio-economic results of the survey and their implications are discussed in seminars.

17.3 TYPES OF SAMPLING

A critical step in selecting a sample is determining the method of selecting the sample from the population. Broadly, there are two principal types of sampling procedures: (a) Probability (or random) sampling, (b) Non-probability (or non-random) sampling.

17.3.1 Probability Sampling

In probability sampling, all the units in the population have a chance of being included in the sample. The probability sampling is also called random sampling. The probability sampling is an objective procedure in which the probability of selection of a sampling unit is known in advance. Here the probability of including each unit in the sample is non-zero.

The advantage of using probability sampling procedure is that we do not play a role in determining which specific population units are chosen to be included in the sample. The probability sampling procedure specifies an objective scheme for choosing the sample — independent of personal preferences or biases. Because of this reason probability sampling procedures take a lot of time to ensure that each element has non-zero probability of getting included in the sample. Another advantage of using probability sampling procedure is that it is possible to quantify the magnitude of the sampling error in inference made. The estimation of the expected error helps us in many situations in building up confidence in the inference. When a probability sampling procedure is properly applied, it contains no bias and is therefore relatively representative of the population. Practically, we can never

be 100% certain that the results measured from the sample are also true for the population. For this reason, often it is enough for us to know that the risk of a deviation from a population is 1% or 5% as we will learn while calculating *confidence intervals* (see Unit 18).

17.3.2 Non-Probability Sampling

The non-probability sampling procedure relies more on human judgment than objectiveness of the procedure. The non-probability sampling procedure is also often referred to as non-random sampling procedure. Judgment plays a major role in determining the sampling units to be included. In non-probability sampling the probability of selection for each population unit is unknown before hand. Therefore, we cannot ensure that the sample is representative. Also, we cannot determine the probability of selection. Based on the experience, expertise and performance of the researcher or selecting authority the units to be included in the sample are identified. At times this may involve personal bias also.

17.4 SELECTION OF A SIMPLE RANDOM SAMPLE

Simple random sampling (SRS) is the basic sampling procedure where we select a sample from the population. In this procedure each unit is chosen entirely through chance mechanism and each unit of the population has an equal chance of being included in the sample. Every possible sample of a given size has the same chance of selection. That is, each unit of the population is equally likely to be chosen at any stage in the sampling process and selection of one unit should in no way influence the selection of another unit in the population.

Simple random sampling should be used with a homogeneous population. That is, all the units in a population should possess the same attributes that we are interested in measuring. The characteristics of homogeneity may include age, sex, income, social status, geographical region, etc.

There are two most commonly used methods to extract a simple random sample. The first is lottery method and the second is a random numbers selection method. Irrespective of which method we decide to use, every element in the sampling frame should be assigned an identifying number.

17.4.1 Lottery Method

The simplest method of selecting a random sample is drawing lottery. In this method, the unit identification numbers are placed in a container and mixed together. Finally, someone draws out numbers at random from the container until the desired sample size is attained. Suppose, we want to select n sample units out of N population units. We assign the numbers 1 to N; one number to each of the population unit; and write these numbers on N slips. The slips are made as homogeneous as possible in shape, size, colour, etc. These slips are then put in a container and thoroughly shuffled. Finally, n slips are drawn one by one. The n units corresponding to numbers on slips drawn, will constitute a random sample.

Example17.1

Let us assume that you are doing some research with a bank branch that wishes to assess customers views on the quality of service in the bank branch. You are asked to select 100 customers as the sample using simple random sampling

procedure with lottery method. First, you have to get the sampling frame organized. For this, you will go through the bank records to identify the account holders. You then assign the serial numbers to all the account holders. Suppose there are 1000 account holders and you want to draw 100/1000=10% sample. You could print the serial numbers, tear them into separate strips, put the strips in a container, mix them up real good, close your eyes and pull out the first 100 strips.

The disadvantage of the lottery method is that it would be tedious and the quality of the sample would depend on how thoroughly you mixed them up and how randomly you picked them. Also, as the population size increases, it becomes more and more difficult to draw samples using lottery method.

17.4.2 Random Numbers Table Method

The random numbers are a collection of digits generated through a probability mechanism. The random numbers have the following properties:

a) The probability that each digit (0,1,2,3,4,5,6,7,8 or 9) will appear at any place is the same, that is 1/10.

b) The occurrence of any two digits in any two places is independent of each other.

In this method each unit in the population is assigned a unique number in a sequence. To draw the sample we use a table of random numbers. You can find the random number table (RNT), among other places, in Fisher and Yates (1963): Statistical *Tables for Biological, Agricultural and Medical Research.* An example of a random numbers table is shown in Table 17.1.

Table 17.1: Random Numbers Table

39634	62349	74088	65564	16379	19713	39153	69459	17986	24537
14595	35050	40469	27478	44526	67331	93365	54526	22356	93208
30734	71571	83722	79712	25775	65178	07763	82928	31131	30196
64628	89126	91254	24090	25752	03091	39411	73146	06089	15630
42831	95113	43511	42082	15140	34733	68076	18292	69486	80468
80583	70361	41047	26792	78466	03395	17635	09697	82447	31405
00209	90404	99457	72570	42194	49043	24330	14939	09865	45906
05409	20830	01911	60767	55248	79253	12317	84120	77772	50103
95836	22530	91785	80210	34361	52228	33869	94332	83868	61672
65358	70469	87149	89509	72176	18103	55169	79954	72002	20582
72249	04037	36192	40221	14918	53437	60571	40995	55006	10694
41692	40581	93050	48734	34652	41577	04631	49184	39295	81776
61885	50796	96822	82002	07973	52925	75467	86013	98072	91942
48917	48129	48624	48248	91465	54898	61220	18721	67387	66575
88378	84299	12193	03785	49314	39761	99132	28775	45276	91816
77800	25734	09801	92087	02955	12872	89848	48579	06028	13827
24028	03405	01178	06316	81916	40170	53665	87202	88638	47121
86558	84750	43994	01760	96205	27937	45416	71964	52261	30781
78545	49201	05329	14182	10971	90472	44682	39304	19819	55799
14969	64623	82780	35686	30941	14622	04126	25498	95452	63937
58697	31973	06303	94202	62287	56164	79157	98375	24558	99241
38449	46438	91579	01907	72146	05764	22400	94490	49833	09258
62134	87244	73348	80114	78490	64735	31010	66975	28652	36166

Contd..

72749	13347	65030	26128	49067	27904	49953	74674	94617	13317
81638	36566	42709	33717	59943	12027	46547	61303	46699	76243
46574	79670	10342	89543	75030	23428	29541	32501	89422	87474
11873	57196	32209	67663	07990	12288	59245	83638	23642	61715
13862	72778	09949	23096	01791	19472	14634	31690	36602	62943
08312	27886	82321	28666	72998	22514	51054	22940	31842	54245
11071	44430	94664	91294	35163	05494	32882	23904	41340	61185
82509	11842	86963	50307	07510	32545	90717	46856	86079	13769
07426	67341	80314	58910	93948	85738	69444	09370	58194	28207
57696	25592	91221	95386	15857	84645	89659	80535	93233	82798
08074	89810	48521	90740	02687	83117	74920	25954	99629	78978
20128	53721	01518	40699	20849	04710	38989	91322	56057	58573
00190	27157	83208	79446	92987	61357	38752	55424	94518	45205
23798	55425	32454	34611	39605	39981	74691	40836	30812	38563
85306	57995	68222	39055	43890	36956	84861	63624	04961	55439
99719	36036	74274	53901	34643	06157	89500	57514	93977	42403
95970	81452	48873	00784	58347	40269	11880	43395	28249	38743
56651	91460	92462	98566	72062	18556	55052	47614	80044	60015
71499	80220	35750	67337	47556	55272	55249	79100	34014	17037
66660	78443	47545	70736	65419	77489	70831	73237	14970	23129
35483	84563	79956	88618	54619	24853	59783	47537	88822	47227
09262	25041	57862	19203	86103	02800	23198	70639	43757	52064

Source: Adapted from Table of Random Numbers at http://www.mrs.umn.edu/~sungurea/introstat/public/instruction/ranbox/randomnumbersII.html

The above random numbers table contains 450 (5 digit) random numbers.

17.4.3 Steps in the Use of RNT

We need to follow the following steps while selecting a SRS by using RNT.

1) Determine the population size (N).
2) Determine the sample size (n).
3) List all the units of the population. Assign the numbers in a serial order. Suppose there are 100 units in the population, assign the serial numbers from 00 to 99.
4) Determine the starting point of selecting the sample by picking up a page from the random number tables and dropping your finger on a number in the page blindly.
5) Choose the direction in which you want to read the numbers (say from left to right or right to left or top to bottom or bottom to top).
6) Suppose you are looking for two digit numbers (00 to 99) you may not get these numbers by direct reading from the tables since they are 5 digit numbers (see Table 17.1). You can either look at the last two digits or first two digits of the numbers. For example, if the 5 digit number you have chosen is 54245 (that is, the number in the 29th row and 10th column of the random number table given at Table 171). Then, the two digit number will be 45 if you chose the last two digits of the number.
7) Look only at the numbers assigned to each population unit. If the number represents one of the unit of the population it becomes part of the sample. Suppose you want to select 10 sample units, the other numbers you will be choosing are 71(11071), 30(44430), 64(94664), 94(91294), 63(35163),

82(32882), 04(23904), 40(41340), 85(61185). Observe that you have omitted 94(05494) since you have already chosen this number.

8) Once a number is chosen, do not select it again.

9) If you reach the end point of the table before obtaining the required sample, pick another starting point in the random number table and select the remaining units for the sample.

Example 17.2

Suppose you have to select 100 account holders as a sample out of total 1000 account holders in the population using random numbers table. Here, first you assign each account holder a number from 000 to 999. To draw a sample of 100 account holders, you need to find 100 three digit numbers in the range 000 to 999. Pick up any row or column in the random numbers table given in Table 171. Suppose you have selected the fourth row and first column as starting point to draw the sample, the first digit number is 628(64628) if you chose last 3 digits as the number for your purpose. Here, you read the last 3 digits of the number. If the number is within the range (000 to 999) include the number in the sample. Otherwise skip the number and read the next number in some identified direction. If a number is already selected omit it. In this example since you have started with fourth row and first column and moving from left to right direction the following 100 numbers are selected for the sample.

628	126	254	090	752	091	411	146	089	630
831	113	511	082	140	733	076	292	486	468
583	361	047	792	466	395	635	697	447	405
209	404	457	570	194	043	330	939	865	906
409	830	911	767	248	253	317	120	772	103
836	530	785	210	228	869	332	868	672	358
469	149	509	176	169	954	002	582	249	037
192	221	918	437	571	995	006	694	692	581
050	734	652	577	631	184	295	776	885	796
822	973	822	467	013	072	942	917	129	624

If the number of units in the population is very large, neither of the above two methods is feasible. These days by using a computer we can select a random sample in a much easier way. There are many computer program which can generate a series of random numbers if we have the units of the population listed in a computer.

We will explain one way of selecting a sample using computer generated random numbers. In our example, let us assume we can copy and paste the list of account holders into a column in an EXCEL spreadsheet. Then, in the column right to it we paste the function =RAND() which is EXCEL's way of putting a random number between 0 and 1 in the cells. Then, all we have to do is take the first 100 names in the sorted list. The entire process takes a minute if we are familiar with using EXCEL program in computer.

17.4.4 Advantages of SRS

1) In simple random sample we assure population units to be homogeneous and thus do not require additional information on the characteristics of the population.

2) Using simple random sampling we can select an unbiased sample. This is because, in SRS the chance of including each unit of the population in the sample is equal. The bias due to human preferences is completely eliminated.
3) Through estimation of sampling error we can assess the accuracy of the results.
4) If the population size is not too large, simple random sampling is a simple and easily implementable sampling procedure.

17.4.5 Limitations of SRS

1) The greatest limitation of simple random sampling is that if the population size is too large then we need to spend a lot of time in listing the units of the population.
2) The simple random sampling procedure will be efficient only when we have a homogeneous population. Suppose we have a population with characteristics such as gender, age, social status, etc. Then, we need a larger sample size to accommodate a representative sample with all those characteristics of the population units. A better way to tackle this issue is to use stratified sampling procedure which you will learn later in this unit.

17.5 SELECTION OF SYSTEMATIC RANDOM SAMPLE

The systematic random sampling procedure is somewhat similar to the simple random sampling procedure. In this sampling procedure, we select a starting point at random and then systematically select the sample units from the population units at a specified sampling interval.

The starting point and the sampling interval are based on the required sample size. The sampling interval will be represented as K. The selection of a sample using systematic random sampling procedure is very simple. Suppose the population consists of N units and you have decided to select a sample of n units using systematic random sampling procedure. Follow the following steps.

1) Number the units in the population from 1 to N (suppose you have 1000 units).
2) Decide the sample size n you need (suppose you want to select 100 units).
3) Determine the sampling interval by dividing the population by the sample size. $K = N/n$ = the interval size (here K=1000/100 = 10).
4) Select a unit at random from the first K units (1 to K) (suppose you have selected unit number 5 as your first sample unit).
5) Then select the subsequent sample units by adding K to the previous unit (the subsequent samples will be 15 (=5+10), 25 (=15+10), …, 995 (=985+10).

Example 17.3

From a population consisting of 500 units draw a sample of 60 units using systematic random sampling procedure.

