UNIT 6 MEASUREMENT OF VARIABLES

Structure

6.0 OBJECTIVES

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

- define a variable;
- distinguish between various types of variables;
- measure ordinal variables through scaling techniques;
- distinguish between census and sample survey;
- explain the steps involved in carrying out a sample survey; and
- distinguish between various types of sampling.

6.1 INTRODUCTION

During the course of research you come across situations where you have to measure various characteristics. These characteristics could be of various types, viz., age, height or income level of visitors to a library; educational qualifications, social status or reading habits of a person; gender, religion, or area of interest of a library user. Note that all these characteristics are not similar from the point of measurement. While age, height or income can be measured in quantitative terms (in number of years, in centimetres, in rupees) religion or gender can be put to certain categories only. In this Unit we discuss the issue of measurement of qualitative variables, particularly the scaling techniques.

Another issue discussed in the Unit pertains to collection of data on the basis of sample survey. Very often it is not possible to survey all the units bearing the characteristic under study. The constraints could be inadequate funds, time limit, and manpower. In such situations we survey only a subset of the population, called sample. We discuss various concepts associated with sampling procedure.

We begin with the types of variables and their measurement.

6.2 TYPES OF VARIABLES

Let us begin with the concept of a variable. It is a characteristic of the *sample* or the *population* that we intend to measure. Thus age of the reader is a variable, so is gender, educational level or mother tongue. As we have mentioned earlier all variables are not similar.

Variables can be of two types: qualitative and quantitative. Qualitative variable is one that cannot be expressed in numerical terms. For example, marital status is a qualitative variable. Here we can have two categories: married and single. Of course, if you want a more detailed categorisation you can further divide single in to widow/widower, divorcee and never married. Similarly, gender (male or female), mother tongue (Hindi, Bengali, Oriya, Tamil, Urdu, etc.), subject categories (economics, history, physics, medicine, etc.), religion (Buddhism, Christianity, Hinduism, Islam, etc.) are examples of qualitative variables. Here we study an attribute or quality that cannot be quantified, but can be divided into various categories. Moreover, we cannot say that one category is higher or greater than another category. Such variables are also called nominal variables.

There is another type of qualitative variable where we can divide the observations into various categories and also say that one category is higher or greater than another category. An example could be the educational qualification of a visitor to a library. Here we can divide the visitors on the basis of their educational qualification into categories such as 'secondary', 'senior secondary', 'graduate' and 'post-graduate'. In this case, obviously, the category 'Senior Secondary' is higher than the category 'Secondary' in terms of number of years of schooling and expected mental maturity. In this case we arrange the categories in an ascending or descending order. This sort of variables are called ordinal variables.

In the case of nominal variables we cannot perform any mathematical operations (such as addition, subtraction, multiplication, division,) or logical operations (greater than, less than) across categories. We can simply count the number of observations in each category. In the case of ordinal variables we can say that one category is greater than another category. But we cannot quantify the difference between categories. For example, we cannot express numerically the difference between two categories (say secondary and senior secondary). Also we cannot say that the difference between two categories (say secondary and senior secondary) is the same as the difference between two other categories (say graduate and post-graduate).

A quantitative variable can be expressed in numerical terms. Hence it is also called numerical variable. Examples of numerical variable could be age, income, weight, height, distance travelled, etc. This category of variables can be subjected to various mathematical and logical operations. Thus we can express the monthly income of a librarian in Rupees and also say by what percentage it exceeds the salary of a library assistant.

Numerical variables can be of two types: discrete and continuous. Discrete variable is one where the observations assume values in complete numbers. For example, the number of children in family can only be whole numbers; it cannot be fractions. On the other hand, continuous variables can assume any value in an interval. For example, weight of a person can be measured to any precision and thus can take any value in between two points.

Let us distinguish between variable and data. We obtain data by measuring a variable (qualitative or quantitative) on certain individuals or units. For example, suppose we measure the height of 50 employees in a library. Here height is the variable and we

Thus if 50 observations. These 50 numerical values that we obtain are our data. Thus we have discrete data or continuous data depending upon whether the variable is discrete or continuous.

Similarly there are primary data and secondary data. Primary data refers to data collected by the researcher by undertaking a field survey. On the other hand, secondary data refers to collection of data from published sources, e.g., census, budget, handbooks, etc. Thus when you undertake a field survey, collect data, analyse the results and present it in some forum, it is primary data. But when I use that data for further analysis it becomes secondary data for me.

6.3 MEASUREMENT OF QUALITATIVE DATA

The measurement of quantitative data does not pose problems as these are expressed in numerical terms. Measurement of qualitative data, however, is a complex issue and needs to be discussed further. There are two types of variables that are usually measured: i) social behaviour and personality, and ii) cultural and social environment. The purpose is to study socioeconomic status and its impact on various issues, attitude towards a particular event or behaviour, participation in certain activity, etc.

In order to measure these types of issues no readymade scale of measurement is available. Secondly, there is a lot of ambiguity and inter-personal variability in the definition of these qualities. However, attempts have been made to construct 'attitude scales' to measure these qualities. These scales are in the form of equal-appearing intervals. Such scales are widely used in sociology, psychology, education and administration apart from other branches of study.

In designing such scales we should keep one thing in mind that it should be 'valid' and 'reliable'. In other words, i) it should measure what we intend to measure, and ii) it should yield the same consistent results when applied under the same conditions.

In designing attitude scales we form a number of statements and ask the respondents to react to these statements. The statements should be brief, unambiguous and relevant. It should be expressed in such a form that it could be endorsed or rejected in terms of definitely expressed attitude.

A widely used scale is 'Likert Scale' which is also referred to as 'technique of summated ratings'. The basic steps in construction of Likert scale are given below:

A series of propositions representing attitudes are compiled. The attitudes of persons towards these propositions could range from extremely negative to extremely positive.

The statements or propositions express values rather than facts. Each statement indicates the position of a person towards the issue concerned. For example, a statement could be 'there should be separate reading rooms for boys and girls in a library'. The response to such a statement could range from 'strongly agree' to 'strongly disagree' depending upon the attitude of the respondent towards the issue.

Each statement is so formulated that the response of persons can be given in any of the five terms such as *i) strongly approve, ii) approve, iii) undecided, iv) disapprove, v) strongly disapprove*. Many times expressions such as *i) strongly agree, ii) agree, iii) cannot say, iv) disagree, v) strongly disagree* are also used. Although a five-point continuum is common, there can be three-, four-, or six-point continuum also. For example, a seven-point continuum could be *i) always, ii) almost always, iii) frequently, iv) occasionally, v) rarely, vi) almost never, vii) never*.

Weights are assigned to each response for a statement. For example, in a five-point continuum the weights that we attach to each response could be 1, 2, 3, 4, and 5. These weights indicate the intensity of the attitudes of a respondent.

When we have a number of statements, we can arrange the statements according to their relative intensity. Also groups of individuals can be arranged in a rank order on the basis of scale scores. Moreover, the scale scores can be used to predict any outside variable.

Self Check Exercise

1) State the type of variable (nominal, ordinal and numerical) for the following:

   a) tribes in India

   b) height of children

   c) number of books

   d) accession number of books

   e) subject codes

2) With an example explain the use of Likert scale.

Note: i) Write your answers in the space given below.

ii) Check your answers with the answers given at the end of the Unit.

.....

.....

.....

.....

6.4 CENSUS VERSUS SAMPLE SURVEY

There are quite a few methods for collection of data that we will discuss in detail in Block 3 (questionnaire method, personal interview, participatory observation, etc.). An important issue is whether we should collect data on all units present or only a part of it. For example, if we wanted to know the reading habits of economics students of Sambalpur University and for that purpose we designed a questionnaire. We found that there are 800 economics students and it would not be possible to survey all the students due to time and money constraints. Thus, we decided to administer the questionnaire to 100 students only.

In formal statistical language we distinguish between 'population' and 'sample'. Since our objective is to study the reading habits of economics students in Sambalpur University, all 800 economics students qualify to be studied by us and thus constitute our population. If we decide not to collect data on all students but study only 100 students, these selected 100 students constitute our sample.

Thus, population is a collection of individuals or objects having the desired characteristics we are interested in a sample which is a part of the population. Obviously we can draw more than one sample from the same population. If you go to Sambalpur University on a different day and select 100 students your sample may be different from what another person had selected.

Collection of information on all units of a population is called census. On the other hand, collection of information on a sample is called sample survey. The process of selection of a sample is called 'sampling'. The advantages of sample survey are:

- Sample survey is less expensive than a census.
- Sample survey requires lesser time and manpower than a census.
- Sample survey can be monitored closely and more accurate information can be colleted than in census.
- Many times we do not require information on all units. For finding out the effectiveness of a new medicine we do not have to test it on all patients; only a representative sample will do.

In sampling we are surveying only a part of the population. Remember that we do not know the exact population characteristics, as we do not survey the population. But there are scientific methods by which we can estimate a population characteristic (called parameter) on the basis of sample characteristic (called statistic). This topic will be covered in Unit 17 while dealing with estimation and hypothesis testing.

There are two types of errors in a survey: sampling error and non-sampling error. Sampling error is due to the fact that only a part of the population is being surveyed. It is the difference between the *parameter* and the *statistic*. It can be reduced by adoption of scientific sampling procedure, specifically probability sampling so that a random sample is selected. Non-sampling error is due to errors in measurement, non-response by the selected units, wrong recording of data, and personal bias of the investigators. Thus, non-sampling error is present in both census and sample survey. In a census, however, many investigators are involved and large number of units are to be surveyed. So non-sampling error could be very high. On the other hand, in a sample survey greater care can be taken in collection and recording of data since lesser number of units are surveyed. Thus non-sampling error can be minimised in a sample survey. The total error is the sum of sampling and non-sampling errors. It may so happen that the non-sampling error in a census is more than the sampling and non-sampling errors in a sample survey! In such cases sample survey would give better results than census.