To use systematic random sampling, the first thing we need to do is listing of the population units in a random order by giving numbers from 1 to 500. The sampling interval is K=500/60 = 8.3 or say 8. Now we select the first sample unit at random from the first 8 population units. Suppose the first unit selected is 5. The subsequent

sample units selected are : 13, 21, 29, 37.....477. Therefore, following are the population units selected in the sample.

5	13	21	29	37	45	53	61	69	77
85	93	101	109	117	125	133	141	149	157
165	173	181	189	197	205	213	221	229	237
245	253	261	269	277	285	293	301	309	317
325	333	341	349	357	365	373	381	389	397
405	413	429	429	437	445	453	461	469	477

Thus, in the systematic random sampling procedure, the first sample unit is selected at random and this sample unit in turn determines the subsequent sample units to be selected. However, it is essential that the units in the population are randomly ordered. In certain cases we prefer using systematic random sampling procedure to simple random sampling procedure because it is easier to select sample units. For example, if you want to find out the yield of coconut trees in a field, select a tree at random, other trees are automatically selected at a gap equivalent to sampling interval.

17.5.1 Advantages of Systematic Random Sampling

a) The main advantage of using systematic random sample is that the time taken and work involved in systematic random sampling is less than simple random sampling procedure. It is frequently used in exit polls on voting behaviour and obtaining the opinions and views of consumers in marketing research.

b) The other advantage of systematic random sampling procedure is that this method can be used even when no formal list of the population units is available. For example, suppose if we are interested in knowing the opinion of consumers on improving the services offered by a bank, we may simply choose every k^{th} account holder visiting a bank branch, provided that we know how many account holders are there : (1) For example, there are 2000 account holders in the population and we want to have 200 account holders as sample size. Then, K=2000/200=10) and we select every 10^{th} account holder visiting the bank.

17.5.2 Disadvantages of Systematic Random Sampling

a) The main disadvantage of systematic random sampling procedure is that if there is a periodicity in the occurrence of units of a population, the use of systematic random sampling procedure gives a highly unrepresentative sample. For example, suppose you are interested in obtaining the views/opinions of consumers of a store in your locality. You may arrange all the consumers of the store according to their date of visit and start selecting a sample of customers who visit the store on 1^{st} of every month. You know that the 1^{st} day of every month cannot be representative of the whole month.

b) The other disadvantage of systematic random sampling procedure is that every unit of the population does not have an equal chance of being selected. Rather the selection of population units in the sample depends on the initial unit of selection. Regardless of how we select the first unit of the sample, subsequent units are automatically determined. This lacks complete randomness.

17.6 SELECTION OF STRATIFIED RANDOM SAMPLE

In some cases the population may not be homogenous, that is, all the units may not be equal with respect to the characteristic we intent to survey. The characteristics of the population under study may be male/female, rural/urban, literate/illiterate, high income/low income groups, etc. In situations where these units vary widely, the simple random sampling procedure or the systematic random sampling procedure will not provide us with a representative sample. In such situations by using stratified random sampling we can obtain a representative sample.

In stratified sampling, we divide the population into different strata in such a way that units are homogenous within each stratum. Moreover, each stratum is different. Suppose we want to stratify the population on the basis of gender distribution then we list the population units separately according to males or females. Subsequently, we decide the sample size to be drawn from each stratum. There are two approaches to decide the sample size from each stratum. These are: (a) proportional stratified sample, and (b) disproportional stratified sample. We will discuss these two procedures below.

17.6.1 Proportional Stratified Sample

When we take a sample from a population with several strata, we require to take samples from each stratum. Such sample could be in proportion of the stratum population size to the total population size. Suppose we divide the population (N) into K non-overlapping strata N_1, N_2, N_3..... N_K such that $N_1+N_2+N_3+.....+N_K = N$. We decide to draw a sample of the size n. Then the sample proportions of different strata are given by:

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \frac{n_3}{N_3} = \cdots = \frac{n_{12}}{N_K}$$

Example 17.4

Suppose we want to draw a sample of 200 units from a population consisting of 1000 units. The population is heterogeneous in nature in terms of high income or low income and rural or urban. The strata population sizes are given as follows:

High income - urban	=	200
Low income - urban	=	400
High income- rural	=	100
Low income-rural	=	300

To have a representative sample each stratum in the sample should represent the corresponding stratum in the population. For this we should take a different sample size from each stratum depending upon the stratum size. The deciding factor in each of the stratum is same as the proportion of total sample to the population. In our example, to have a sample of 200 units, the proportion of sample to the population in each stratum is

$$\frac{n}{N} = \frac{200}{1000} = 0.2$$

Observe that we are considering the same proportion for each stratum. Then the sample from each stratum will be as follows:

Stratum Category	Stratum Population size(N_i)	Sample to population proportion	Stratum sample size
(1)	(2)	(3)	(4)=(2)×(3)
High income - urban	200	0.2	40
Low income -urban	400	0.2	80
High income-rural	100	0.2	20
Low income-rural	300	0.2	60
Total	1000	0.2	200

There are several advantages of stratified sampling over simple random sampling. The stratified sampling ensures sample representation of not only the entire population, but also each stratum. This is important where the stratum size is small. Moreover, stratified sampling generally has more statistical precision than simple random sampling.

17.6.2 Disproportional Stratified Sampling

In proportional stratified sampling, we assumed that each stratum in the population is homogeneous. Consequently, we expect that variability within stratum is lower than the variability for the population as a whole. On the other hand, if the variability within each stratum is not small then we use disproportional stratified sampling. In disproportional stratified sampling, the strata allocation is based on size and variability (that is, the standard deviation of the characteristic under study). In this procedure a larger sample is drawn from the stratum having higher variability. This procedure is also sometimes called double weighing scheme and provides the most efficient sample and most precise/reliable estimates for a given sample size. The only requirement is that we should have knowledge/estimate of the standard deviation of the characteristic under study within each stratum.

Follow the steps given below for using disproportional stratified sampling.

1) Divide the population into strata based on the chosen characteristic (example, Rural/Urban, Male/Female, etc.)
2) The number of units taken from each stratum is directly proportional to the relative size of the stratum and standard deviation s_i of the characteristic under consideration. Suppose, if $s_1, s_2, s_3, \ldots s_k$ are the standard deviations of k strata and $P_1, P_2, P_3, \ldots P_k$ are the stratum proportions to the total population, and $n\ (= n_1 \div n_2 + \ldots\ldots + n_k)$ is the sample size required. Then the stratum sample size using disproportional stratified sampling procedure is

$$n_i = \frac{P_i \times s_i \times n}{\sum P_i s_i}$$

3) Choose the sample from each stratum using either simple random sampling or systematic random sampling.

Let us go back to Example 17.4, where we have divided the population into 4 strata. We observe that there are small number of households in high income strata and large number of households in low income strata. Assume that the variance of income among higher income groups is higher than the variance among the lower income groups. Therefore, in order to avoid under-representation of higher income groups in the sample, a disproportional sample is taken in each stratum. That means, if the variability within the stratum is higher, we must have larger sample size of

that stratum to increase the precision of the estimates. Similarly, if the variability within the stratum is lower, we must have smaller sample size of that stratum. That is, higher the stratum variance larger the stratum sample size and lower the stratum variance smaller the sample size. This is in addition to the fact that larger stratum size requires a larger sample size.

Example 17.5

Consider Example 17.4 again. Suppose the stratum variances are given as follows:

Stratum	Variance (s^2)
High income urban	6.5
Low income urban	2.5
High income rural	4.5
Low income rural	2.0

Use the disproportional stratified sampling procedure to chose a sample of size 200 from the four strata.

For this example the disproportional sample size for each stratum is given below:

Stratum	Stratum population	Stratum population proportion (P_i)	Stratum Variance (s_i^2)	Stratum standard deviation ($s_i = \sqrt{s_i^2}$)	$P_i \times s_i$	Sample size $\frac{P_i \times s_i \times n}{\sum P_i s_i}$
High income-urban	200	0.20	6.5	2.5	0.50	56
Low income-urban	400	0.40	2.5	1.6	0.64	72
High income-rural	100	0.10	4.5	2.1	0.21	24
Low income-rural	300	0.30	2.0	1.4	0.42	47
Total	1000				1.77	200

17.6.3 Advantages of Stratified Sampling

a) In stratified random sampling the sample is drawn from each stratum of the population. Therefore, the stratified random sampling procedure is more representative.

b) The stratified random sampling procedure is more precise than simple random sampling. Therefore, to a great extent this procedure avoids sample selection bias.

c) As we have seen in simple random sampling and systematic random sampling procedures, when there is heterogeneity of population we need to have a large sample size to have a fairly representative sample. However, in stratified random sampling this objective can be achieved with a smaller sample size. This saves a lot of time, money and other resources for data collection.

17.6.4 Disadvantages of Stratified Sampling

a) The main disadvantage of stratified random sampling procedure is that we need a detailed knowledge of the distribution of the characteristics in the population.

If we cannot accurately identify the homogeneous groups, it is better to use simple random sample since improper stratification can lead to serious error.

b) The other disadvantage of stratified random sampling is that we need to prepare a list of population units for each stratum separately. As the list of population units may not be readily available for each characteristic, the preparation of lists may be a very difficult task.

17.7 SELECTION OF A CLUSTER SAMPLE

Very often population units are spread over a vast geographical area. In that case collection of data through simple random sampling requires a lot of time, money and manpower as we have to cover the entire geographical area for collecting data on the selected units. Imagine taking a sample of respondents spread all over Uttar Pradesh in order to conduct personal interviews. Using simple random sample, the respondents will be spread all over the state and you have to travel and spend lot of money meeting the respondents. In such situation cluster sampling will be much useful.

The basic principles of cluster sampling are:

i) The differences or variability within a cluster should be as large as possible. As far as possible the variability within each cluster should be the same as that of population.

ii) The variability between clusters should be as small as possible. Once the clusters are selected, all the units in the selected clusters are included in the sample for obtaining data.

In cluster sampling we divide the population into groups called clusters. Then we select a sample of clusters using a simple random sampling. The population units in each of the clusters are assumed to be as heterogeneous as those in the total population. That is, each cluster itself is a representative of the population.

17.7.1 Steps in Cluster Sampling

In cluster sampling, we follow the steps given below:

1) Divide the population into a number of clusters.
2) Determine the number of clusters needed for your sample.
3) Randomly select the sample of clusters.
4) Survey all units within the sampled clusters.

Suppose the division of clusters is based on the geographical boundaries of the population, then it is called *area sampling*. You have observed that in the case of cluster sampling the clusters are selected using random sampling method. Subsequently all the population units within each sampled cluster are included in the sample. Suppose instead of including all the population units within each selected cluster you chose to include only a sample of units within each cluster. Then you can clearly understand that there are two stages.

In the first stage you select the clusters and in the second stage you select the sample units within each sampled clusters. This sampling procedure is called *two-stage sampling*. Here, the clusters are called primary units and the units within the sampled clusters are called secondary units.

Example 17.6

Suppose we are interested in finding the options of ATM customers of a Bank in Uttar Pradesh state. We can divide the state into say 30 clusters (may be we can consider district as a unit and include one or two districts in one cluster). Here, we assume that each of these clusters will represent the opinions of the ATM customers of Uttar Pradesh as a whole. We then select a sample of clusters and obtain the opinion of all the ATM customers in each of the cluster.

17.7.2 Advantages of Cluster Sampling

a) The main advantage of cluster sampling is that it takes less travel time and related data collection costs.

b) Since the researcher need not cover all the clusters and only sample of clusters are covered, it is a more practical method which facilitates fieldwork.

17.7.3 Disadvantages of Cluster Sampling

a) In cluster sampling we assume that each cluster represents the heterogeneity of the population units of all clusters. However, this assumption may not be true in many cases, because often the tendency is that the units in the clusters are more homogeneous than the units of the entire population. That means it is difficult to form heterogeneous clusters.

b) The cluster sampling has a lower sampling efficiency for a given sample size than random sampling and stratified sampling. This method is cost effective but not statistically efficient.

17.8 MULTISTAGE SAMPLING

We have seen in cluster sampling that when we select a sample instead of covering all the units from each cluster, we call it two-stage sampling. The multistage sampling is as extension of two-stage sampling.

The four methods we have covered so far, namely, (a) simple random sampling, (b) systematic random sampling, (c) stratified sampling, and (d) cluster sampling are the simplest probability (or random) sampling procedures. However, in real-life, we use sampling methods that are more complex than the above four methods. The basic principle in multistage sampling is that we can combine these simple methods in a variety of useful ways to address our sampling needs. We call it multistage sampling when we combine two or more of the above sampling methods.

Example 17.7

Consider the case of interviewing school students in Haryana in order to grade the schools according to socio-economic background of the parents. For this problem, in the first stage we need to apply cluster sampling. We divide the state of Haryana into a number of clusters, say districts. Then we select a sample of districts (clusters) using simple random sampling method. In the second stage we divide the schools using stratified sampling. Here the strata may be government schools, government-aided schools, central schools, and public schools. We select a sample of schools in each stratum using either a simple random sampling or a systematic random sampling. In the third stage we again use simple random sampling and select a sample of classes in each sampled school for face-to-face interviews with the students. In the fourth stage of sampling we consider selecting

a sample of students from each sampled class using simple random sampling or systematic random sampling.

In multi stage sampling it is possible to consider as many stages as necessary to achieve a representative sample. In each stage a suitable method of sampling is used. Each stage results in a reduction of the sample size.

Advantages

a) Multistage sampling procedure provides cost gains by reducing the data collection costs.

b) Multistage sampling is more flexible and allows us to use different sampling procedures in different stages of sampling.

c) If the population is spread over a very wide geographical area, multistage sampling is the most appropriate sampling method.

Disadvantages

If the sampling units selected at different stages are not representative, multistage sampling becomes less precise and less efficient.

Check Your Progress 1

1) Which of the following is a procedure of selecting samples from a population?
   a) Random sampling
   b) Non-random sampling
   c) Stratified sampling
   d) All the above

2) Suppose you are applying a stratified random sampling procedure on a population. How do you make your sample selection?
   a) Select at random an equal number of units from each stratum
   b) Draw equal number of units from each stratum and weigh the results
   c) Select the sample at random from each stratum proportional to the population
   d) b) and c)
   e) a) and c)

3) Say whether the following statements are true or false.
   a) A sampling procedure that selects units from a population at uniform intervals is called simple random sampling.
   b) A sampling procedure that divides the population into well-defined groups from which random samples are drawn is known as stratified sampling.

4) A population is made up of groups that have wide variation between groups but little variation within each group. In this situation the appropriate type of sampling procedure to use is
   a) Cluster sampling
   b) Systematic sampling
   c) Stratified sampling
   d) Multistage sampling

5) Obtain the equation of the line of regression of yield of rice (y) on water (x) from the data given in the following table :

Water in inches (x)	12	18	24	30	36	42	48
Yield in tons (y)	5.27	5.68	6.25	7.21	8.02	8.71	8.42

Estimate the most probable yield of rice for 40 inches of water.

.....

.....

.....

.....

.....

.....

9.8 MULTIPLE REGRESSION

So far we have considered the case of the dependent variable being explained by one independent variable. However, there are many cases where the dependent variable is explained by two or more independent variables. For example, yield of crops (Y) being explained by application of fertilizer (X_1) and irrigation water(X_2). This sort of models is termed multiple regression. Here, the equation that we consider is

$$Y = \alpha + \beta X_1 + \gamma X_2 + e \qquad ...(9.13)$$

Where Y is the explained variable, X_1 and X_2 are explanatory variables, and e is the error term. In order to make the presentation simple we have dropped the subscripts. A regression equation can be fitted to (9.13) by applying the method of least squares discussed in Section 9.3. Here also we minimise $\sum e^2$ and obtain the normal equations as follows:

$$\begin{aligned} \Sigma Y &= n\alpha + \beta\Sigma X_1 + \gamma\Sigma X_2 \\ \Sigma X_1 Y &= \alpha\Sigma X_1 + \beta\Sigma X_1^2 + \gamma\Sigma X_1 X_2 \\ \Sigma X_2 Y &= \alpha\Sigma X_2 + \beta\Sigma X_1 X_2 + \gamma\Sigma X_2^2 \end{aligned} \qquad \ldots (9.14)$$

By solving the above equations we obtain estimates for α, β and γ. The regression equation that we obtain is

$$\hat{Y} = \alpha + \beta X_1 + \gamma X_2 \qquad ...(9.15)$$

Remember that we obtain predicted or forecast values of Y (that is $\hat{y}$) through (9.15) by applying various values for X_1 and X_2.

In the bivariate case (Y,X) we could plot the regression line on a graph paper. However, it is quite complex to plot the three variable case (Y, X_1, X_2) on graph paper because it will require three dimensions. However, the intuitive idea remains the same and we have to minimise the sum of errors. In fact when we add all the error terms ($e_1, e_2, \ldots\ldots e_n$) it sum up to zero.

In many cases the number of explanatory variables may be more than two. In

such cases we have to follow the basic principle of least squares: minimize Σe^2.

Thus if $Y = a_0 + a_1X_1 + a_2X_2 + \ldots\ldots\ldots\ldots + a_nX_n + e$ then we have to minimize $\Sigma e^2 = \Sigma(Y - a_0 - a_1X_1 - a_2X_2 \ldots\ldots\ldots - a_nX_n)^2$ and find out the normal equations.

Now a question arises, 'How many variables should be added in a regression equation?' It depends on our logic and what variables are considered to be important. Whether a variable is important or not can be identified on the basis of statistical tests also. These tests will be discussed later in Block 7.

We present a numerical example of multiple regression below.

Example 9.2

A student tries to explain the rent charged for housing near the University. She collects data on monthly rent, area of the house and distance of the house from the university campus and fits a linear regression model.

Rent (in Rs.'000) Y	Area (in sq.mt.) X_1	distance (in Km.) X_2
20	65	5.7
25	66	3.2
26	70	7.5
28	70	6.5
30	75	5
31	76	4
32	72	6
33	75	6.2
35	78	3.5
40	103	2.4

In the above example rent charged (Y) is the dependent variable while area of the house (X_1) and distance of the house from the university campus (X_2) are independent variables. The steps involved in estimation of regression line are:

i) Find out the regression equation to be estimated. In this case it is given by $Y = \alpha + \beta X_1 + \gamma X_2 + e$.

ii) Find out the normal equations for the regression equation to be estimated. In this case the normal equations are

$$\Sigma Y = n\alpha + \beta\Sigma X_1 + \gamma\Sigma X_2$$
$$\Sigma X_1Y = \alpha\Sigma X_1 + \beta\Sigma X_1^2 + \gamma\Sigma X_1X_2$$
$$\Sigma X_2Y = \alpha\Sigma X_2 + \beta\Sigma X_1X_2 + \gamma\Sigma X_2^2$$

iii) Construct a table as given in Table 9.4.

iv) Put the values from the table in the normal equations.

v) Solve for the estimates of α, β and γ.

Table 9.4: Computation of Multiple Regression

Y	X_1	X_2	X_1Y	X_2Y	X_1^2	X_2^2	X_1X_2	$\hat{Y}$	e_i
20	65	5.7	1300	114	4225	32.49	370.5	25.49	-5.49
25	66	3.2	1650	80	4356	10.24	211.2	25.71	-0.71
26	70	7.5	1820	195	4900	56.25	525	27.94	-1.94
28	70	6.5	1960	182	4900	42.25	455	27.85	0.15
30	75	5	2250	150	5625	25	375	30.00	0.00
31	76	4	2356	124	5776	16	304	30.37	0.63
32	72	6	2304	192	5184	36	432	28.72	3.28
33	75	6.2	2475	204.6	5625	38.44	465	30.11	2.89
35	78	3.5	2730	122.5	6084	12.25	273	31.24	3.76
40	103	2.4	4120	96	10609	5.76	247.2	42.58	-2.58
300	750	50	225000	15000	562500	2500	37500	300	0

By applying the above mentioned steps we obtain the estimated regression line as $\hat{Y} = -4.80 + 0.45X_1 + 0.09X_2$

9.9 NON-LINEAR REGRESSION

The equation fitted in regression can be non-linear or curvilinear also. In fact, it can take numerous forms. A simpler form involving two variables is the quadratic form. The equation is

$$Y = a + bX + cX^2$$

There are three parameters here viz., a, b and c and the normal equations are:

$$\sum Y = n\alpha + b\sum X + c\sum X^2$$

$$\sum XY = \alpha\sum X + b\sum X^2 + c\sum X^3$$

$$\sum X^2Y = \alpha\sum X^2 + b\sum X^3 + c\sum X^4$$

By solving for these equation we obtain the values of a, b and c.

Certain non-linear equations can be transformed into linear equations by taking logarithms. Finding out the optimum values of the parameters from the transformed linear equations is the same as the process discussed in the previous section. We give below some of the frequently used non-linear equations and the respective transformed linear equations.

1) $Y = a\ c^{bx}$

   By taking natural log (ln), it can be written as

   $\ln Y = \ln a + bX$

   or $Y' = \alpha + \beta X'$

   Where, $Y' = \ln Y$, $\alpha = \ln a$, $X' = X$ and $\beta = b$

2) $Y = aX^b$

   By taking logarithm (log), the equation can be transformed into

   $\log Y = \log a + b \log X$

   or $Y' = a + \beta X'$

   where, $Y' = \log Y$, $\alpha = \log a$, $\beta = b$ and $X' = \log X$

3) $Y = \frac{1}{a+bX}$

If we take $Y' = \frac{1}{Y}$ then

$Y' = a + bX$

4) $Y = a + b\sqrt{X}$

If we take $X' = \sqrt{X}$ then

$Y = a + bX'$

Once the non-linear equation is transformed, the fitting of a regression line is as per the method discussed in the beginning of this unit. We derive the normal equations and substitute the values calculated from the observed data. From the transformed parameters, the actual parameters can be obtained by making the reverse transformation.

Check Your Progress 2

1) Using the data on scores in Statistics and Economics of Table 9.8, compute the regression of Y on X and X on Y and check that the two lines are different. On the scatter diagram, plot both these regression lines. Check that the product of the regression coefficients is the square of the correlation coefficient.

.....

.....

.....

.....

.....

.....

2) Suppose that the least squares linear regression of family expenditure on clothing (Rs. Y) on family annual income (Rs. X) has been found to be $Y = 100 + 0.09X$, in the range $1000 < X < 100000$. Interpret this regression line. Predict the expenditure on the clothing of a family with an annual income of Rs. 10,000. What about families with annual income of Rs. 100 and Rs. 10,00,000?

.....

.....

.....

.....

.....

.....

9.10 LET US SUM UP

In this Unit we discussed an important statistical tool, that is, regression. In regression analysis we have two types of variables: dependent and independent. The dependent variable is explained by independent variables. The relationship

between variable takes the form of a mathematical equation. Based on our logic, understanding and purpose of analysis we categorise variables and identify the equation form.

The regression coefficient enables us to make predictions for the dependent variable given the values of the independent variable. However,.prediction remains more or less valid within the range of data used for analysis. If we attempt to predict for far off values of the independent variable we may get insensible values for the dependent variable.

9.11 KEY WORDS

Coefficient of Determination : It is given as r^2, i.e., the square of the correlation coefficient. It shows the percentage variation in the dependent variable Y explained by the independent variable X.

Normal Equations : A set of simultaneous equations derived in the application of the least squares method, for example in regression analysis. They are used to estimate the parameters of the model.

Regression : It is a statistical measure of the average relationship between two or more variables in terms of the original units of the data.

9.12 SOME USEFUL BOOKS

Nagar, A.L. and R.K. Das, 1989 : *Basic Statistics*, Oxford University Press, Delhi.

Goon, A.M., M.K. Gupta and B. Dasgupta, 1987 : *Basic Statistics*, The World Press Pvt. Ltd., Calcutta.

9.13 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) + 0.98 ; $Y = 0.64X + 0.54$; 8.2

2) $X = 0.95Y - 6.4$; $Y = 0.95X + 7.25$

3) $X = 2.23Y - 12.70$; $Y = 0.39X + 7.33$

(i) 29.6 (ii) 18.9

4) $Y = 0.613X + 14.81$; $X = 1.360y - 5.2$

5) $Y = 3.99 + 0.103X$; 8.11 tons

Check Your Progress 2

1) i) $Y = a + bX = 5.856 + 0.676x$

   ii) $X = \alpha + \beta Y = 29.848 + 0.799Y$

   iii) $r = 0.73$

   iv) $0.676 \times 0.799 = 0.54$

2) Expenditure on clothing, when family income is Rs. 10,000, is Rs. 1,000. In the case of income below Rs. 1,000 or above Rs. 1,00,000 the regression line may not hold good. In between both these figures, one rupee increase in income increases expenditure on clothes by 9 paise.

UNIT 18 STATISTICAL ESTIMATION

Structure

18.0 OBJECTIVES

After going through this Unit you will be in a position to:

- explain the concept of estimation;
- distinguish between point estimate and interval estimate;
- estimate confidence interval for a parameter; and
- explain the concept of confidence level.

18.1 INTRODUCTION

Many times due to certain constraints such as inadequate funds or manpower or time we are not in a position to survey all the units in a population. In such situations we take resort to sampling, that is, we survey only a part of the population. On the basis of the information contained in the sample we try to draw conclusions about the population. This process is called statistical inference. We must emphasise that statistical inference is widely applied in economics as well as in many other fields such as sociology, psychology, political science, medicine, etc. For example, before election process starts or just before declaration of election results many newspapers and television channels conduct exit polls. The purpose is to predict election results before the actual results are declared. At that point of time, it is not possible for the surveyors to ask all the voters about their preferences for electoral candidates — the time is too short, resources are scarce, manpower is not available, and a complete census before election defeats the very purpose of election!

In the above example the surveyor actually does not know the result, which is the outcome of votes cast by all the voters. Here all the voters taken together comprise the population. The surveyor has collected data from a representative sample of the population, not all the voters. On the basis of the information contained in the sample, (s)he is making forecast about the entire population.

In this Unit we deal with the concept of statistical inference and methods of statistical estimation. Parameter, as you know, is a function of population units while statistic is a function of sampling units. There could be a number of *parameters* and corresponding

statistics. However, in order to keep our presentation simple, we will confine ourselves mostly to arithmetic mean.

18.2 STATISTICAL BACKGROUND

In the previous two blocks we have discussed two important aspects: theoretical probability distributions and sampling techniques. These two aspects form the basis of statistical inference.