Self-Check Exercises

3) What are the advantages of sampling over census?

4) Distinguish between sampling error and non-sampling error.

Note: i) Write your answers in the space given below.

ii) Check your answers with the answers given at the end of the Unit.

.....

.....

.....

.....

6.5 SAMPLING PROCEDURE

While carrying out a sample survey there are certain steps that we should follow. These are as follows:

Specification of objectives: The objectives of carrying out sampling is the first and foremost step in sampling procedure. Because all other steps will follow from this objective.

Definition of population: In this step we define the units that should be included in the population. Many times there are certain border cases where proper definition is important. For example, if you want to survey the library personnel, it has to be decided whether you should include part-time employees in the population or not.

Preparation of 'sampling frame': Once you have defined the units to be included in the population, the next step is to prepare a list of the units from which the sample is to be drawn. Many times problems come up because the sources from which you want to prepare the sampling frame may be incomplete or obsolete.

Identification of sampling procedure: Sampling procedure refers to the method of selecting the sample. There are quite a few sampling procedures available. We should select a method that: i) gives us a representative sample, ii) is feasible to carry out keeping in view our constraints, and iii) is cost effective. There are broadly two types of sampling: probability sampling and non-probability sampling. We will discuss these types in the next Section.

Determination of sample size: The next step is the determination of sample size. The factors that influence sample size are i) population size, ii) variance of units, iii) desired precision level, iv) response rate, and v) availability of resources. The sample size should be relatively larger if population size is larger, variability among units is higher, more precise results are required, and lesser is the response rate. However, we should settle for smaller sample size if there are constraints such as availability of funds or manpower or time. Sample is considered to be large if its size is greater than 30. If sample size is less than or equal to 30 it is considered as small sample. The procedures of drawing inferences (See Unit 17) for large and small samples are different.

Selection of sampling units: Once you have decided on the sampling procedure and the size of the sample, the next step is draw the sampling units (that is, units included in the sample) from the sampling frame.

6.6 TYPES OF SAMPLING

As mentioned earlier there are two types of sampling: probability and non-probability. In probability sampling some element of randomness is involved in selection of units, so that personal judgement or bias is not there. Here we have simple random sampling, systematic random sampling, stratified random sampling and multistage sampling. Remember that the word 'random' here does not mean haphazard or unsystematic, rather it implies lack of bias or control in selection. In non-probability sampling we have judgement sampling, quota sampling, and snowball sampling.

Simple Random Sampling (SRS): It is the basic sampling procedure where each unit in the population gets an equal chance of being included in the sample. There are two commonly used methods to draw a simple random sample, viz., i) lottery method, and ii) random numbers selection method. In both the methods we assign a unique number to each unit in the sampling frame. In lottery method we mix up the numbers very well and draw the numbers one by one. In random number selection method we refer to 'random number tables (RNT)' available from various sources (including the Internet) and select the units which are there in the RNT. Simple random sampling should be used with a homogeneous population. That is, all the units in a population possess the same attributes that the researcher is interested in measuring. The characteristics of

homogeneity may include age, sex, income, social status, geographical region, etc. If the population is not homogeneous then we need a larger sample size to obtain a representative sample. But we can divide the population into various homogeneous strata and obtain a stratified random sample. A limitation of the simple random sampling is that if the population size is too large, selection of the sampling units is time consuming. These days, of course, computer has been very useful in selection of random numbers and reduced time spent in selection process.

Systematic Random Sampling: In this case we select the units in a fixed interval. For example, in a library you have to check the quality of maintenance of books. For this purpose you decide take physical verification of 100 books out of the 5000 books available in the library. Here population size is 5000 and sample size is 100. In systematic random sampling procedure you take a random starting point and then select the 50^{th} subsequent unit. Thus if 12^{th} book is the starting point, then you go on checking books located at the positions: 62^{nd}, 112^{th}, 162^{nd}, …… and so on. This procedure is useful when preparation of sampling frame is difficult or not possible.

Stratified Random Sampling: This procedure is practiced when the population is not homogeneous but can be divided into various homogeneous groups (called 'strata'). Here we select sub-samples from each stratum and add them together to obtain our desired sample. Therefore, the stratified random sampling procedure is a better representative of the population and sampled units reflect the population characteristics more accurately. The steps we should follow are:

- Divide the population into strata based on some characteristic chosen by you (example, Rural/Urban, Male/Female, etc.)
- The number of units taken from each stratum is proportional of the relative size of the stratum and standard deviation of the characteristic within the stratum. If stratum size is then a larger sub-sample should be taken. Similarly if you find that variability among units is more in a stratum than other strata, then you should take a larger sub-sample from that.
- Choose the sub-sample from each stratum using simple random sampling.

Multistage Sampling: As the name suggests, sampling procedure in this case is divided into two or more stages. Let us take an example of a two-stage sampling. Suppose you have to survey the aptitude of college students, who can be categorised into three streams: arts, science and commerce. Here in the first stage you divide colleges according to geographical regions and you select specified number of colleges from each region on the basis of random sampling. In the second stage you select students by a random sampling method from the colleges selected at the first stage only, not from all colleges. Let us try to explain the difference between two-stage sampling and stratified sampling. In stratified random sampling you divided the population (all college students) into there strata: arts, science and commerce. From each stratum you selected a sub-sample by simple random sampling procedure. The selected students can be from any college and you cannot rule out a visit to any college. In two-stage sampling you are excluding certain colleges at the first stage itself, which may prove to be economical.

Judgement Sampling: It is a non-probability sampling procedure. It is also called purposive sampling, where 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.

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 to the size of the group in the population. Do not confuse the quota sampling with stratified random sampling discussed earlier. In stratified random sampling you select random samples from each stratum or group whereas in quota sampling the researcher/interviewer has a fixed quota and selection of units is based on judgement. Quota sampling has the advantage that cost and time involved in selection of units is reduced considerably. However, it has many disadvantages also. In quota sampling, the samples are selected according to the convenience of the investigator. Therefore, the selected sample may be biased.

Snowball Sampling: In this procedure you begin by identifying someone who meets the criteria for inclusion in your study. You then ask them to recommend others who they may know 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 you are trying to reach population that are inaccessible or hard to find. For example, if you are studying the homeless or people suffering from a particular disease you are may not be able to find good lists respondents. However, if you identify one or two, you may find that they know others in a similar condition.

Self Check Exercise

5) What are the important steps in carrying out a sample survey?

6) Explain the procedure of drawing a stratified random sample.

7) Explain the following concepts:

   a) Systematic random sampling

   b) Parameter and statistic

   c) Multistage sampling

Note: i) Write your answers in the space given below.

ii) Check your answers with the answers given at the end of the Unit.

.....

.....

.....

.....

6.7 SUMMARY

A variable is a characteristic that we are interested in analysing. It can be nominal, ordinal or numerical. Numerical variable can be discrete or continuous. While numerical variable is amenable to mathematical and logical operations nominal and ordinal variables are not. In certain cases it is difficult to measure ordinal variables because we do not

have standard measurement. In such cases we need to construct a measurement scale. In this we discussed one such scaling techniques, that is summated rating.

Another issue that we covered in this Unit relates to sampling techniques. Because of inadequate resources or infeasibility we often resort to sampling instead of census of all the units in the population. There are two broad types of sampling procedures: probability sampling and non-probability sampling. In probability sampling we have simple random sampling, systematic random sampling, stratified random sampling and multistage sampling. On the other hand, non-probability sampling procedures are judgment sampling, quota sampling and snowball sampling.

6.8 ANSWERS TO SELF CHECK EXERCISES

1) a) Nominal, b) Numerical, c) Numerical, d) Nominal, and e) Nominal

2) Example of the use of Likert Scale: Show your agreement with the following:

   Libraries will ever remain important for the advancement of the society

   Strongly Agree; Agree; Undecided; Disagree; Strongly Disagree

3) Sampling is advantageous over statistic in that it saves cost and time to be devoted to the survey. It involves fewer personnel to be deployed and also results in more precision in the results.

4) Sampling error is due to the faulty sample selected. It may be due to non- probability sampling techniques adopted. Non- sampling errors are due to the errors in data measurement or analysis methods.

5) The steps to be followed in carrying out a sample survey are:
   specification of objective, preparation of sampling frame, identification of sampling procedure, determination of sample size, and selection of sampling units.

6) The procedure of drawing a stratified random sample is:

   - Divide the population into strata based on some characteristic chosen by you (example, Post graduate/ Under graduate, Male/Female, etc.)
   - Decide the number of units to be taken from each stratum proportional to the relative size of the stratum and standard deviation of the characteristic within the stratum. If stratum size is then a large a sub-sample should be taken. Similarly if you find that variability among units is more in a stratum than other strata, then you should take a larger sub-sample from that.
   - Choose the sub-sample from each stratum using simple random sampling.

7) a) Systematic random sampling is a type of random sampling where the bias is minimised. Here the random sample is selected in a systematic way, e.g., in case a sample of 100 is to be selected from a population of 1000, and then we may select the first member of the population and every subsequent 20^{th} member.

   b) Parameter is a summary value of the population while statistic is that of the sample.

   c) Multistage sampling is sampling done in two or more stages. In case we have to survey reading habits of users in public libraries, we may first take a random sample on the basis of geographical regions and then further take a sample on the basis of age groups. This is an example of multistage sampling.

6.9 KEYWORDS

Convenience Sampling : It refers to the method of obtaining a sample that is most conveniently available to the researcher.

Judgement Sampling : In this sampling procedure the selection of sample is based on the researcher's judgement about some appropriate characteristics required of the sample units.

Multistage Sampling : The sample selection is done in a number of stages.

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.

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.

Sample : It is a sub-set of the population. Therefore, it is a collection of some units from the population.

Simple Random Sampling: This is the basic sampling procedure where all units in the population have an equal chance of being included in the sample.

Snowball Sampling : It relies on referrals from initial sampling units to generate additional sampling units.

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 parameter.

Stratified Random Sampling : In this sampling procedure the population is divided into groups called strata. Subsequently sub-samples are selected from each stratum using a random sampling method.

Systematic Sampling : In this procedure the units are selected from the population at uniform interval (in time, order or space).

6.10 REFERENCES AND FURTHER READING

Bhardwaj, R. S. (1999). *Business Statistics.* New Delhi: Excel Books.

IGNOU Study Material (2005). *EEC-13: Elementary Statistical Methods and Survey Techniques*, Block 6.

Kothari, C.R. (1985). *Research Methodology: Methods and Techniques*. New Delhi: Wiley Eastern.

Young, P. V. (1988). *Scientific Social Surveys and Research*, Prentice Hall of India: New Delhi.

UNIT 7 DATA PRESENTATION

Structure

7.0 OBJECTIVES

After going through this unit you will be able to:

- arrange discrete data in a tabular form;
- identify class intervals and arrange continuous data in tabular form;
- present data in the form of histogram and frequency curves;
- arrange data in the form of bar diagrams; and
- represent data in the form of pie chart.

7.1 INTRODUCTION

Let us imagine a situation where you have to collect data on the number of visitors to a library and their age, income level, gender, education level and areas of interest. Your objective is to revamp the library policy. For example, you may like to decide on a membership fee keeping in view the ability to pay of users, or to build stock of books depending on the age profile and interest of users, or to increase the sitting capacity on the basis of the number of visitors, or to provide for separate facilities on the basis of gender. In order to carry on such a project you requested each member visiting the library to enter the above particulars on a register and compiled the same on a daily basis. No doubt you end up with a large volume of data very soon. Unless these data are subjected to statistical analysis it would not be of much help in policy formulation.

Once data collection is completed, your efforts should be geared towards bringing these raw data into a presentable form. You may think of presentation of the data in the form of a table or a chart. The task before you is to decide on the structure of the table or the shape of the graph. Let us look in to the options available.

In the previous Unit we had classified variables into nominal, ordinal and numerical types. Let us begin with tabulation of data pertaining to different types of variables and later on we will move on to graphical presentation. The first step in the analysis and interpretation of data is its classification and tabulation. The process of arranging data into groups according to their common characteristics is known as its classification. On the other hand tabulation implies a systematic presentation of data in rows and columns according to some salient features or characteristics.

In this Unit we will first discuss tabulation of data pertaining to discrete and continuous variables. Subsequently we will take up graphical presentation. There are basically three forms in which data can be presented: i) graphs such as line graph, histogram, frequency polygon and frequency curve, ii) bar diagrams (single bars, component bars, multiple bars), and iii) pie charts. We explain each of these through appropriate examples.

7.2 PREPARATION OF A TABLE

There are certain things we should take care of while preparing a table.

- It is required to give a table number for identification of the particular table.
- There should be a title of the table that indicates the type of information contained in the table. Title should be brief and precise. Avoid expressions like 'Table presents…' or Table contains….'
- If necessary give a head note. It should be given in parentheses and should appear on the right side top just below the title. See, for example, the expression (in Rupees) given in Table 7.4.
- Stub head describes the nature of stub entry, e.g., 'class interval' in Table 7.4
- Stub entries describe the rows.
- Caption describes the nature of data presented in columns followed by column heads and sub-heads. In certain tables it may not be necessary to give sub-heads.
- The main body of the table contains numerical information.
- Below the table there may be footnote. The purpose of footnote is to caution the readers about the limitations of the table.
- Source of the table may be the last component. It is quite important in the case of secondary data. It provides opportunity to the readers to check the data if they desire and get more of it.
- Remember that you have to design your own table, keeping your requirements in view. In Table 7.1 we have summarised different parts of a table.

Table No. 7.1: (————————TITLE————————)

(Head note)

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

Footnote:
Source:

7.3 TABULAR PRESENTATION

7.3.1 Nominal and Ordinal Data

You may recall that nominal and ordinal data can be classified into categories. Thus we can count the number of observations in a category, note them down and present in a table. Such an arrangement of data is called 'frequency distribution' because we are counting the frequency with which each category is repeated.

Let us take a concrete example. Suppose, in order to collect data on areas of interest of visitors to the library you identified 5 subject areas, viz., economics, history, political science, sociology and public administration. Thus there are five categories. Very often the number of observations may be large enough. In such situations we use tally bars. In this method all the categories (in our example, five) are written in a column. For every observation, a tally bar denoted by (|) is noted against its corresponding category. 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 category repeated fourteen times will be marked as (𝍸 𝍸 IIII).

Table 7.2: Frequency Distribution of Areas of Interest of Visitors

Area of Interest	Tally Sheet	Frequency
Economics	𝍸 𝍸 𝍸	15
History	𝍸 𝍸 II	12
Political Science	𝍸 𝍸 𝍸 𝍸 II	22
Public Administration	𝍸 𝍸 𝍸 III	18
Sociology	𝍸 𝍸 𝍸 𝍸 III	23
Total		**90**

In Table 7.2 we have constructed a frequency distribution of 90 visitors to a library according to their area of interest. Here we have given the tally sheet also. But while making the final presentation we delete the second column, that is the tally sheet, and provide only the first and third columns so that the table does not look cluttered.

At times we are interested in 'relative frequency'. In this case the percentage share of

each category is given in addition to actual frequency. For example, for the data given in table 7.2 we obtain relative frequency by dividing each frequency by the total and then multiplying by 100. For economics we obtain relative frequency as $\frac{15}{90} \ 100 =$ 16.7. The total of all frequencies is 100. Relative frequency gives the percentage share of each category in the total.

Table 7.3: Relative Frequency Distribution

Area of Interest	Frequency	Relative Frequency (%)
Economics	15	16.7
History	12	13.3
Political Science	22	24.4
Public Administration	18	20.0
Sociology	23	25.6
Total	**90**	**100.0**

7.3.2 Numerical Data

In the case of numerical data we have two types: discrete and continuous. The frequency distribution of discrete data is not much different from the method discussed above for nominal and ordinal data. Here we make a list of all possible values that the characteristic is likely to assume and count the frequency of occurrence for each value. In place of categories in Table 7.2 we write down the discrete values and construct a similar table. For example, if you have to prepare a frequency distribution of 'number of books issued' to 100 members of the library you may obtain a frequency distribution as given in Table 7.3.

Table 7.4: Books Issued to Borrowers

Number of books issued	Number of borrowers
0	10
1	23
2	25
3	17
4	15
5	10
Total	**100**

In the case of continuous data preparation of a frequency distribution is somewhat different because of the following: i) Recall that an observation can assume any value (implies infinite number of values) within a range in the case of continuous data. Hence it may not be possible to list out all the possible values. ii) Repetition of the same value for two or more observations may be a rare coincidence.

In order to resolve the above issue we divide the range of values that is the difference between the highest and the lowest values, into certain 'class intervals'. In each class

interval we count the number of observations and report it. For example, let us consider the monthly expenditure on purchase of books by 175 persons. We find that the lowest monthly expenditure is Rs. 135 and the highest is Rs. 750. Thus the range is Rs. 750 - Rs. 135 = Rs. 615. We divide this range into 7 class intervals and prepare a frequency distribution as given in Table 7.5.

Table 7.5: Monthly Expenditure on Purchase of Books (in Rupees)

Class Interval	Frequency	Relative Frequency
100-200	21	12.00
200-300	32	18.29
300-400	49	28.00
400-500	33	18.86
500-600	23	13.14
600-700	12	6.86
700-800	5	2.86
Total	**175**	**100.00**

At this point two questions may be shaping up in your mind:

- How many class intervals should be taken?
- What should be the width of each class interval?

Let us discuss some of the issues that we should be careful about.

Number of class intervals: There is no hard and fast rule regarding the number of class intervals. However, it should not be too small, neither too large. If the number of class intervals is too small, then there is a chance of losing valuable information due to grouping. For example, in Table 7.5, we do not know the exact amount spent by the 21 persons whose monthly expenditure is between Rs. 100 and Rs. 200. Note that when we have lesser number of class intervals, the width of class intervals will increase. 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. Based on experience, it has been observed that the minimum number of classes should not be less than 5 and in any case, there should not be more than 20 classes. A decision on the number of class intervals should also take into account the number of observations - higher the number of observations, higher the number of class intervals.

Width of Class Intervals: As far as possible, all the classes should be of equal width. However, when a frequency distribution, based on equal class intervals, does not reveal a regular pattern of behaviour, it might become necessary to re-group the observations into class intervals of unequal width. By a regular pattern of behaviour we mean that observations should not be distributed among classes in an erratic manner. In other words, there should not be situations where frequency is zero in one class and very high in the adjoining class.

Open-ended Class Intervals: In many cases a few observations may be very high or very low in value. For example, in Table 7.5, suppose one person has a monthly expenditure of Rs. 1150 while others have less than Rs. 800. Here if we provide four extra class intervals, namely, 800-900, 900-1000, 1000-1100, and 1100-1200, then the frequency in the classes 800-900, 900-1000, 1000-1100 classes will be zero each

and in the class 1100-1200 will be one. In order to manage such cases we often resort to open-ended class intervals. In Table 7.5, instead of having class limits for the last class as 700-800 we may modify it as 'more than 700'. This class may include any observation above Rs. 700. There may be another situation where we need to modify the first class interval as 'less than 200' if one or two observations are lees than Rs. 100.