In Unit 14, Block 5 we explained the concept of a random variable. We learnt that X is a random variable if it assumes values $x_1, x_2, \ldots\ldots, x_n$ with corresponding probabilities $p_1, p_2, \ldots\ldots, p_n$ attached to it. Here the probability of occurrence of x_1 is p_1, the probability of occurrence of x_2 is p_2, and so on. If the values $x_1, x_2, \ldots\ldots, x_n$ are discrete we call X a discrete random variable and find out the probability for isolated values of X. On the other hand, if X is a continuous random variable we can find out the probability of X within certain range such that $P(a \leq X \leq b) = p_1$.

In Units 14 and 15 of Block 5 we discussed theoretical discrete probability distributions (such as binomial and Poisson) and continuous probability distributions (such as normal and t). We learnt that if the range of X increases infinitely then these probability distributions approach normal distribution. Thus normal distribution is a limiting case of these probability distributions and is considered as a sort of ideal among probability distributions.

The normal distribution is defined by two parameters: mean (μ) and standard deviation (σ). If the probabilities associated with a random variable are distributed according to normal distribution (that means, if X follows normal distribution), we can find out the probability of $P(a \leq X \leq b) = p_1$ by using the equation for its probability distribution function.

A problem encountered here is that μ and σ can take any values and finding out corresponding probability is time consuming. This problem is tackled by subtracting μ from the normal variable and dividing it by σ. This way we obtain the 'standard normal variate', $z = \frac{x-\mu}{\sigma}$, which has mean = 0 and standard deviation = 1. By plotting the probabilities for different values of z on a graph paper we obtain 'standard normal curve' which is symmetrical and area under the curve is = 1. Remember that in the case of standard normal curve we measure $z = \frac{x-\mu}{\sigma}$ on the x-axis and probability of occurrence of z, that is $p(z)$, on the y-axis. Thus if we consider a particular segment of the normal curve (bounded by two values of z, say, z_1 and z_2) the area under the curve gives its probability. Remember that normal curve is different from the frequency curve considered in Block 1 of this course. Area under the normal curve does not give frequencies; it gives probabilities.

In Unit 16 of Block 6 we leant that very often it is not possible to study the entire population and we undertake a sample survey. If the sample is drawn in a random manner through appropriate probability attached to each population unit and the sample size is not very small, the sample can be a representative one of the population. Recall that we can draw a number of samples from a given population and each sample provides us with a sample mean. Thus the sample means can be arranged in the form of a frequency distribution, called the 'sampling distribution'.

We know from Unit 16, Block 6 that sample mean ($\bar{x}$) assumes different values and for each value we can attach a probability. Thus sample mean can be considered as a random variable. In real life situations we have a finite population and the number of samples (and therefore the number of sample means) is finite. In this case $\bar{x}$ is a discrete random variable but when there are infinite number of samples, $\bar{x}$ could be a continuous random variable.

Now let us consider another important concept discussed in Unit 16: the central limit theorem. It says that sampling distribution of $\bar{x}$ is normal if the population from which the sample is drawn is normal. However, sampling distribution of $\bar{x}$ is approximately normal if sample size (n) is large, even if the parent population (that is, population from which it is drawn) is not normal. If the parent population is approximately normal then sampling distribution of sample means is approximately normal even when sample size is small.

We know that dispersion of sample means is smaller in value than dispersion of the parent population from which the sample is drawn. Recall that the standard deviation of the sampling distribution is called standard error. Thus if the population has a standard deviation σ then the standard error of sample mean is $\frac{\sigma}{\sqrt{n}}$.

From the above we learn that sample mean can be considered as a random variable and it approximates normal distribution when sample size is large. Usually we consider a sample to be large in size if $n > 30$. For small samples ($n \leq 30$), sampling distribution of sample means is similar to student's t distribution. Recall that in the case of t distribution the shape of the probability curve changes according to its 'degrees of freedom'.

18.3 CONCEPT OF STATISTICAL INFERENCE

As mentioned earlier, statistical inference deals with the methods of drawing conclusions about the population characteristics on the basis of information contained in a sample drawn from the population. Remember that population mean is not known to us, but we know the sample mean. In statistical inference we would be interested in answering two types of questions. First, what would be the value of the population mean? The answer lies in making an informed guess about the population mean. This aspect of statistical inference is called 'estimation'. The second question pertains to certain assertion made about the population mean. Suppose a manufacturer of electric bulbs claims that the mean life of electric bulbs is equal to 2000 hours. On the basis of the sample information, can we say that the assertion is not correct? This aspect of statistical inference is called hypothesis testing.

Thus statistical inference has two aspects: estimation and hypothesis testing. We will discuss about statistical estimation in the present Unit while testing of hypothesis will be taken up for discussion in the next Unit.

Fig. 18.1 below summarises different aspects of statistical inference. A crucial factor before us is whether we know the population variance or not. Of course when we do not know the population mean, how do we know the population variance? We begin with the case where population variance is known, because it will help us in explaining the concepts. Later on we will take up the more realistic case of unknown population variance.

Estimation could be of two types: point estimation and interval estimation. In point estimation we estimate the value of the population parameter as a single point. On the other hand, in the case of interval estimation we estimate lower and upper bounds around sample mean within which population mean is likely to remain.

The assertion or claim made about the population mean would be in the form of a null hypothesis and its counterpart, alternative hypothesis. We will explain these concepts and the methods of testing of hypothesis in the next Unit.

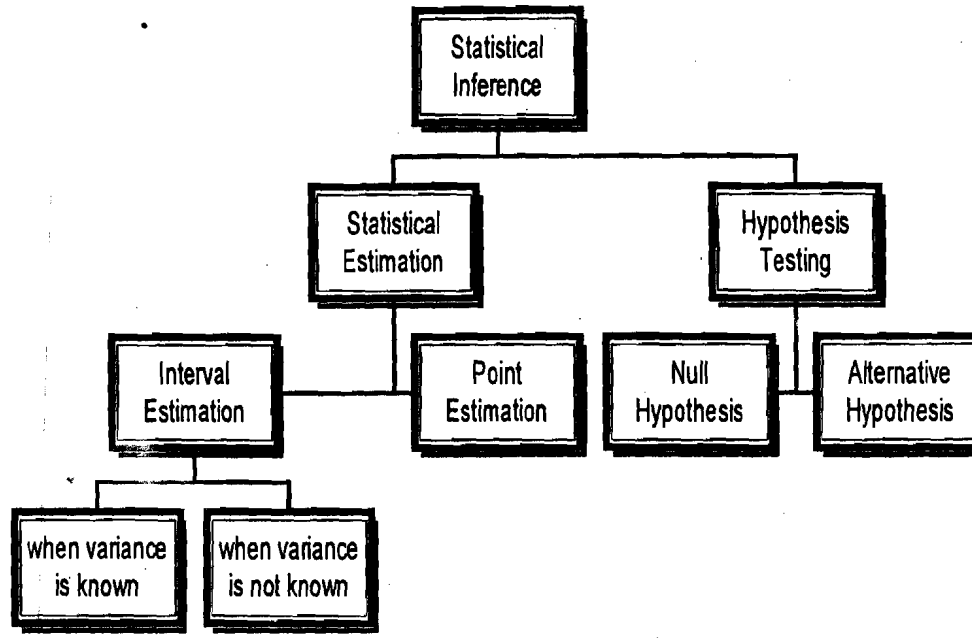


Fig. 18.1: Statistical Inference

Check Your Progress 1

1) Explain the following concepts.
   a) standard normal variate
   b) random variable
   c) sampling distribution
   d) central limit theorem

.....

.....

.....

.....

.....

.....

2) State whether the following statements are true or false.
   a) Normal distribution is a limiting case of binomial distribution.
   b) Standard deviation of sampling distribution of a statistic is termed as standard error.
   c) Poisson distribution is an example of continuous distribution.
   d) Statistical estimation is a part of statistical inference.

.....

.....

.....

.....

.....

18.4 POINT ESTIMATION

As mentioned earlier we do not know the parameter value and want to guess it by using sample information. Obviously the best guess will be the value of the sample statistic. For example, if we do not know the population mean the best guess would be the sample mean. Here in this case we use a single value or point as 'estimate' of the parameter.

In Unit 16 we have explained the concepts of estimate and estimator. Also we have pointed out the distinction between the two. Recall that estimator is the formula and estimate is the particular value obtained by using the formula. For example, if we use sample mean for estimation of population mean, then $\frac{1}{n}\sum_i x_i$ is the estimator. Suppose I collect data on a sample, and put the sampling units to this formula and obtain a particular value for sample mean, say 120. Then 120 is an estimate of population mean. It is possible that you draw another sample from the same population, use the formula for sample mean, that is $\frac{1}{n}\sum_i x_i$, and obtain a different value, say 123. Here both 120 and 123 are estimates of population mean. But in both the cases the estimator is the same, which is $\frac{1}{n}\sum_i x_i$. Remember that the term statistic, which is used to mean a function of sample values, is a synonym for estimator.

There may be situations when you would find more than one potential estimator (alternative formulae) for a parameter. In order to choose the best among these estimators, we need to follow certain criteria. Based on these criteria an estimator should fulfill certain desirable properties. There are quite a few desirable properties for an estimator, but the most important is its unbiasedness.

Unbiasedness means that an estimate may be higher or lower than the unknown value of the parameter. But the expected value of the estimate should be equal to the parameter. For example, sample mean ($\bar{x}$) may fluctuate from sample to sample but on an average it would be equal to population mean. In other words, $E(\bar{x}) = \mu$.

However, $\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is not an unbiased estimator of the population variance $\sigma^2 = \frac{1}{N}\Sigma(X_i - \bar{X})^2$. In fact, if we define $s^2 = \frac{1}{n-1}\Sigma(x_i - \bar{x})^2$, then s^2 is an unbiased estimator of σ^2. Usually a sample is less dispersed than the population from which it is drawn. Therefore, there is a tendency for the sample standard deviation s to be little less than population standard deviation σ. In order to rectify this condition we artificially inflate s by dividing by a smaller number $(n-1)$, instead of n.

The point estimate is quite important for testing of hypothesis, as we will see in Unit 19.

18.5 CONFIDENCE INTERVAL FOR KNOWN VARIANCE

We have seen above that in point estimation, we estimate the parameter by a single value, usually the corresponding sample statistic. The point estimate may not be realistic in the sense that the parameter value may not exactly be equal to it. An alternative procedure is to give an interval, which would hold the parameter with certain probability.

Here we specify a lower limit and an upper limit within which the parameter value is likely to remain. Also we specify the probability of the parameter remaining in the interval. We call the *interval* as 'confidence interval' and the *probability* of the parameter remaining within this interval as 'confidence level' or 'confidence coefficient'.

Let us take an example. Suppose you are asked to estimate the average income of people in Raigarh district of Chhattisgarh state. You collected data from a sample of 500 households and found the average income (say, $\bar{x}$) to be Rs. 18250 per annum. This sample average may not be equal to the actual average income of Raigarh district of Chhattisgarh (μ) because of sampling error. Thus we are not sure whether average income of the above district is Rs. 18250 or not. On the other hand, it will be more sensible if we say that average income of Raigarh district of Chhattisgarh is between Rs. 17900 and Rs. 18600 per annum. Also we may specify that the probability that average income will remain within these limits is 95 per cent. Thus our confidence interval in this case is Rs. 17900-18600 and the confidence level or confidence coefficient is 95 per cent.

Here a question may be shaping up in your mind, 'How do we find out the confidence interval and confidence coefficient?' Let us begin with confidence coefficient. We know that the sampling distribution of $\bar{x}$ for large samples is normally distributed with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$, where n is the size of the sample. By transforming the sample mean ($z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$) we obtain *standard normal variate*, which has zero mean and unit variance. The standard normal curve is symmetrical and therefore, the area under the curve for $0 \le z \le \infty$ is 0.5 which is presented in the form of a table (See Table 15.1 in Unit 15 of Block 5). Let us assume that we want our confidence coefficient to be 95 per cent (that is, 0.95). Thus we should find out a range for z which will cover 0.95 area of the standard normal curve. Since distribution of z is symmetrical, 0.475 area should remain to the right and 0.475 area should remain to the left of $z = 0$. If look into normal area table (Table 15.1) we find that 0.475 area is covered when $z = 1.96$. Thus the probability that z ranges between –1.96 to 1.96 is 0.95. From this information let us work out backward and find the range within which μ will remain.

We find that

$$P(-1.96 \le z \le 1.96) = 0.95 \qquad \ldots(18.1)$$

$$or\ P\left(-1.96 \le \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \le 1.96\right) = 0.95$$

$$or\ P\left(-1.96\frac{\sigma}{\sqrt{n}} \le \bar{x} - \mu \le 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$or\ P\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95 \qquad \ldots(18.2)$$

Let us interpret the above. Recall that each sample would provide us with a different value of $\bar{x}$. Accordingly, the confidence interval would be different. In each case the confidence interval would contain the unknown parameter or it would not. Equation (18.2) means that if a large number of random samples, each of size n,

are drawn from the given population and if for each such sample the interval $\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$ is determined, then in about 95% of the cases, the interval will include the population mean μ.

The confidence coefficient is denoted by $(1-\alpha)$ where α is the level of significance (we will discuss the concept of 'level of significance' in Unit 19). Confidence coefficient could take any value. We can ask for a confidence level of say 81 per cent or 97 per cent depending upon how precise our conclusions should be. However, conventionally two confidence levels are frequently used, namely, 95 per cent and 99 per cent. Of course at times we take 90 per cent confidence level also, though not frequently.

Let us find out the confidence interval when confidence coefficient $(1-\alpha) = 0.99$. In this case 0.495 area should remain on either side of the standard normal curve. If we look into the normal area table (Table 15.1) we find that 0.495 area is covered when z = 2.58.

Thus

$$P(-2.58. \leq z \leq 2.58) = 0.99 \qquad \ldots(18.3)$$

By rearranging the terms in the above we find that

$$P\left(\bar{x} - 2.58\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 2.58\frac{\sigma}{\sqrt{n}}\right) = 0.99 \qquad \ldots(18.4)$$

Equation (18.4) implies that 99 per cent confidence interval for μ is given by

$$\bar{x} \pm 2.58\frac{\sigma}{\sqrt{n}}.$$

By looking into the normal area table you can work out the confidence interval for confidence coefficient of 0.90 and find that

$$P\left(\bar{x} - 1.65\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.65\frac{\sigma}{\sqrt{n}}\right) = 0.90 \qquad \ldots(18.5)$$

We observe from (18.2), (18.4) and (18.5) that as the interval widens, the chance for the interval holding a population parameter (in this case μ) increases.

The two limits of the confidence interval are called *confidence limits*. For example, for 95 per cent confidence level we have the lower confidence limit as $\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}$ and the upper confidence limit as $\bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$. The confidence coefficient can be interpreted as the confidence or trust that we place in these limits for actually holding μ.

Example 18.1

A paper company wants to estimate the average time required for a new machine to produce a ream of paper. A random sample of 36 reams shows an average production time of 1.5 minutes per ream of paper. The population standard deviation is known to be 0.30 minute. Construct an interval estimate with 95% confidence limits.

The information given is

$\bar{x} = 1.5$, $\sigma = 0.30$ and $n = 36$

Since $n = 36\ (> 30)$, we can take the sample as a large sample and accordingly $\bar{x}$ is normally distributed with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$. Now, the standard error is

$$\frac{\sigma}{\sqrt{n}} = \frac{0.30}{\sqrt{36}} = 0.05$$

The 95% confidence interval is given by

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

or $1.5 - 1.96 \times 0.05 \le \mu \le 1.5 + 1.96 \times 0.05$

i.e., $1.402 \le \mu \le 1.598$

Thus with 95% confidence, we can state that the average production time for the new machine will be between 1.402 minutes and 1.598 minutes. Here, 1.402 is the lower confidence limit and 1.598 is the upper confidence limit.

18.6 CONFIDENCE INTERVAL FOR UNKNOWN VARIANCE

In the previous Section we estimated confidence interval for population mean on the assumption that population variance is known. It is a bit unrealistic that we do not know population mean (we want to estimate it) but know population variance. A realistic case would be the assumption that both population mean and variance are unknown. On the basis of sample mean and variance we want to find out confidence interval for population mean.

Since the population standard deviation (σ) is not known we use the sample standard deviation (s) in its place. However, in such a case the sampling distribution of $\bar{x}$ is not normal, rather it follows student's t distribution. The standard error of the sample means would be $\frac{s}{\sqrt{n}}$.

Like the standard normal variate z, the t-distribution has a mean of zero, is symmetrical about mean and ranges between $-\infty$ to ∞. But its variance is greater than 1. Actually its variance changes according to degrees of freedom. However, when $n > 30$ the t-distribution has a variance very close to 1 and thus resembles z-distribution.

The t-statistic, like the z-statistic, is calculated as

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

By looking into the area table for t-distribution (see Table 15.3 in Unit 15) we find the probability values for the confidence level that we require. The degrees of freedom is $(n-1)$. Thus the confidence interval would be

$$\bar{x} - t \cdot \frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t \cdot \frac{s}{\sqrt{n}} \qquad \ldots(18.6)$$

Example 18.2

The mean weight (in kilogram) of 20 children are found to be 15 with a standard deviation of 4. On the basis of the above information estimate 95 per cent confidence

interval for mean weight of the population from which the sample is drawn. Assume that population is normally distributed.

Since population is normal and sample size is small we apply *t*-distribution for estimation of confidence interval. Since $n = 20$ we have degrees of freedom (d.f.) = 19. We move down the first column of Table 15.3 till we reach the row corresponding to 19. Since we need 95 per cent confidence interval we should leave 0.025 area on each side of $t = 0$ as we did in the previous Section. Thus for 19 degrees of freedom and $\alpha = 0.025$ we find that *t*-value is 2.093.

Hence the confidence interval is

$$15 - 2.093 \times \frac{4}{\sqrt{20}} \leq \mu \leq 15 + 2.093 \times \frac{4}{\sqrt{20}}$$

or $\quad 15 - 1.87 \leq \mu \leq 15 + 1.87$

or $\quad 13.13 \leq \mu \leq 16.87$

Similarly you can find out confidence intervals for different sample sizes and confidence coefficients.

Let us summarise the rules for application of z or *t* statistic for estimation of confidence interval.

1) If sample size is large (n>30) apply z-statistic — it does not matter whether i) parent population is normal or not, and ii) variance is known or not.
2) If sample size is small ($n \leq 30$) check whether i) parent population is normal, and ii) variance is known.
   a) If parent population is not normal apply nonparametric tests.
   b) If parent population is normal and variance is known apply *z*.
   c) If parent population is normal and variance is not known apply *t*.

In Fig. 18.2 we present the above in the form of a chart.

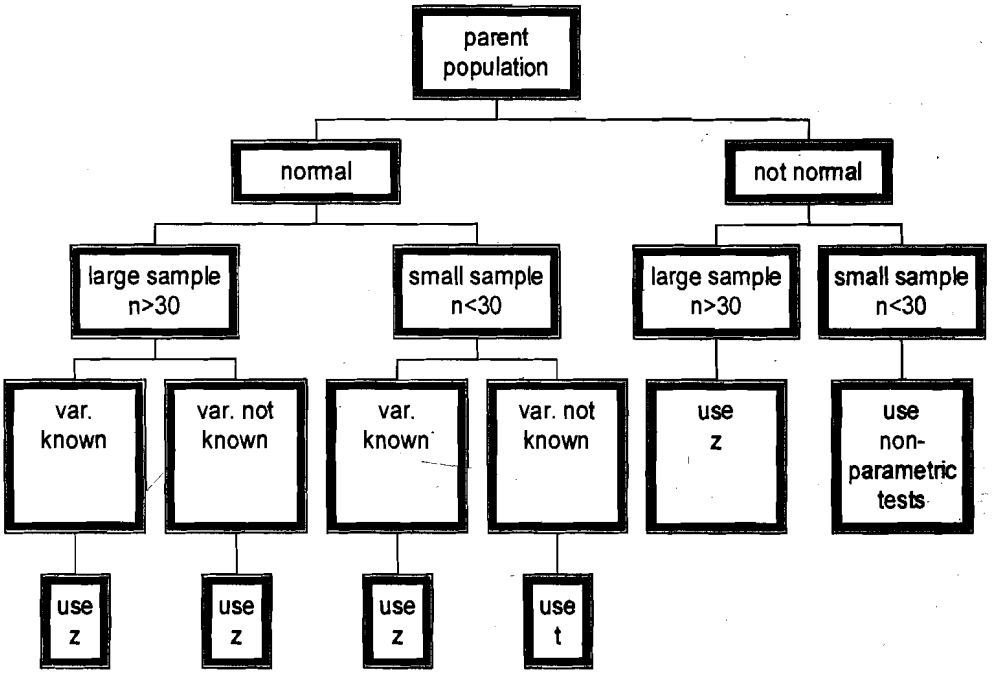


Fig. 18.2: Selection of Proper Test Statistic

Check Your Progress 2

1) A sample of 50 employees were asked to provide the distance commuted by them to reach office. If sample mean was found to be 4.5 km. find 95 percent confidence interval for the population. Assume that population is normally distributed with a variance of 0.36.

.....

.....

.....

2) For a sample of 25 students in school the mean height was found to be 95 cm. with a standard deviation of 4 cm. Find the 99 percent confidence interval.

.....

.....

.....

3) State whether the following statements are true or false.

   a) When parent population is not normal and sample size is small we use t-distribution to estimate confidence interval.

   b) The range of t-distribution is 0 to infinity.

   c) When confidence level is 90 per cent, level of significance is 10 per cent.

.....

.....

.....

18.7 LET US SUM UP

Drawing conclusions about a population on the basis of sample information is called statistical inference. Here we have basically two things to do: estimation and hypothesis testing. In this unit we took up the first issue while the second one will be discussed in the remaining units of the block.

An estimate of an unknown parameter could be either a point or an interval. Sample mean is usually taken as a point estimate of population mean. On the other hand, in interval estimation we construct two limits (upper and lower) around the sample mean. We can say with stipulated level of confidence that the population mean, which we do not know, is likely to remain within the confidence interval. In order to construct confidence interval we need to know the population variance or its estimate. When we know population variance, we apply normal distribution to construct the confidence interval. In cases where population variance is not known, we use student's t for the above purpose. Remember that when sample size is large (n>30) t-distribution approximates normal distribution. Thus for large samples, even if population variance is not known, we can use normal distribution for construction of confidence interval on the basis of sample mean and sample variance.

18.8 KEY WORDS

Confidence Level : It gives the percentage (probability) of samples where the population mean would remain within the confidence interval around the sample mean. If α is the significance level the confidence level is $(1-\alpha)$.

Estimation : It is the method of prediction about parameter value on the basis of statistic.

Estimator : It is another name given to statistic in the theory of estimation.

Parameter : It is a measure of some characteristic of the population.

Population : It is the entire collection of units of a specified type in a given place and at a particular point of time.

Random Sampling : It is a procedure where every member of the population has a definite chance or probability of being selected in the sample. It is also called probability sampling.

Sample : It is a sub-set of the population. It can be drawn from the population in a scientific manner by applying the rules of probability so that personal bias is eliminated. Many samples can be drawn from a population and there are many methods of drawing a sample.

Sampling Error : In the sampling method, we try to approximate some feature of a given population from a sample drawn from it. Now, since in the sample all the members of the population are not included, howsoever close the approximation is, it is not identical to the required population feature and some error is committed. This error is called the sampling error.

Significance Level : There may be certain samples where population mean would not remain within the confidence interval around sample mean. The percentage (probability) of such cases is called significance level. It is usually denoted by α. When $\alpha = 0.05$ (that is, 5 percent) we can say that in 5 per cent cases we are likely to reach an incorrect decision or commit Type I error. Level of significance could be at any level but it is usually taken at 5 percent or 1 percent level.

Statistic : It is a function of the values of the units that are included in the sample. The basic purpose of a statistic is to estimate some population parameter.

Sampling Distribution : It is the relative frequency or probability distribution of the values of a statistic when the number of samples tends to infinity.

Standard Error : It is the standard deviation of the sampling distribution of a statistic.

Statistical Inference : It is the process of concluding about an unknown population from a known sample drawn from it.

Problem of Estimation : We may be interested in some feature of the population that is *completely* unknown to us and we want to make some intelligent guess about it on the basis of a random sample drawn from the population. This problem of statistical inference is known as the problem of estimation.

18.9 SOME USEFUL BOOKS

Nagar, A. L. and Das, R. K., 1989, *Basic Statistics:* Oxford University Press, Delhi, Chapter 9.

Newbold, P., 1991, *Statistics for Business and Economics* (Third Edition): Prentice Hall, New Jersey, Chapters 6, 7, 8 and 9.

Keller, G, and B. Warrack, 1991, *Essentials of Business Statistics*, Wordsworth Publishing Co., California, Chapters 7 and 8.

18.10 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Go through Section 18.2 and answer.

2) a) true b) true c) true d) true

Check Your Progress 2

1) Since it is large sample we apply z-statistic. The confidence interval is

$4.40 \leq \mu \leq 4.60$

2) Since it is small sample and population variance is not given we apply t-statistic with degrees of freedom 24. The tabulated value of t at 99 per cent confidence level is 2.49. The confidence interval is $93.01 \leq \mu \leq 96.99$.

3) a) false b) false c) true

UNIT 19 TESTING OF HYPOTHESIS

Structure

19.0 OBJECTIVES

After going through this Unit you will be in a position to:

- explain the concepts of null hypothesis and alternative hypothesis;
- identify critical region based on the level of significance;
- distinguish between Type I and Type II errors;
- test for hypothesis concerning population mean on the basis of one sample; and
- test for the difference in sample means obtained from two samples.

19.1 INTRODUCTION

In the previous Unit we learnt about the estimation of confidence interval for population mean on the basis of sample data. In the present Unit we will look into another aspect of statistical inference, that is, hypothesis testing. Hypothesis is a statement or assertion or claim about the population parameter. For example, suppose we have a hitch that the per capita income of Chhattisgarh state is Rs. 20000 per annum. We can be sure about the truth in the above statement if we undertake a complete census of households in the state. This implies we collect data on the income of all the households in Chhattisgarh and calculate the per capita income of the state. However, constraints such as time, money and manpower may restrict us to go for a sample survey and draw a conclusion about the statement on the basis of sample information. The procedure followed in the above is the subject matter of hypothesis testing.

Hypothesis testing is applied widely in various fields and to various situations. For example, suppose the effectiveness of a new drug in curing tuberculosis needs to be

tested. Obviously all the patients suffering from tuberculosis need not be administered with the new drug to see its effectiveness. What we need is a representative sample and test whether the new drug is more effective than existing drugs. As another example let us take the case of a planner who asserts that the crude birth rate is the same in the states Bihar and Rajasthan. In this case it may not be possible on our part to go for a census survey of all the births that have taken place in Bihar and Rajasthan during the last year and calculate the crude birth rate. Instead a sample survey is undertaken and the assertion made by the planner is put to test.