Mid-value of a Class: If we look into Table 7.5 we find that for each class, there are two class limits - lower limit and upper limit. We assume that the observations are uniformly distributed within the class. Thus we can say that the average value of the observations in a class is equal to the mid-value of the class. In Table 7.5 the mid-value of the first class (100-200) is 150 while that of the second class (200-300) is 250 and so on. Remember that class limits are usually kept as multiple of 5 or 10 so that it is convenient to locate the mid-value of a class.

You may observe that in Table 7.5 the upper limit of the second class - interval (200-300) is equal to the lower limit of the third class interval (300-400). In which class interval do you include a person having monthly expenditure of exactly Rs. 300? We should note that the second class - interval is defined as 'monthly expenditure of Rs. 200 or more but less than Rs. 300'. Similarly, the third class interval is defined as 'monthly expenditure of Rs. 300 or more but less than Rs. 400'. Naturally Rs. 300 would be included in the third class interval (300-400) and not in the second class - interval (200-300).

As in the case of nominal data, we can present the relative frequency for each class interval. It is obtained by presenting the frequencies as percentage of the total so that total frequency is 100 percent (or equal to 1).

Self-Check Exercise

1) Define the following terms.

   a) Class interval

   b) Open-ended class

   c) Frequency distribution

2) What are the factors one should keep in mind while preparing a frequency distribution for continuous data?

Note: i) Write your answers in the space given below.

ii) Check your answers with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

.....

.....

.....

Cumulative Frequency

Many times we are interested to know the number of persons below certain value. For example, for the data given in Table 7.5, we are interested to know the number of persons whose monthly expenditure on books is less than Rs. 500. We obtain it by adding the frequencies of preceding classes and find it to be 21 32 49 33 135.

Table 7.6: Cumulative Frequency (monthly expenditure in Rupees)

Class Interval	Frequency	Cumulative Frequency (less than type)
Less than 200	21	21
Less than 300	32	53
Less than 400	49	102
Less than 500	33	135
Less than 600	23	158
Less than 700	12	170
Less than 800	5	175
Total	**175**	

We can construct cumulative frequencies for the number of persons having a monthly expenditure of more than a particular value. For example, suppose we have to find out the number of persons having monthly expenditure of more than Rs. 400. We can obtain it by adding the frequencies of the succeeding classes and find it to be . Similarly, you can find out cumulative frequency for other classes.

33 23 12 5 73

7.4 GRAPHICAL PRESENTATION

Collected data are very often presented through graphs and diagrams for greater clarity.

7.4.1 Line Graph

Suppose you are provided with data on number of books issued in a library (month-wise for the year 2004) as given in Table 7.1.

Table 7.4.1: Number of Books Issued in a Library

Month	Number of visitors	Month	Number of visitors
January	76	July	105
February	85	August	108
March	86	September	110
April	90	October	115
May	82	November	118
June	98	December	106

Line graph is appropriate when we need to present the movement or variation in a variable. It is quite simple to draw and indicates the increase or decrease in a variable over time or across observations. Line graphs can be used for discrete data. Recall that in the case of continuous data we assumed that the average value of each class is its mid-point. Thus we can plot the frequencies for each class against its mid-point and join these points to obtain a line graph.

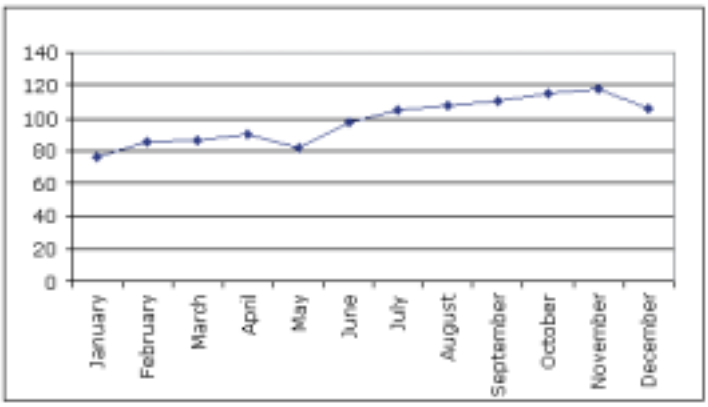


Fig. 7.1: Number of Books Issued in a Library

In Fig. 7.1 we take the variable 'months' on x-axis and the number of books issued on the y-axis and plot it as line graph. You may observe that the number of books issued has increased over time except for the months of May and December.

7.4.2 Histogram

Histogram is a rectangular diagram where the area of each rectangle is proportional to the frequency of the respective class. Remember that histogram is appropriate for continuous data arranged into class intervals. It is not used for discrete data.

The steps followed are:

- On a graph paper we mark class intervals such as 100-200, 200-300, etc. on the horizontal axis.
- Similarly we mark frequencies on the vertical axis.
- We draw rectangles as shown in Fig. 7.2.
- When class intervals are equal the height of rectangles are equal to the frequency of classes.
- When class intervals are not equal the frequencies are adjusted so that area of rectangle is proportional to class frequency. For example, if the interval of one class is double that of other classes, then we need to divide the frequency of the former by two.

Let us construct histogram for the data given in Table 7.5.

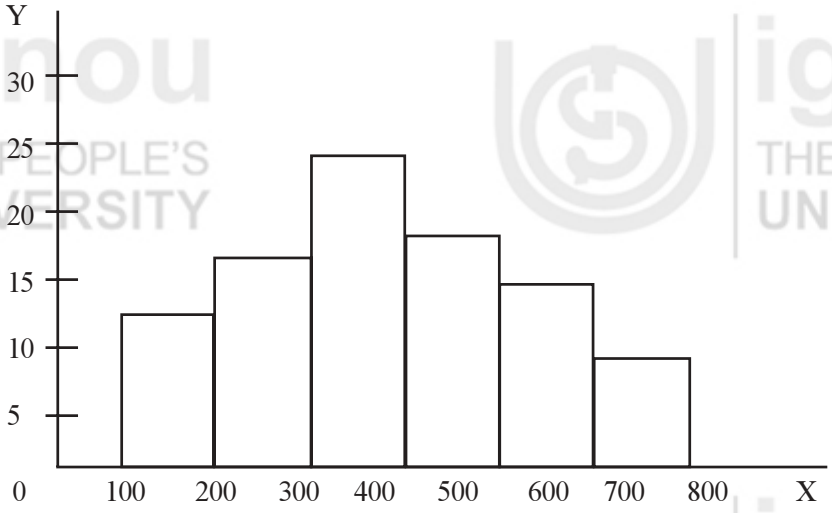


Fig. 7.2: Histogram

7.4.3 Frequency Polygon

It is obtained from a histogram by joining the mid-value of the top of the rectangles with the help of straight lines as shown in Fig. 7.3. Remember that the area under the frequency polygon should be same as the area under the histogram. Hence, we draw two additional class intervals, one on each end of the histogram. For the histogram given in Fig. 7.2 we follow the steps given below.

- Draw two class intervals, viz., 0-100, and 800-900.
- Take the frequency for these two classes to be 0.
- Join the mid-values of all the classes, including 0-100, and 800-900.
- The frequency polygon obtained is as given in Fig. 7.3.

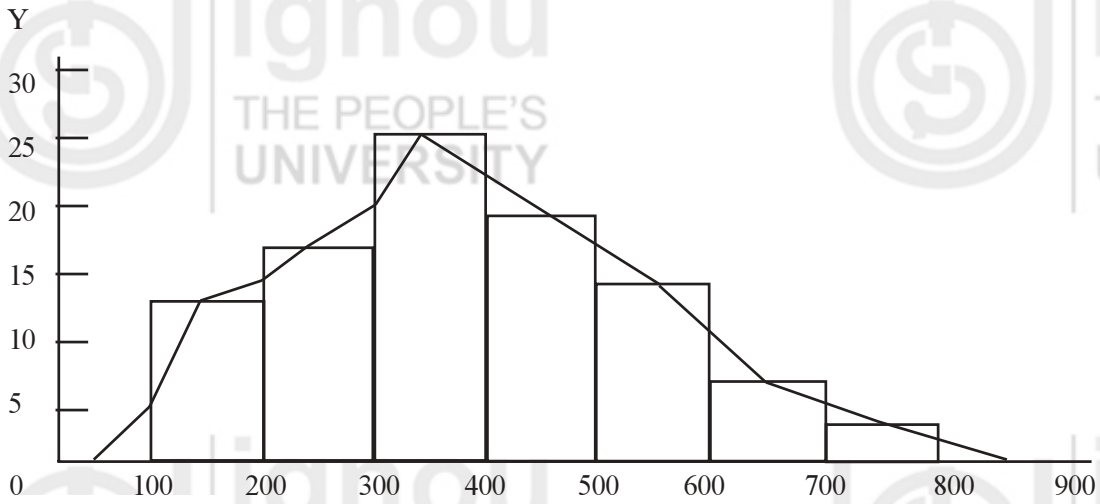


Fig. 7.3: Frequency Polygon

y Curve

frequency polygon by smoothening the edges of the frequency
n Fig. 7.4.