In hypothesis testing we try to answer questions of the following types: Is the sample under consideration drawn from a particular population? Is the difference between two samples significant enough that they cannot belong to the same population?

19.2 FORMULATION OF A HYPOTHESIS

A hypothesis is a tentative statement about a characteristic of a population. It could be an assertion or a claim also. For example, official records for recent years claim that female literacy in Orissa is 51 per cent. Here a statement or a claim about the rate of female literacy is being made. Thus it could be considered as a hypothesis.

In hypothesis testing there are four important components: i) null hypothesis, ii) alternative hypothesis, iii) test statistic, and iv) interpretation of results. We discuss each of these below.

Usually statistical hypotheses are denoted by the alphabet *H*. There are two types of hypothesis: null hypothesis and alternative hypothesis. A null hypothesis is a statement that we consider to be true about the population and put to test by a test statistic. Usually we denote null hypothesis by H_0. In the example on female literacy in Orissa our null hypothesis is

$$H_0 : \mu = 0.51 \qquad \ldots(19.1)$$

where μ is the parameter, in this case female literacy in Orissa.

There is a possibility that the null hypothesis that we intend to test is not true and female literacy is not equal to 51 per cent. Thus there is a need for an alternative hypothesis which holds true in case the null hypothesis is not true. We denote alternative hypothesis by the symbol H_A and formulate it as follows:

$$H_A : \mu \neq 0.51 \qquad \ldots(19.2)$$

We have to keep in mind that null hypothesis and alternative hypothesis are mutually exclusive, that is, both cannot be true simultaneously. Secondly, both H_0 and H_A exhaust all possible options regarding the parameter, that is, there cannot be a third possibility. For example, in the case of female literacy in Orissa, there are two possibilities — literacy rate is 51 per cent or it is not 51 per cent; a third possibility is not there.

It is a rare coincidence that sample mean ($\bar{x}$) is equal to population mean (μ). In most cases we find a difference between $\bar{x}$ and μ. Is the difference because of sampling fluctuation or is there a genuine difference between the sample and the population? In order to answer this question we need a test statistic to test the difference between the two. The result that we obtain by using the test statistic needs to be interpreted and a decision needs to be taken regarding the acceptance or rejection of the null hypothesis.

The development of test statistic for hypothesis testing and interpretation of results requires elaboration. Before discussing further on these two steps we present another concept — critical or rejection region.

19.3 REJECTION REGION AND TYPE OF ERRORS

The underlying idea behind hypothesis testing and interval estimation (discussed in the previous Unit) is the same. Recall from Unit 18 that a confidence interval is built around sample mean with certain confidence level. A confidence level of 95 per cent implies that in 95 per cent cases the population mean would remain in the confidence interval estimated from the sample mean. It is implicit that in 5 per cent cases the population mean will not remain within the confidence interval. Note that when the population mean does not remain within the confidence interval we should reject the null hypothesis.

19.3.1 Rejection Region for Large Samples

Let us explain the concept of critical region for large samples. Later on we will extend the concept to small samples.

As you already know from previous Units, sampling distribution of sample mean ($\bar{x}$) follows normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. Thus $\bar{x}$ can be transformed into a standard normal variate, z, so that it follows normal distribution with mean 0 and standard deviation 1. Recall that $\frac{\sigma}{\sqrt{n}}$ is the standard error of $\bar{x}$.

In notations $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ and $z \sim N(0,1)$.

Recall from Unit 15, Block 5 that area under the standard normal curve gives the probability for different range of values assumed by z. These probabilities can also be presented in a table (see Table 15.1 in Unit 15, Block 5).

Let us look into the standard normal curve presented in Fig. 19.1, where the x-axis represents the variable z and the y-axis represents the probability of z, that is $p(z)$. We should note the following points.

- When sample mean is equal to population mean (that is, $\bar{x} = \mu$) we find that $z = 0$. When $\bar{x} > \mu$ we find that z is positive. On the other hand, when $\bar{x} < \mu$ we find that z is negative.
- Note that we are concerned with the difference between $\bar{x}$ and μ. Therefore, negative or positive sign of z does not matter much.
- Higher the difference between $\bar{x}$ and μ, higher is the absolute value of z. Thus z-value measures the discrepancy between $\bar{x}$ and μ, and therefore can be used as a test statistic.
- We should find out a critical value of z beyond which the difference between $\bar{x}$ and μ is significant.
- If the absolute value of z is less than the critical value we should not reject the null hypothesis.
- If the absolute value of z exceeds the critical value we should reject the null hypothesis and accept the alternative hypothesis.

Thus in the case of large samples the absolute value of z can be considered as test statistic for hypothesis testing such that

$$z = \frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}} = \frac{|\bar{x} - x|}{S.E.(\bar{x})} \quad \ldots(19.3)$$

Let us explain the concept of critical region through the standard normal curve given in Fig. 19.1 below. When we have a confidence coefficient of 95 per cent, the area covered under the standard normal curve is 95 per cent. Thus 95 per cent area under the curve is bounded by $-1.96 \leq z \leq 1.96$. The remaining 5 per cent area is covered by $z \leq -1.96$ and $z \geq 1.96$. Thus 2.5 per cent of area on both sides of the standard normal curve constitute the rejection region. This area is shown in Fig. 19.1. If the sample mean falls in the rejection region we reject the null hypothesis.

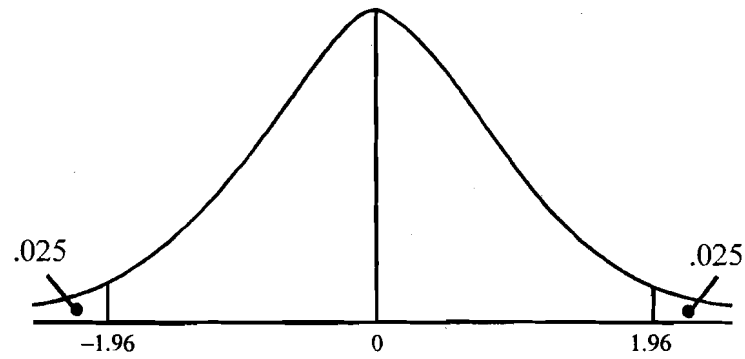


Fig. 19.1: Critical Regions

19.3.2 One-tail and Two-tail Tests

In Fig. 19.1 we have shown the rejection region on both sides of the standard normal curve. However, in many cases we may place the rejection region on one side (either left or right) of the standard normal curve.

Remember that if α is the level of significance, then for a two-tail test $\frac{\alpha}{2}$ area is placed on both sides of the standard normal curve. But if it is a one-tail test then α area is placed on one-side of the standard normal curve. Thus the critical value for one-tail and two-tail tests differ.

The selection of one-tail or two-tail test depends upon the formulation of the alternative hypothesis. When the alternative hypothesis is of the type $H_A : \bar{x} \neq \mu$ we have a two-tail test, because $\bar{x}$ could be either greater than or less than μ. On the other hand, if alternative hypothesis is of the type $H_A : \bar{x} < \mu$, then entire rejection is on the left hand side of the standard normal curve. Similarly, if the alternative hypothesis is of the type $H_A : \bar{x} > \mu$, then the entire rejection is on the right hand side of the standard normal curve.

The critical values for z depend upon the level of significance. In Table 19.1 these critical values for certain specified levels of significance (α) are given for the tests to be conducted under the assumption of normal distribution. The values are given for both two-tail and one-tail tests.

Table 19.1: Critical Values for z-statistic

Significance Level (α)	0.10	0.05	0.01	0.005
two-tail test	1.65	1.96	2.58	2.81
one-tail test	1.28	1.65	2.33	2.58

Note: The table is derived from Table 15.1.

19.3.3 Type I and Type II Errors

In hypothesis testing we reject or do not reject a hypothesis with certain degree of confidence. As you know, a confidence coefficient of 0.95 implies that in 95 out of 100 samples the parameter remains within the acceptance region and in 5 per cent cases the parameter remains in the rejection region. Thus in 5 per cent cases the sample is drawn from the population but sample mean is too far away from the population mean. In such cases the sample belongs to the population but our test procedure rejects it. Obviously we commit an error such that H_0 is true but gets rejected. This is called 'Type I error'. Similarly there could be situations when the H_0 is not true, but on the basis of sample information we do not reject it. Such an error in decision making is termed 'Type II error' (see Table 19.2).

Note that Type I error specifies how much error we are in a position to tolerate. Type I error is equal to the level of significance, and is denoted by α. Remember that confidence coefficient is equal to $1-\alpha$.

Table 19.2: Type of Errors

	H_0 **true**	H_0 **not true**
Reject H_0	Type I Error	Correct decision
Do not reject H_0	Correct decision	Type II Error

19.3.4 Rejection Region for Small Samples

Let us go back to Fig. 18.2 in Unit 18 where we have given certain criteria for use of proper test statistic in interval estimation. We see that in the case of small samples ($n \leq 30$), if population standard deviation is known we apply *z*-statistic for hypothesis testing. On the other hand, if population standard deviation is not known we apply *t*-statistic. The same criteria apply to hypothesis testing also.

In the case of small samples if population standard deviation is known the test statistic is

$$z = \frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}} \qquad \ldots(19.4)$$

On the other hand, if population standard deviation is not known the test statistic is

$$t = \frac{|\bar{x} - \mu|}{s/\sqrt{n}} \qquad \ldots(19.5)$$

In the case of *t*-statistic, however, the area under the curve (which implies probability) changes according to degrees of freedom. Thus while finding the critical value of *t* we should take into account the degrees of freedom. When sample size is *n*, degrees of freedom is $n-1$. Thus we should remember two things while finding critical value of *t*. These are: i) significance level, and ii) degrees of freedom.

Check Your Progress 1

1) Distinguish between the following:
   a) Null hypothesis and Alternative hypothesis
   b) One-tail and Two-tail tests

c) Confidence level and Level of significance

d) Type I and Type II errors

.....

.....

.....

.....

.....

2) Suppose a sample of 100 students has mean age of 12.5 years. Show through diagram the rejection region at 5 per cent level of significance to test the hypothesis that the sample has a mean age greater than the population mean. Assume that population mean and standard deviation are 10 years and 2 years respectively.

.....

.....

.....

.....

.....

19.4 TESTING OF HYPOTHESIS FOR A SINGLE SAMPLE

In many situations we are asked to judge whether a sample is significantly different from a given population. For example, let us assume that we surveyed a sample of 400 households of Raigarh district of Chhattisgarh state and calculated the per capita income of these households. Subsequently, our task is to test the hypothesis that per capita income calculated from the sample is not different from the per capita income of the district.

In the above example we can have two different situations: i) population (in this case all the households of the district) variance is known, ii) population variance is not known to us. We explain the steps to be followed in each case below.

19.4.1 Population Variance is Known

Let us consider the case that we know the per capita income of Raigarh district of Chhattisgarh as well as its variance. Suppose the data available in official records show that per capita income of Raigarh district is Rs. 10000 and standard deviation of per capita income is Rs. 1500. However, we did a sample survey of 400 households and found that their per capita income is Rs. 10500. Do we accept the data provided in official records?

In this case μ = Rs. 10000

σ = Rs. 1500

$\bar{x}$ = Rs. 10500

$n = 400$

From the central limit theorem we know that when sample size is large, sample is approximately normally distributed. This is true even in cases where the parent population is normally distributed. Thus the example is appropriate for application of normal distribution.

Our null hypothesis in this case is

$$H_0 : \bar{x} = \mu$$

The null hypothesis suggests that sample mean is equal to population mean. In other words, per capita income obtained from the sample is the same as the data provided in official records.

Our alternative hypothesis is

$$H_A : \bar{x} \neq \mu$$

Suppose we do not have any reason to say that per capita income obtained from the sample ($\bar{x}$) is greater than or smaller than the per capita income available in official records. Thus our alternative hypothesis is that $\bar{x}$ could be on either side of μ. Therefore, we should go for two-tail test so that rejection region is on both sides of the standard normal curve and the test statistic is

$$z = \frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}}. \qquad \ldots(19.6)$$

By substituting values in the above we obtain

$$z = \frac{|10500 - 10000|}{1500/\sqrt{400}} = \frac{500}{1500/20} = \frac{500}{75} = 6.67$$

Recall the standard normal curve and the area for different values of z (See Table 15.1 in Unit 15). We notice that when $z = 1.96$ the area covered under standard normal curve is 0.4750. Therefore, the level of significance is 5 per cent. Similarly, when $z = 2.58$ the area covered under standard normal curve is 0.4950. Therefore, the level of significance is 1 per cent.

In the above case since $z = 6.67$, the sample lies in the critical region and we reject the null hypothesis. Thus the per capita income obtained from the sample is significantly different from the per capita income provided in official records.

The steps you should follow are:

1) Specify the null hypothesis.
2) Find out whether it requires one-tail or two-tail test. Accordingly identify your critical region. This will help in specification of alternative hypothesis.
3) Apply sample values to z-statistic given at (19.6).
4) Find out from z-table the critical value according to level of significance.
5) If you obtain a value lower than the critical value do not reject the null hypothesis.
6) If you obtain a value greater than the critical value reject the null hypothesis and accept the alternative hypothesis.