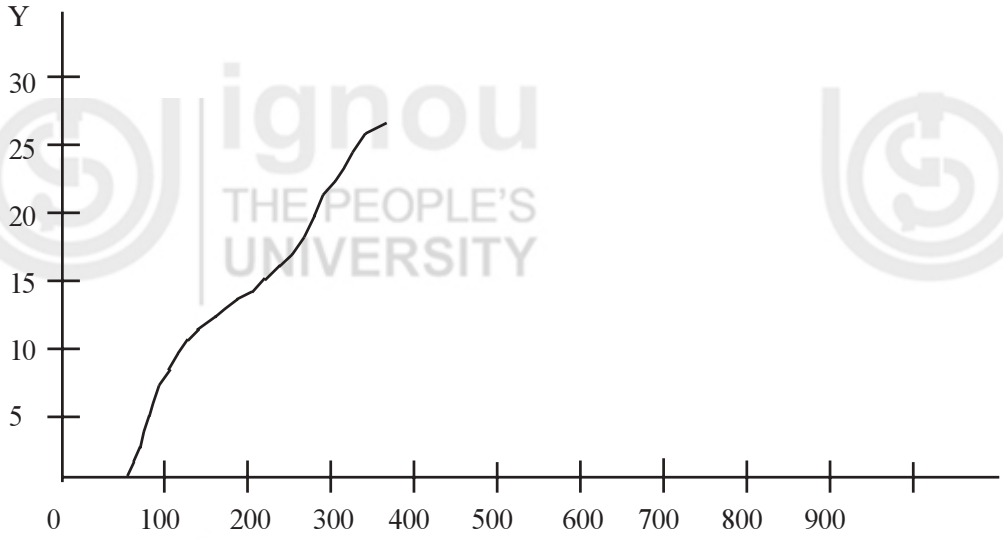


Fig. 7.4: Frequency Curve

Self Check Exercise

3) Construct a more than type cumulative frequency distribution for the data given in Table 7.5.

Note: i) Write your answer in the space given below.

ii) Check your answer with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

7.4.5 Bar Diagrams

Diagram is a visual form of presentation data. Diagrams can be bars, squares, circles, maps and pictures. In this section we discuss the procedure of drawing bar diagrams. Bar diagrams are more appropriate for nominal data where certain categories are distinguished.

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. Remember that width of the bars does not mean anything. Moreover, bars are separated from each other with equal gaps. Thus it is different from histogram, which is more appropriate for quantitative data and area of the bars is important. 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 7.4.2: Number of Students in a College

Stream	Number of students
Arts	800
Science	460
Commerce	520

The bar diagram of the above data is drawn in Fig. 7.5. To make the bar diagrams beautiful we can either fill in colour in the bars or shade them in different ways. This is left to the aesthetic taste of the investigator.

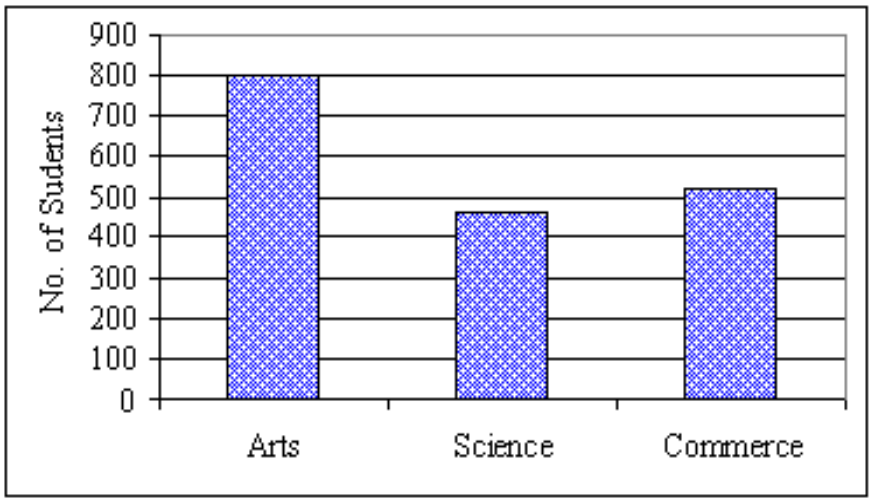


Fig. 7.5: Simple Bars

Component Bar Diagram

A component bar diagram presents the components of a phenomenon so that a comparison can be made. It is more appropriate for qualitative data. The bars corresponding to each category is divided into various components. The portion of the bar occupied by each component denotes its share in the total.

Suppose we have additional information on the number of boys and girls admitted in a college (Table 7.9). We present component bar diagram for such data in Fig. 7.6. Remember that 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.

Table 7.4.3: Number of Students in a College

	Boys	Girls
Arts	400	400
Science	260	200
Commerce	350	170

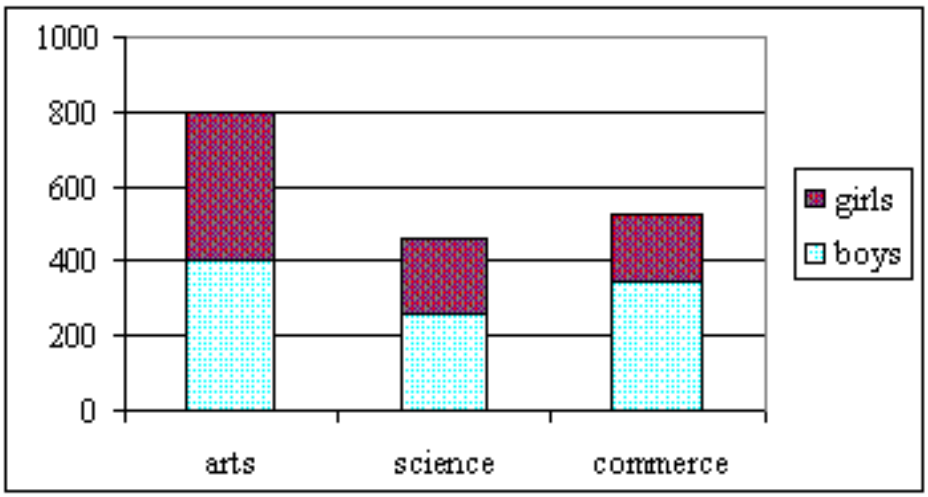


Fig. 7.6: Component Bar Diagram

Multiple Bar Diagram

It is used when comparisons are to be shown between two or more sets of data. A set of bars for a period or place or a related phenomenon is drawn side by side without gap. Different bars are distinguished from one another by different shades or colours. A multiple bar diagram for the hypothetical data given in Table 7.10 is drawn in Fig. 7.7. Suppose our purpose is to show the changes in the number of girl students admitted to different of a college over the years.

Table 7.4.4: Number of Girls Students during past three years

Stream/Year	2001	2002	2003
Arts	130	135	125
Science	45	57	65
Commerce	20	26	30

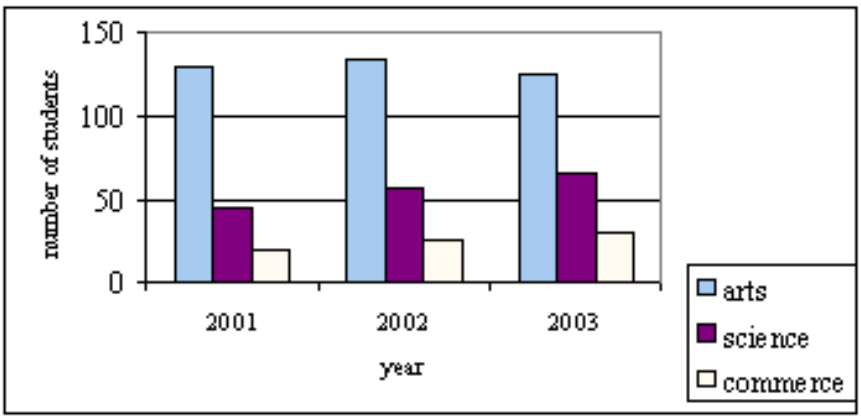


Fig. 7.7: Multiple Bar Diagram

7.4.6 Pie Chart

Pie chart is widely used to show share of different components in a variable. For example, expenditure of a library on different heads can be shown in the form a pie chart. Suppose for the financial year 2004-05 you have budget data of a library as given in Table 7.11. Recall that a circle has 360^0. This area is divided into different components according respective share. Therefore, we first calculate the share of each component and convert it to degrees.

Table 7.4.5: Heads of Expenditure in a Library

(in Rs. Thousand)

Heads of Expenditure	Budget	Ratio of the component	Degrees
Salary	246	0.24	86.3^0
Purchase of books	325	0.32	114.0^0
Purchase of journals	175	0.17	61.4^0
Purchase of furniture	200	0.19	70.2^0
Maintenance	50	0.05	17.5^0
Contingency	30	0.03	10.5^0
Total	**1026**	**1.00**	**360^0**

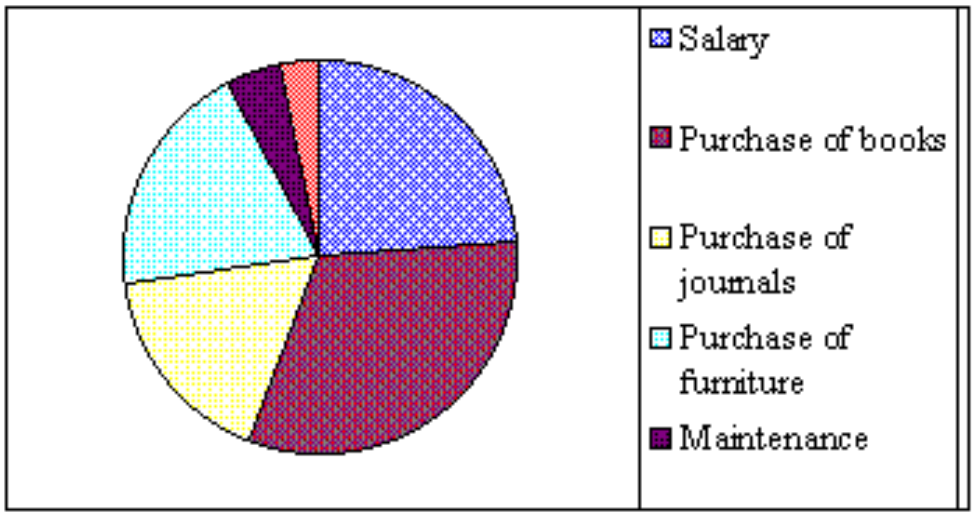


Fig. 7.8: Pie Chart

Steps to be followed in the construction of Pie diagram:

- Find the total of all components (in this case =1026).
- Find ratio of each component to total. For example, for 'salary' the ratio is $\frac{246}{1026}$ 0.24.
- Multiply the ratio of each component by 360^0. For example, for 'salary' it is $0.24 \quad 360^0 \quad 86.3^0$.
- Draw a circle of a suitable size.
- Use protractor to draw the angles you have obtained. Preferably start with the largest one.
- Shade the different segments with different colours.
- Identify the colour or shade for different components on the right hand side.
- At times we write down the share of each component inside the circle.