Example 19.1

Suppose the voltage generated by certain brand of battery is normally distributed. A random sample of 100 such batteries was tested and found to have a mean voltage of 1.4 volts. At 0.01 level of significance, does this indicate that these batteries have a general average voltage, that is different from 1.5 volts? Assume that population standard deviation is 0.21 volts.

Here, H_0: $\mu = 1.5$

Since average voltage of the sample can be different from average voltage of the population if it is either less than or more than 1.5 volts, our rejection region is on both sides of the normal curve. Thus it is a case of two-tail test and alternative hypothesis is

$H_1 : \mu \neq 1.5$

Since the population standard deviation σ is known, the test statistic is

$$z = \frac{|\bar{x} - \mu|}{\frac{\sigma}{\sqrt{n}}} = \frac{|1.4 - 1.5|}{\frac{0.21}{\sqrt{100}}} = 4.8$$

From the table for the area under the standard normal curve, we find that the critical value at the 1 per cent level of significance is 2.58. Since the observed value of z is greater than 2.58 we reject the null hypothesis at 1% level and accept the alternative hypothesis that the average life of batteries is different from 1.5 volts.

19.4.2 Population Variance not Known

The assumption that population standard deviation (σ) is known to us is unrealistic, as we do not know population mean itself. When σ is unknown we have to estimate it by sample standard deviation (s). In such situations there are two possibilities depending upon the sample size. If the sample size is large ($n > 30$) we apply z-statistic, that is,

$$z = \frac{|\bar{x} - \mu|}{s/\sqrt{n}} \qquad \ldots(19.7)$$

In case the sample size is small ($n \leq 30$) we apply t-statistic with $n-1$ degrees of freedom. The test statistic is

$$t = \frac{|\bar{x} - \mu|}{s/\sqrt{n}} \qquad \ldots(19.8)$$

The steps you should follow are:

1) Specify the null hypothesis.
2) Find out whether it requires one-tail or two-tail test. Accordingly identify the rejection region in the standard normal curve. This will help in specification of the alternative hypothesis.
3) Check whether sample size is large ($n > 30$) or small ($n \leq 30$).
4) In case $n > 30$, apply z-statistic given at (19.7).
5) Find out from z-table the critical value according to level of significance (α).
6) In case $n \leq 30$, apply t-statistic given at (19.8).
7) Find out from t-table (given in Table 15.3 in Block 5) the critical value for $n-1$ degrees of freedom and level of significance (α).
8) If you obtain a value lower than the critical value do not reject the null hypothesis.
9) If you obtain a value greater than the critical value reject the null hypothesis and accept the alternative hypothesis

Example 19.2

A tablet is supposed to contain on an average 10 mg. of aspirin. A random sample of 100 tablets show a mean aspirin content of 10.2 mg. with a standard deviation of 1.4 mg. Can you conclude at the 0.05 level of significance that the mean aspirin content is indeed 10 mg.?

Here, the null hypothesis is H_0: $\mu = 10$

The rejection region is on both sides of 10 mg. Thus it requires a two-tail test and $H_A : \mu \neq 10$.

Also, the sample mean is $\bar{x} = 10.2$ and the sample size $n = 100$. Since population standard deviation is not known we estimate it by sample standard deviation s.

Since sample size is large, our test statistic is $z = \dfrac{|\bar{x} - \mu|}{s/\sqrt{n}}$. By applying relevant values from the sample we obtain

$$z = \frac{|\bar{x} - \mu|}{\frac{s}{\sqrt{n}}} = \frac{|10.2 - 10|}{\frac{1.4}{\sqrt{100}}} = 1.43$$

At 5 per cent level of significance the critical value of z is 1.96. Since the z value that we have obtained is less than 1.96, we do not reject the null hypothesis. Therefore the mean level of aspirin is 10 mg.

Example 19.3

The population of Haripura district has a mean life expectancy of 60 years. Certain health care measures are undertaken in the district. Subsequently, a random sample of 25 persons shows an average life expectancy of 60.5 years with a standard deviation of 2 years. Can we conclude at the 0.05 level of significance that the average life expectancy in the district has indeed gone up?

Here, H_0: $\mu = 60$

We have to test for an increase in life expectancy. Thus it is a case of one-tail test and the rejection region will be on the right-hand tail of the curve.

Hence our alternative hypothesis is

$H_A : \mu > 60$

Here population standard deviation σ is not known and we estimate it by the sample standard deviation s. Here the sample size is small hence we have to apply t-statistic given at (19.8).

$$t = \frac{|\bar{x} - \mu|}{\frac{s}{\sqrt{n}}} = \frac{|60.5 - 60|}{\frac{2}{\sqrt{25}}} = 1.25$$

Since sample size is 25, degrees of freedom is $25 - 1 = 24$. From the t-table we find that for 24 degrees of freedom, 5 per cent level of significance, and one-tail test the t-value is 1.71.

Since t-value obtained above is less than the critical value we do not reject the hypothesis. Therefore, life expectancy has not changed Therefore, we accept the alternative hypothesis that life expectancy for the district has not changed after the health care measures.

Check Your Progress 2

1) A report claimed that in the 'School Leaving Examination', the average marks scored in Mathematics was 78 with a standard deviation of 16. However, a random sample of 37 students showed an average of 84 marks in Mathematics. In the light of this evidence, can we conclude that actually the average was more than 78? Use 0.05 level of significance.

.....

.....

.....

.....

.....

2) A passenger car company claims that average fuel efficiency of cars is 35 km. per litre of petrol. A random sample of 50 cars shows an average of 32 km. per litre with a standard deviation of 1.2 km. Does this evidence falsify the claim of the passenger car company at 0.01 level of significance?

.....

.....

.....

.....

.....

.....

3) A random sample of 200 tins of coconut oil gave an average weight of 4.95 kg per tin with a standard deviation of 0.21 kg. Do we accept the hypothesis of net weight of 5 kg per tin at 0.01 level of significance?

.....

.....

.....

.....

.....

.....

4) According to a report, the national average annual income of the government employees during a recent year was Rs. 24,632 with a standard deviation of Rs. 1827. A random sample of 49 government employees during the same year shows an average annual income of Rs. 25,415. On the evidence of this sample, at 0.05 level of significance, can we conclude that the national average annual income of government employees is indeed Rs. 24,632?

.....

.....

.....

.....

.....

.....

19.5 TESTS FOR DIFFERENCE BETWEEN TWO SAMPLES

Many times we need to test for the difference between two samples. The objective may be to ascertain whether both samples are drawn from the same population or to check whether a particular characteristic is the same in two populations. For example, we formulate a hypothesis that the production per worker in plant A is the same as the production per worker in plant B. We discuss below the procedure for testing of such a hypothesis.

Here again we deal with two different situations: whether variance of both the populations are known. Another consideration is sample size: large or small.

The null hypothesis is the statement that population means of both the populations are the same. In notations

$$H_0 : \mu_1 = \mu_2 \qquad \ldots(19.9)$$

The alternative hypothesis is the statement that both the population means are different. In notations

$$H_A : \mu_1 \neq \mu_2 \qquad \ldots(19.10)$$

19.5.1 Population Variance is Known

When standard deviations (positive square root of variance) of both the populations are known we apply z statistic specified as follows:

$$z = \frac{\left|(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad \ldots(19.11)$$

In (19.11) above, subscript 1 refers to the first sample and subscript 2 refers to the second sample. By applying relevant data in (19.11) we obtain the observed value of z and compare it with the critical value for specified level of significance.

Example 19.4

A bank wants to find out the average savings of its customers in Delhi and Kolkata. A sample of 250 accounts in Delhi shows an average savings of Rs. 22500 while a sample of 200 accounts in Kolkata shows an average savings of Rs. 21500. Tt is known that standard deviation of savings in Delhi is Rs. 150 and that in Kolkata is Rs. 200. Can we conclude at 1 per cent level of significance that banking pattern of customers in Delhi and Kolkata is the same?

In this case the null hypothesis is $H_0 : \mu_1 = \mu_2$

and the alternative hypothesis is $H_A : \mu_1 \neq \mu_2$

We are provided with the information that

$\bar{x}_1$ = Rs. 22500 $\qquad$ σ_1 = Rs. 150

$\bar{x}_2$ = Rs. 22400 $\qquad$ σ_2 = Rs. 200

$n_1 = 250$ $\qquad$ $n_2 = 200$

Since σ_1 and σ_2 are known we apply z-test.

The test statistic is $z = \dfrac{|(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)|}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$.

By applying the information provided above we obtain

$$z = \frac{|22500 - 22400|}{\sqrt{\dfrac{150^2}{250} + \dfrac{200^2}{200}}} = \frac{100}{\sqrt{90 + 200}} = 5.87$$

We find that at 1 per cent level of significance the critical value obtained from Table 19.2 is 2.58.

Since the observed value of t is greater than the critical value of t the null hypothesis is rejected and the alternative hypothesis is accepted. Thus the banking pattern of customers in Delhi and Kolkata are different.

19.5.2 Population Variance is not Known

When population variance (σ^2) is not known we estimate it by sample variance (s^2). If both samples are large in size (n>30) then we apply z statistic as follows:

$$z = \frac{|(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)|}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \qquad \ldots(19.12)$$

On the other hand, if samples are small in size ($n \leq 30$) then we apply t-statistic as follows:

$$t = \frac{|(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)|}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \qquad \ldots(19.13)$$

Degrees freedom for t-test = $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$

Example 19.5

A mathematics teacher wants to compare the performance of Class X students in two sections. She administers the same set of questions to 25 students in Section A and 20 students in Section B. She finds that Section A students have a mean score of 78 marks with standard deviation of 4 marks while Section B students have a mean score of 75 marks with standard deviation of 5 marks. Is the performance of students in both Sections different at 1 per cent level of significance?

In this case the null hypothesis is $H_0 : \mu_1 = \mu_2$

and the alternative hypothesis is $H_A : \mu_1 \neq \mu_2$

We are provided with the information that

$\bar{x}_1 = 78$ $\quad$ $s_1 = 4$

$\bar{x}_2 = 75$ $\quad$ $s_2 = 5$

$n_1 = 25$ $\quad$ $n_2 = 20$

Since σ_1 and σ_2 are not known and sample sizes are small we apply *t*-test.

$$t = \frac{|(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{|78 - 75|}{\sqrt{\frac{4^2}{25} + \frac{5^2}{20}}} = \frac{3}{1.37} = 2.18$$

The degrees of freedom in this case is 25+20–2 = 43.

We find out from Table 15.3 that at 1 per cent level of significance the *t*-value for 43 degrees of freedom is 2.69.

Since the critical value of *t* is less than the observed value of *t* we reject the null hypothesis and accept the alternative hypothesis. Therefore, students in Section A and Section B are different with respect to their performance in mathematics.

Check Your Progress 3

1) You are given the following information.

n_1=50 n_2=50

$\bar{x}_1$=52.3 $\bar{x}_2$=52.3

$\sigma_1 = 6.1$ $\sigma_2 = 6.1$

Test the following hypothesis.

$H_o : \mu_1 = \mu_2$

$H_A : \mu_1 \neq \mu_2$

$\alpha = 0.05$

.....

.....

.....

.....

2) From two normal populations two samples are drawn. The following information is obtained.

n_1=15 n_2=10

$\bar{x}_1$=140 $\bar{x}_2$=150

$s_1 = 10$ $s_2 = 15$

Test the hypothesis at 1% level of significance that there is no difference between both the populations.

.....

.....

.....

.....

3) Suppose that samples of size n_1=20 and n_2=15 are drawn from two normal populations. The sample statistics are as follows:

$\bar{x}_1$=110 $\bar{x}_2$=125

$s_1^2 = 225$ $s_2^2 = 150$

Can we conclude at the 5% level of significance that $\mu_1 < \mu_2$?

.....

.....

.....

.....

.....

19.6 LET US SUM UP

In the present Unit we discussed about the methods of testing a hypothesis and drawing conclusions about the population. Hypothesis is a statement about a parameter. In order to test a hypothesis we formulate test statistic on the basis of the information available to us. In this Unit we considered two situations: i) description of a single sample, and ii) comparison between two samples.

Construction of the test statistic depends on the knowledge about the population variance and sample size. When population variance is known to us or the sample size is large we apply normal distribution and use z statistic to test the hypothesis. On the other hand, when we do not know the population variance and sample size is small we construct the test statistic on the basis of t distribution. Remember that for large samples t distribution approximates normal distribution and therefore we can use z statistic.

19.7 KEY WORDS

Estimator : It is another name given to statistic in the theory of estimation.

Parameter : It is a measure of some characteristic of the population.

Problem of Hypothesis Testing : Some times we may have some information about certain features of the population and we want to examine whether the information is tenable in the light of a random sample drawn from the population. This problem of statistical inference is called the problem of hypothesis testing.

Sampling Distribution : It is the relative frequency or probability distribution of a statisatic.

Sampling Error : In the sampling method, we try to approximate some feature of a given population from a sample drawn from it. Since in the sample all the members of the population are not included, howsoever close the approximation is, it is not identical to the required population feature and some error is committed. This error is called the sampling error.

Sampling Fluctuation : It is the variation in the values of a statistic computed.

Statistic : It is a function of the values of the units that are included in the sample. The basic purpose of a statistic is to estimate some population parameter.

Standard Error : It is the standard deviation of the sampling distribution of a statistic.