Self Check Exercise

4) Draw a simple bar diagram for the data given in Table 7.3.

5) For the data given in Table 7.3 draw a pie chart.

Note: i) Write your answers in the space given below.
ii) Check your answers with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

.....

7.5 SUMMARY

In this unit we discussed the method of presentation of data. We began with preparation of frequency distribution for qualitative and quantitative data. Also we prepared relative frequencies and cumulative frequencies for numerical data.

Data can be presented in the form of tables and graphs. There are certain parts of a table that we should mention while preparing a table. Graphs can be in the form of histogram, frequency polygon and frequency curve. Diagrams could be in the form simple bar, component bar, multiple bar or pie chart.

7.6 ANSWERS TO SELF CHECK EXERCISES

1) a) The difference between the upper limit and the lower limit of a class is called the class interval.

b) A class with no upper limit is called an open- ended class.

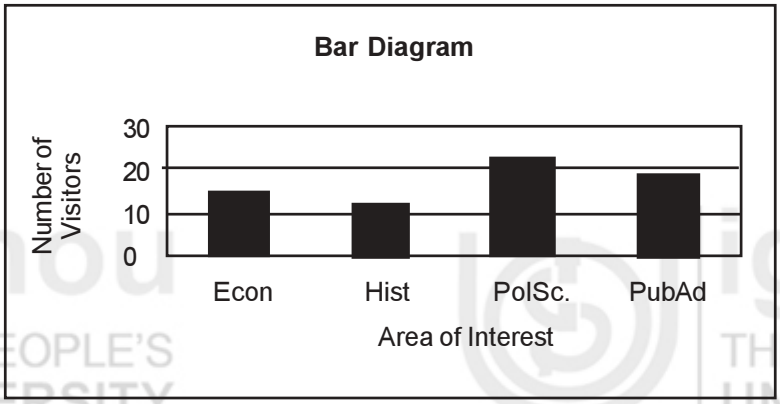
c) A tabular presentation of nominal or ordinal data as classes along with the frequency of their occurrence is called frequency distribution.

2) While preparing frequency distribution we should be keep in mind: number of class intervals, their width and whether they have to be closed or open.

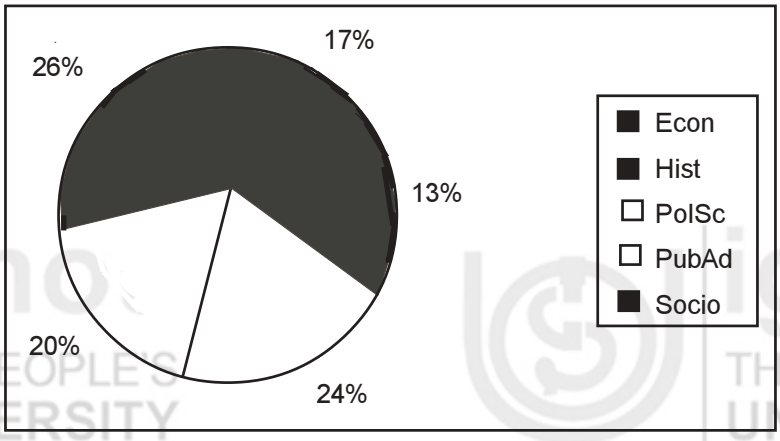
3) The 'more than' type cumulative frequency is given below.

Class Interval	Frequency	Cumulative Frequency
More than 100	21	175
More than 200	32	154
More than 300	49	122
More than 400	33	73
More than 500	23	40
More than 600	12	17
More than 700	5	5
Total	175	

4)



5)



7.7 KEYWORDS

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.

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 is the difference of two limits of a class. It is equal to upper limit minus lower limit. It is also called class width.

Continuous Frequency Distribution : A continuous frequency distribution in formed where the variable can take any value between two numbers like height and weight, income, temperatures, etc.

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.

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 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.

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

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.

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

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

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

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

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

7.8 REFERENCES AND FURTHER READING

Chase, Clinton I. (1988). *Elementary Statistical Procedures*. 3rd ed. New York: McGraw Hill.

Elhance, D.N. (1988). *Fundamentals of Statistics*. Allahabad: Kitab Mahal.

Gupta, C.B. (1985). *An Introduction to Statistical Methods*. Delhi: Vikas.

Hraglin, David C. [et.al.] (1985). *Exploring Data Tables, Trends and Shapes*. New York: Wiley.

Simpson, I.S. (1988). *Basic Statistics for Librarians*. London: Library Association.

UNIT 8 STATISTICAL TECHNIQUES

Structure

8.0 OBJECTIVES

After going through this Unit you should be in a position to:

- explain various measures of central tendency such as arithmetic mean, median and mode;
- explain various measures of dispersion such as range, variance, standard deviation, and coefficient of variation;
- explain correlation and regression techniques; and
- analyse time series data.

8.1 INTRODUCTION

In the previous Unit we explained the methods of presenting data in the form of tables and graphs. However, many times we need a single summary value that would describe a series. For example, we have data on the number of visitors to a library on a daily basis. Such data can be presented in the form of a table or in the form of a line graph. But if we want a single summary figure, arithmetic mean would give us the average number of visitors to the library.

Statistical techniques are more suitable for quantitative data although certain techniques do exist for qualitative data also. Recall that quantitative data can be of two types: discrete and continuous. We will explain various techniques for both types of variables. We begin with measures of central tendency.

8.2 MEASURES OF CENTRAL TENDENCY

Measures of central tendency provide us with a summary that describes some central or middle point of the data. There are five important measures of central tendency, viz., i) arithmetic mean, ii) median, iii) mode, iv) geometric mean, and v) harmonic mean. Out of these the last two measures, viz., geometric mean and harmonic mean, have very specific uses and thus less frequently used. Therefore, we will discuss the first three measures in this Unit.

Before dealing with these measures let us be familiar with certain notations that we will use. The standard notation is: X is a variable that takes values $X_1, X_2, X_3 \cdots X_N$. Let us consider the data given in Unit 7, Table 7.3 on the number of books issued to borrowers. You know that it is a discrete variable and the number of books that can be issued to each borrower varies between 0 and 5, viz., 0, 1, 2, 3, 4, and 5. The corresponding frequency for each observation is: 10, 23, 25, 17, 15 and 10.

In the above example, we denote the variable 'number of books issued' as X and the values assumed by it as $X_1, X_2, X_3, X_4, X_5, X_6$. The corresponding frequencies are $f_1, f_2, \cdots, f_6$. We call a typical observation as the i^{th} observation and denote it as X_i with frequency f_i. In the example on the number of books issued i ranges between 0 and 5.

In the case of continuous variable we take the mid-values of class intervals as $X_1, X_2, X_3 \cdots X_n$ and the corresponding frequencies as $f_1, f_2, \cdots, f_N$.

8.2.1 Arithmetic Mean

Arithmetic mean is also called 'mean' or 'average'. It is denoted by a bar above the variable being averaged. It is defined as the sum of all observations divided by the number of observations.

Let us calculate arithmetic mean for observations arranged in a frequency distribution. If $X_1, X_2, X_3 \cdots X_N$ are the observations and the corresponding frequencies are $f_1, f_2, \cdots, f_N$ then arithmetic mean is given by $\overline{X}$ (Read it as 'X-bar') and defined as

$$\overline{X} \quad \frac{1}{N} f_1X_1 \quad f_2X_2 \quad \cdots \quad f_nX_n \ .$$

It can be abbreviated as $\bar{X} = \frac{1}{N} \sum f_i X_i$ ...(8.1)

where N is the total number of observations and is equal to $\sum f_i$. The symbol $\sum$ (read it as 'sigma') denotes the sum of a variable.

When observations are classified into class intervals, in the case of continuous data, individual observations within a class interval are not separately identifiable. To avoid this difficulty, it is assumed that every observation falling into a class interval has a value equal to the *mid-value* of the class interval.

Example 8.1

Calculate the average number of books issued to borrowers on the basis of the following data.

Number of books issued	Number of borrowers
0	10
1	23
2	25
3	17
4	15
5	10
Total	100

We prepare a table for calculating arithmetic mean.

Number of books issued (X_i)	Number of borrowers (f_i)	$f_i X_i$
0	10	0
1	23	23
2	25	50
3	17	51
4	15	60
5	10	50
Total	$\sum f_i = 100$	$\sum f_i X_i = 234$

Secondly, we fill in the values from the table in the formula (8.1) that is $\bar{X} \quad \frac{1}{N} \quad \sum f_i X_i$.

Here $N = 100$, and $\sum f_i X_i = 234$.

Therefore, $\bar{X} \quad \frac{1}{100} \quad 234 \quad 2.34$

Thus, the average number of books issued to borrowers is 2.34.

Example 8.2

Given below is the monthly expenditure in Rupees on purchase of books by 100 persons. What is the monthly average expenditure?

Class Interval	Frequency
100-200	12
200-300	18
300-400	28
400-500	19
500-600	13
600-700	7
700-800	3
Total	100

In the case of continuous data we have to find out the mid-value of the class intervals and then apply formula (8.1).

Class Interval	Mid-value	Frequency	$f_i X_i$
100-200	150	12	1800
200-300	250	18	4500
300-400	350	28	9800
400-500	450	19	8550
500-600	550	13	7150
600-700	650	7	4550
700-800	750	3	2250
Total		$f_i = 100$	$f_i X_i = 38600$

By applying the relevant values from the above table in (8.1), that is,

$$\bar{X} \quad \frac{1}{N} \quad f_i X_i$$

Here $N = 100$, and $f_i X_i = 38600$.

Therefore, $\bar{X} \quad \frac{1}{100} \quad 38600 \quad 386$

Thus the average monthly expenditure on purchase of books by the groups of individuals is Rs. 386.

8.2.2 Median

Median gives us with the middle-most observation in a series so that half of the observations remain on each side of the median. For example, if you have 5 observations, viz., 2, 5, 9, 14 and 20, then 9 is the middle observation and 2 observations remain on both sides of it. Thus median of the above series is 9. Let us consider

another series where there are 6 observations: 3, 8, 15, 25, 35, and 43. In this case the median could be any number between 15 and 25 because 2 observations will remain on both sides. Conventionally we take the average of the middle-most two numbers. Here it would be $\frac{15 \quad 25}{2}$ 20. Thus in this case the median is 20.

However, when the number of observation is too large or data is arranged in a frequency distribution, it is not that simple to locate the median. If there are N observations the median observation should correspond to the th observation. We first find out the cumulative frequency of the distribution (see Unit 7) and secondly find out the class interval in which the $\frac{N}{2}$ th observation lies. This class interval is our 'median class'. Thirdly, we apply the following formula to get the median value.

$$M_d \quad l_m \quad \frac{\frac{N}{2} \quad C}{f_m} \quad h, \qquad \ldots(8.2)$$

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,

f_m is the frequency of median class, and

h is the width of median class.

$\frac{N}{2}$

Example 8.3

For the data given in Example 8.2 above, find out the median monthly expenditure on purchase of books.

To solve the above problem we go by the following steps:

1) calculate the cumulative frequency distribution

2) find out the median class

3) apply formula (8.2)

Class Interval	Frequency	Cumulative Frequency
100-200	12	12
200-300	18	30
300-400	28	58
400-500	19	77
500-600	13	90
600-700	7	97
700-800	3	100
Total	100	

There are 100 observation. So the median value corresponds to 50th observation, which lies in the class interval 300-400. Therefore, median will remain somewhere between 300-400. Thus the median class is 300-400. In this case

l_m =300, C = 30, N = 100, f_m = 28, h = 100,

By applying (8.2) we obtain the median value as

$$M_d \quad 300 \quad \frac{100/2 \quad 30}{28} \quad 100 = \text{Rs. } 371.43$$

8.2.3 Mode

Mode is the observation with the highest frequency. For discrete data it is easier to find out the mode. But in the case of continuous data we have to identify the 'modal class', that is the class interval having highest frequency. We have to see that the width of the classes is the same. Otherwise, large class intervals are likely to include large number of observations and smaller class intervals are likely to have few observations. Mode is computed by the following formula:

$$M_o \quad l_m \quad \frac{1}{1 \quad 2} \quad h, \qquad \ldots(8.3)$$

where

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

$_1 \quad f_m \quad f_{m\ 1}$ is the difference of the frequencies of the modal class and its preceding class,

$_2 \quad f_m \quad f_{m\ 1}$ is the difference of the frequencies of the modal class and its following class, and

h is the width of modal class.

Example 8.4

Calculate the mode for the data set given in Example 8.2. The steps involved are

1) Identify the modal class. This is the class interval with highest frequency. In this case the modal class is 300-400.

2) Calculate l_m, $_1 \quad f_m \quad f_{m\ 1}$, $_2 \quad f_m \quad f_{m\ 1}$, and h

3) Apply formula (8.3)

We find that mode is $M_o \quad l_m \quad \frac{1}{1 \quad 2} \quad h \quad 300$.

$$= 352.63$$

Note that mean, median and mode assume different values for the same data. In the case of data relating to monthly expenditure on purchase of books given in Example 8.2, we find that mean, median and mode are Rs. 386, Rs. 371.43 and Rs. 352.63 respectively.

8.3 MEASURES OF DISPERSION

Measures of central tendency provide us with a summary figure for the data set. However, in many situations these measures do not represent the distribution of data. For example, look into the following three sets of data:

1) Set A: 2, 5, 17, 17, and 44.

2) Set B: 17, 17, 17, 17, and 17.

3) Set C: 13, 14, 17, 17, and 24.

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

The word dispersion gives 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, only 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 should capture such variability in data.

There are quite a few measures of dispersion. We will discuss range, mean deviation, variance and standard deviation in this Section.

8.3.1 Range

Range 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. In the case of grouped data individual observations are not identifiable. In such cases we take the difference between two extreme boundaries of the classes.

8.3.2 Variance

Variance is the most widely used measure of dispersion. It is denoted by the symbol ^{2} (read as 'sigma-squared') and is defined as

$$\text{Variance} = {}^2 \quad \frac{1}{N} \quad (X_i \quad \bar{X})^2 \qquad \ldots(8.4)$$

In the case of frequency distribution variance is given by

$$^2 \quad \frac{1}{N} \quad f_i(X_i \quad \bar{X})^2 \qquad \ldots(8.5)$$

where $N \quad \sum_{i \ 1}^{n} f_i$, the total number of observations.

In order to simplify calculation we use the following formula

$$^2 \quad \frac{1}{N} \quad f_i X_i^2 \quad \bar{X}^2 \qquad \ldots(8.6)$$

Remember that we obtain the same value whether we apply (8.5) or (8.6).

8.3.3 Standard Deviation

Standard deviation is another widely used measure of dispersion. It is defined as the positive square root of variance and denoted by . Remember that standard deviation cannot be negative.

Example 8.5

Given below is the range of marks obtained by students in a class. Find out the standard deviation.

Marks	Number of students
15-25	8
25-35	12
35-45	20
45-55	10
55-65	6
65-75	4
Total	60

In order to calculate the standard deviation we go by the following steps:

1) calculate the mid-values of the classes
2) calculate arithmetic mean
3) Apply formula (8.5) or (8.6) to find out variance
4) Prepare a table with the required columns. We have prepared one for applying (8.5).
5) Find out variance
6) Find out the positive square root of variance

Marks	Number of students f_i	Mid-value X_i	X_i	$(X_i \quad \overline{X})$	$(X_i \quad \overline{X})^2$	$f_i(X_i \quad \overline{X})^2$
1	2	3	4	5	6	7
15-25	8	20	160	-21	441	3528
25-35	12	30	360	-11	121	1452
35-45	20	40	800	-1	1	20
45-55	10	50	500	9	81	810
55-65	6	60	360	19	361	2166
65-75	4	70	280	29	841	3364
Total	60		2460			11340

In the above table we first find out the arithmetic mean, which is 41. Next we find out variance as

$$^2 \quad \frac{1}{N} \quad f_i(X_i \quad \overline{X})^2 = \frac{1}{60} \quad 11340 \quad 189$$

Hence standard deviation is

$$\sqrt{189} = 13.75 \text{ marks}$$

Remember that standard deviation is always expressed in the unit of measurement. It is not a pure number. The difficulty with variance is that it is expressed in the square of the unit of measurement, which does not have any meaning.

8.3.4 Coefficient of Variation

Many times we have to compare the variability among different series of data. If the units are measured in different units we obtain different values for standard deviation. For example, let us compare the economic status of households in two villages. The summary figures of monthly calorie intake of households are given below for the two villages.

	Village A	**Village B**
Number of households	817	561
Mean calorie intake	2417	2235
S. D. of calorie intake	418	232

The problem is to identify the village that has more inequality as far as calorie intake is concerned. We find that village A has higher mean calorie intake but has larger standard deviation and larger number of households compared to village B. Thus village A may actually have more number of poorer households than in village B. In order to compare such situations we use the coefficient of variation (c.v.). It is defined as percentage standard deviation per unit of mean, i.e.,

$$\text{c.v.} = \frac{}{\overline{X}} \quad 100 \qquad \ldots(8.7)$$

Since and have same unit of measurement, c.v. is a pure number and it is not affected by the choice of unit of measurement.

For village A, c.v. $= \frac{418}{2417} \quad 100 = 17.29$ and for village B, c.v. $= \frac{232}{2235} \quad 100 = 10.38$.

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.

8.4 CORRELATION

So far we have dealt with a single characteristic of data. But, there may be cases when we would be interested in analysing more than one characteristic at a time. For example, you may like to study the relationship between the age and the number of books a person reads. Such data, having two characteristics under study are called bivariate data. One of the measures to find out the extent or degree of relationship between two variables is correlation coefficient.

An analysis of the co-variation of two or more variables is usually called correlation. If two characteristics vary in such a way that movement in one is accompanied by movement in the other, these characteristics are correlated. For example, there is relationship between price and supply, income and expenditure, etc. With the help of correlation analysis we can measure in one figure the degree of relationship existing between two variables.

8.4.1 Scatter Diagram

If we are interested in finding out the relationship between two variables, the simplest way to visualise it is to prepare a dot chart called scatter diagram. Using this method, the given data are plotted on a graph paper in the form of dots. For example, for each pair of X and Y values, we put a dot and thus obtain as many point as the number of observations. Now, by looking into the scatter of various dots, we can ascertain whether the variables are related or not. The greater the scatter of the plotted points on the chart, the lesser is the relationship between the two variables. The more closely the points come to a straight line, the higher the degree of relationship. Here are some illustrations of some correlations between two variables.