19.8 SOME USEFUL BOOKS

Nagar, A. L. and Das, R. K., 1989, *Basic Statistics:* Oxford University Press, Delhi, Chapter 9.

Newbold, P., 1991. *Statistics for Business and Economics* (Third Edition): Prentice Hall, New Jersey, Chapters 6, 7, 8 and 9.

Keller, G, and B. Warrack, 1991, *Essentials of Business Statistics*, Wordsworth Publishing Co., California, Chapters 7 and 8.

19.9 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Go through Sections 18.2 and 18.3, and answer.
2) It is large sample and σ is unknown. It requires one-tail test. Thus rejection region is to the right hand tail of standard normal curve. Accordingly draw the diagram.

Check Your Progress 2

1) Since it is large sample with known variance, we apply z-statistic. Since alternative hypothesis is $\mu > 78$, we apply one-tail test. The observed value of z is 2.28 and critical value of z at 5% level of significance is 1.65. Since the observed value is greater than the critical value we reject the null hypothesis. Therefore, we conclude that the average marks was more than 78.
2) It is a large sample with unknown variance. It requires two-tail test. The observed value of z is 17.68 and critical value of z at 1% level of significance is 2.58. Since the observed value is greater than the critical value, the null hypothesis is rejected.
3) It is a large sample with unknown variance. Requires two-tail test with z-statistic. Observed value of z is 3.37. Null hypothesis is rejected.
4) Since it is large sample with known standard deviation, we apply z-statistic. Requires two-tail test. Observed value of z is 3.00. critical value of z at 5% level of significance is 2.58. null hypothesis is rejected. Therefore, the national average of annual income of government employees is different from Rs. 24632.

Check Your Progress 3

1) The samples sizes are large and population standard deviations are known. Hence we apply z-statistic and observed value of z is 2.58. Since the alternative hypothesis is $\mu_1 \neq \mu_2$, we have a two tail test at $\sigma = 0.05$ the critical value of z is 1.96. Null hypothesis is rejected.
2) The sample sizes are small and σ is not known. Hence we apply t-statistic and observed value of t is 0.61. The hypothesis are H_o: $\mu_1 = \mu_2$ against H_A: $\mu_1 \neq \mu_2$. Thus, it requires a two-tail test. For 23 d.f., at 1% level of significance, the critical value of t is 2.50. H_o is not rejected.
3) The sample sizes are small and σ is not known, t-statistic is applied. Observed value of t is 0.72. H_o is $\mu_1 = \mu_2$ and H_A is $\mu_1 < \mu_2$. Hence one-tail test is required. Thus critical value of t at 33 d.f. for 5% level of significance is 2.00. H_o is not rejected.

UNIT 20 CHI-SQUARE TEST FOR NOMINAL DATA

Structure

20.0 OBJECTIVES

After going through this Unit you will be in a position to:

- present nominal data in the form of contingency table;
- explain the chi-square test statistic; and
- apply chi-square test to contingency tables.

20.1 INTRODUCTION

In the previous two Units we discussed the procedures of drawing conclusions about population parameters on the basis of sample information. In many cases, particularly for nominal variables, we do not have parameters. Here a variable (or attribute) assumes values pertaining to a finite number of categories and we can count the number of observations in each category. Drawing inferences in the case of nominal data is the subject matter of the present Unit.

The method of hypothesis testing discussed in the previous Unit requires certain assumptions about the population from which the sample is drawn. For example, application of t-test for small samples requires that the parent population is normally distributed. Similarly the hypothesis is formulated by specifying a particular value for the parameter. Hence these tests are called parametric tests.

When it is not possible to make any assumption about the value of a parameter the test procedure described in the previous Unit fails. In situations where the population does not follow normal distribution or where it is not possible to specify the parameter value, we use non-parametric tests.

There are many types of non-parametric tests depending upon our need. However, we confine ourselves to a common procedure, that is, chi-square test for test of independence between variables.

20.2 CONTINGENCY TABLE

Contingency table is a rectangular table in which observations from the population are classified according to two characteristics. It is also called a two-way table, which is

discussed in Unit 7. Recall that qualitative data can be arranged into categories and presented in the form of a two-way table.

In order to explain the application of chi-square test let us take a concrete example. In Unit 7 we had given an example on occupation of father and number of children. We divided occupation into five categories - i) unemployed, ii) unskilled labour, iii) skilled labour, iv) self-employed, and v) professional. Similarly we divided families into five categories according the number of children - i) no child, ii) one child, iii) two children, iv) three children, and v) more than three children. For a sample of 650 families the data obtained is presented in Table 20.1.

Table 20.1: Observed Frequency on Occupation and Number of Children

Number of Children	*Occupation*					*Total*
	Unemployed	*Unskilled Labour*	*Skilled Labour*	*Self-Employed*	*Professional*	
	(1)	(2)	(3)	(4)	(5)	
0	10	15	10	12	11	58
1	35	25	17	18	25	120
2	22	33	45	40	43	183
3	11	40	48	58	30	187
≥ 4	11	33	30	19	9	102
Total	89	146	150	147	118	650

Table 20.1 is a called contingency table, because we are trying to find whether the number of children is contingent upon the occupation of the father.

Our purpose is to test for possible relationship between the number of children and the occupation of father. Thus the null hypothesis is specified as

H_0 : Number of children and occupation of father are independent

against the alternative hypothesis

H_A : Number of children and occupation of father are dependent

In Table 20.1 we have presented the observed frequency for each cell in the table. What should be the expected frequency when there is no relationship between the variables under consideration? We will answer this question below.

20.3 EXPECTED FREQUENCIES

As mentioned above, expected frequency is calculated under the assumption that there is no relationship between the number of children and the occupation of father. For each cell in Table 20.1 the expected frequency is obtained by multiplying the sample size n by the cell probability. In order to calculate the cell probability we first find out the marginal frequencies for each row and column. As you know from Unit 7 'row marginal' are given by the row totals. Similarly, 'column marginal' are given by column totals.

For the rows, we can find out the 'marginal row probability'. For row 1 the marginal row probability, $p(r_1)$, is given by

$$p(r_1) = \frac{58}{650} = 0.09 \qquad \ldots(20.1)$$

Marginal row probabilities for other rows are

$$p(r_2)=\frac{120}{650}=0.18 \qquad p(r_3)=\frac{183}{650}=0.28$$

$$p(r_4)=\frac{187}{650}=0.29 \qquad p(r_5)=\frac{102}{650}=0.16$$

Similarly for column 1 the marginal column probability, $p(c_1)$, is given by

$$p(c_1)=\frac{89}{650}=0.14 \qquad \ldots(20.2)$$

Marginal column probabilities for other columns are

$$p(c_2)=\frac{146}{650}=0.22 \qquad p(c_3)=\frac{150}{650}=0.23$$

$$p(c_4)=\frac{147}{650}=0.23 \qquad p(c_5)=\frac{118}{650}=0.18$$

Recall from Unit 13 that if events A and B are independent then the probability of joint occurrence of A and B is given by

$$p(A \cap B)=p(A).p(B)$$

Therefore, if we assume the null hypothesis to be true, then cell probability for the first cell (c_1, r_1) will be

$$p(r_1 \cap c_1)=p(r_1).p(c_1)=\frac{58}{650}\times\frac{89}{650}=0.0892\times 0.1369=0.0122 \qquad \ldots(20.3)$$

Hence, expected frequency for the first cell will be

$$E_{11}=n.p(r_1 \cap c_1)=650\times 0.0122=7.94 \qquad \ldots(20.4)$$

In general terms we can say that the expected frequency of cell ij is

$$E_{ij}=\frac{(\text{Row } i \text{ total})\ (\text{Column } j \text{ total})}{\text{Sample size}} \qquad \ldots(20.5)$$

By applying (20.5) we calculate the expected cell frequency for each cell and prepare a table as given in Table 20.2.

Table 20.2: Calculation of Expected Frequency for Each Cell

Number of Children		*Occupation*					*Total*
		Unemployed	*Unskilled Labour*	*Skilled Labour*	*Self-Employed*	*Professional*	
		c_1	c_2	c_3	c_4	c_5	
0	r_1	7.94	13.03	13.38	13.12	10.53	58.00
1	r_2	16.43	26.95	27.69	27.14	21.78	120.00
2	r_3	25.06	41.10	42.23	41.39	33.22	183.00
3	r_4	25.60	42.00	43.15	42.29	33.95	187.00
≥ 4	r_5	13.97	22.91	23.54	23.07	18.52	102.00
Total		89.00	146.00	150.00	147.00	118.00	650.00

The next step is to compare the observed frequency with the expected frequency.

20.4 CHI-SQUARE TEST STATISTIC

In order to compare the observed frequency with the expected frequency we construct the chi-square statistic, which is given by

$$\chi^2 = \Sigma \frac{(O_i - E_i)^2}{E_i} \qquad \ldots(20.6)$$

where O refers to observed frequency and E refers to expected frequency.

The chi-square statistic has degrees of freedom $(r-1)(c-1)$. For example, if there are 3 rows and 4 columns, then degrees of freedom is $(3-1)(4-1)=6$.

Let us summarise the steps to be followed in chi-square test. These are:

1) specify the null and alternative hypotheses
2) calculate the expected frequency for each cell by using (20.5)
3) calculate the observed value of χ^2 statistic by using (20.6)
4) determine the degrees of freedom according to the formula $(r-1)(c-1)$
5) check the level of significance (α) required
6) from Table (15.2) given in Unit 15, Block 5 find out the critical value of χ^2 for α and relevant degrees of freedom
7) compare the observed value of χ^2 with the critical value of χ^2
8) if the observed value is less than the critical value, then do not reject H_0
9) if the observed value is greater than the critical value, then reject H_0 and accept H_A

For the data given in Table 20.1 let us find out the observed value of χ^2.

Table 20.3: $\frac{(O_i - E_i)^2}{E_i}$ for each Cell

Number of Children		*Occupation*					*Total*
		Unemployed	*Unskilled Labour*	*Skilled Labour*	*Self-Employed*	*Professional*	
		c_1	c_2	c_3	c_4	c_5	
0	r_1	0.53	0.30	0.86	0.10	0.02	1.80
1	r_2	20.99	0.14	4.13	3.08	0.47	28.81
2	r_3	0.37	1.60	0.18	0.05	2.88	5.08
3	r_4	8.33	0.10	0.54	5.84	0.46	15.26
≥ 4	r_5	0.63	4.44	1.77	0.72	4.89	12.46
Total		30.85	6.58	7.48	9.77	8.72	63.41

Since there are 5 rows and 5 columns, the degrees of freedom is $(5-1)(5-1)=16$.

For 16 d.f. the critical value of χ^2 at 5 per cent level of significance (see Table 15.3) is 26.30. We find from Table 20.3 we find that the observed value of χ^2 is 63.41. Since the observed value is greater than the critical value we reject the null hypothesis and accept the alternative hypothesis. Therefore, we conclude that the variables 'number of children' and 'occupation of father' are dependent.

Check Your Progress 1

1) Explain the following concepts.

   a) marginal frequency

   b) cell probability

   c) expected frequency

   d) critical value of χ^2

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

2) There are three brands (orange, cola and lemon) of soft drinks produced by a company. A survey of 160 persons in two states (one form north- Punjab and one from south- Tamil Nadu) on the preference for soft drinkgs provides the following information.

	orange	*cola*	*lemon*
Punjab	33	26	31
Tamil Nadu	17	24	29

Test the hypothesis that there is no preference for particular brand of soft drink in both the states ($\alpha = 0.05$).

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

20.5 LET US SUM UP

In the case of qualitative data we cannot have parametric values. Therefore, hypothesis testing on the basis of z-statistic or t-statistic cannot be performed. Chi-square test is applied to such situations. Chi-square test is a non-parametric test, where no assumption about population is required. There are various types of non-parametric tests beside the chi-square test. Moreover, chi-square test can be applied to many situations besides contingency table.

In contingency table we test the null hypothesis that variables under consideration are independent of each other against the alternative hypothesis that variables are related. Here we compare expected frequency with observed frequency and construct the chi-square statistic. If the observed value of chi-square exceeds the expected value of chi-square we reject the null hypothesis.

20.6 KEY WORDS

Nominal Variable : Such a variable takes qualitative values and do not have any ordering relationships among them. For example, gender which assumes two qualitative values, male and female has no ordering in 'male' and 'female' status. A nominal variable is also called an attribute.

Contingency Table : A two-way table to present bivariate data. It is called contingency table because we try to find whether one variable is contingent upon the other variable

Degrees of Freedom : It refers to the number of pieces of independent information that are required to compute some characteristic of a given set of observations.

Expected Frequency : It is the expected cell frequency under the assumption that both the variables are independent.

20.7 SOME USEFUL BOOKS

Kiess, H. O., 1989, *Statistical Concepts for the Behavioral Sciences*, Allyn and Bacon, Boston.

Keller, G. and B. Warrack, 1991, *Essentials of Business Statistics*, Wadsworth Publishing Co., California.

20.8 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Go through the text and explain these terms.

2) The expected frequencies are

	orange	*cola*	*lemon*
Punjab	28.13	28.13	33.75
Tamil Nadu	21.88	21.88	26.25

The observed value of chi-square statistic is 2.98. Degrees of freedom is 2. The critical value of chi-square at 5 per cent level of significance at 2 degrees of freedom is 5.99. Hence null hypothesis is not rejected and soft drink consumption is independent of the region.