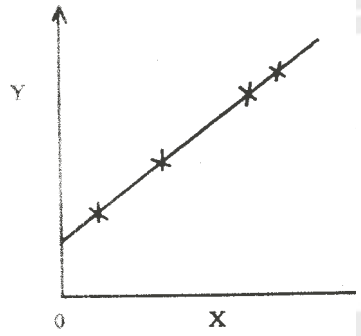


Fig. 8.1: Perfect Positive Correlation r = 1

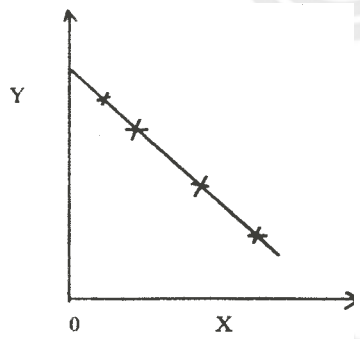


Fig. 8.2:Perfect Negative Correlation r = -1

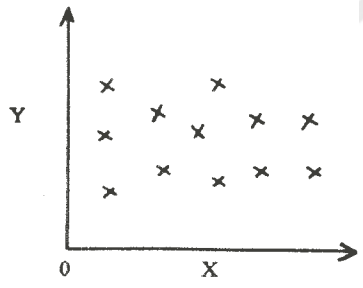


Fig. 8.3: No Correlation r = 0

You see that the scatter diagram gives a rough idea of the degree of relationship between two variables.

As we are considering a relationship between the two variables here, there might be a relationship between more than two variables. In case of 10 variables, we can plot the points on a two-dimensional graph paper, i.e., on a space with x and y axes. But the scatter diagram has the limitation that it cannot be plotted where more than two variables are involved. Secondly, it does not give an exact figure on the degree of relationship between variables.

To reach an exact figure on the extent of relationship between variables and also to overcome the limitation of considering more than two variables at a time, we calculate the correlation coefficient.

Fig. 8.4: High Degree of Positive Correlation

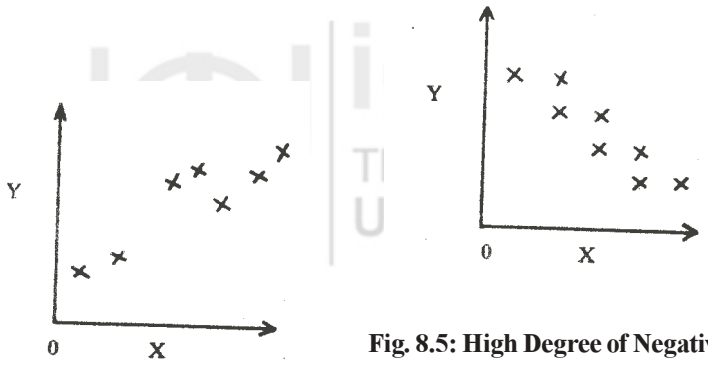


Fig. 8.5: High Degree of Negative Correlation

8.4.2 Pearson's Product Moment Correlation

Of the several methods of measuring correlation, Pearson's Product Moment correlation is mostly used in practice. It is denoted by the symbol r.

The formula is

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}}{\sqrt{\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]\left[\sum_{i=1}^{n}(Y_i\right.}}$$

Another formula for correlation coefficient is:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sigma_X \sigma_Y}$$

Where $_x$ is the standard deviation of X and $_y$ is the standard deviation of Y.

When we get value of r = +1, it means there is perfect positive correlation between variables. When r = -1, there is perfect negative correlation and when r = 0, it means that there is no correlation between the two variables. Correlation coefficient can take any value between +1 and -1, i.e., it cannot exceed +1 and cannot be less than -1. Usually, in real life analysis, we get values, which lie between +1 and -1 such as +0.6, -0.5 etc. The above formula of r can be transformed into the following form, which is easier to apply.

$$r = \frac{\sum_{i\ 1}^{n} y_i}{\sqrt{\sum_{i\ 1}^{n} y_i^2}}$$

Where $x_i \quad (X_i \quad \bar{X})$ and $y_i \quad (Y_i \quad \bar{Y})$

Steps necessary for calculating Coefficient of Correlation are:

i) Take the deviation of X from the mean X and denote this by

ii) Square this deviation and obtain the total, i.e.

iii) Take the deviation of Y from the mean Y and denote this by n

iv) Square this deviation and obtain

v) Multiply x and y obtain the total i.e.

vi) Substitute the values of $x_i y_i$ $x_i^{\ 2}$ and in the above formula.

Remember that correlation indicates the degree of association between variables X and Y. It merely shows whether the variation in one variable is accompanied by the variation in the other or not. Correlation coefficient does not indicate a cause and effect relationship between the variables. We cannot say X causes Y or vice versa. Second thing to remember is that a r = +0.6 does not imply greater relationship than r = 0.6 The positive or negative sign indicates the direction of relationship. If r = +0.6, then it is very likely than an increase in X is accompanied by an increase in Y. On the other hand, r = -0.6 indicates the inverse relationship between X and Y. Thirdly, correlation coefficient is neutral to change in scale and origin. This means that in illustration -8, for example, if we divide the X variable by any figure, say, 5 and variable Y by 2 (or any other figure), we get the same value for r-value. This is called a change in scale. For instance, when you convert the height in cm. To that in inches, the correlation coefficient does not change. Similarly, if you change the origin, i.e., deduct or add certain figure to all the observations, the correlation coefficient does not change. You may find this out yourself by deducting 10 from X and 5 from Y and then calculating the correlation coefficient. Fourthly, correlation coefficient indicates the linear relationship between the variables. For higher order relationships, correlation coefficient does not reflect the proper degree of relationship. For example, for two variables, X and Y, if X = Y2 for every value of Y then r may turn out to be zero. But this does not mean that X and Y are not related. So that we can say is that when two variables are independent r = 0, but from the mere knowledge of r = 0 we cannot infer that the two variables are independent.

Let us take an illustration and find out the value of r.

Illustration 11

The following table shows the data on height and weight of 10 children. Find out the product moment correlation coefficient.

Height (in cm)	Weight (in Kg)	Height (in cm)	Weight (in Kg)
110	26	140	38
110	21	135	30
125	22	130	30
130	24	140	40
145	36	135	43

Let the height (in cm.) be termed X and weight is Y

	Y_i	x_i $(X_i \quad \bar{X})$	y_i $(Y_i \quad \bar{Y})$	$x_i y_i$	x_i^2	y_i^2
110	26	-20	-5	100	400	25
110	21	-20	-10	210	400	100
125	22	-5	-9	45	25	81
130	24	0	-7	0	0	49
145	36	15	5	75	225	25
140	38	1	7	70	100	49
135	30	5	-1	-5	25	1
130	30	0	-1	0	0	1
140	40	10	9	90	100	81
135	43	5	12	60	25	144
1300	310				645	1300

X_i

8.4.3 Spearman's Rank Correlation

In the previous Unit, while mentioning the types of data, we had distinguished between ratio scale and ordinal scale of measurement. Pearson's product moment correlation coefficient can be applied only when the data are measured in the ratio scale. But when, instead of actual magnitude of the observations, we have only the ranks, the Pearson's 'r' becomes inapplicable. In such cases the Spearman's rank correlation (r_s) is used. The method of calculating r, is quite simple. The first step is to identify the rank or the i^{th} observation in X and Y in their respective series of n items. The second step is to find out the difference between the rank in X and the rank in Y. Let this difference be termed D1. The third step is to apply the following formula.

$$r_s \quad 1 \quad \frac{6 \quad \overset{n}{} \quad D_i^2}{\underset{i \ 1}{n(n^2 \quad 1)}}$$

The Spearman's rank correlation also ranges from +1 to –1. Thus, positive values indicate direct relationship between the variables, while negative values indicate inverse relationship. The value r = 0 indicates absence of association between the variables.

One note of caution is that, Spearman's rank correlation should not be used just because it is easier to compute than Pearson's product moment correlation coefficient..r, is less interpretive than r. We cannot strictly say that the change in X is associated with proportionate change in Y because equal differences in ranks do not imply equal differences in the characteristics.

Illustration 12

The following data give rank of 12 journals by two different methods used to compute rank correlation coefficient.

Sl. No.	Rank according to the No. of publication X	Rank according to the No. of citation Y	D_i	D_i^2
1	1	12	-11	121
2	2	9	-7	49
3	3	6	-3	9
4	4	10	-6	36
5	5	3	+2	4
6	6	5	+1	1
7	7	4	+3	9
8	8	7	+1	1
9	9	8	+1	1
10	10	2	+8	64
11	11	11	0	0
12	12	1	+11	121
				D^2 416

Here

Therefore, $r_{s=1}$ - 1.454 = 0.454

Self Check Exercise

1) Find out the standard deviation of the following data.

Marks in statistics	No. of students
0-20	5
20-40	10
40-60	15
60-80	8
80-100	2

2) The following table shows the marks obtained by 10 students in statistics and mathematics. Find out the correlation coefficient.

Sl. No. of Students	Marks in Statistics	Marks in Mathematics
1	52	48
2	60	40
3	45	38
4	38	28
5	35	42
6	62	65
7	68	53
8	28	25
9	50	28
10	62	33

3) Write the formulae for the following concepts:

a) Standard Deviation

b) Variance

c) Product Moment Correlation Coefficient

d) Rank Correlation

Note: i) Write your answers in the space given below.

ii) Check your answers with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

.....

.....

8.5 REGRESSION ANALYSIS

In the previous Unit it has been noted that a correlation coefficient does not reflect cause and effect relationship. The regression analysis, to be discussed in this Unit, seeks to dwell on such a theme. It assumes that one variable is the cause and other(s) the effect. In general terms, we can say, variables are of two types: independent variables and dependent variables. Independent variable is the cause and dependent variable is the effect.

Regression analysis 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.

Let us assume that the number of books in circulation, in a library is related to the number of users. For example, it can be postulated that as the number of users increases, the number of books in circulation also increases. Here, the number of users is the independent variable and the number of books in circulation is the dependent variable. Let us denote the dependent variable as Y and the independent variable as X. In regression analysis, we gather data over a period of time or across units at a point of time. Let us assume that 'n' pairs of observations in X and Y are collected. The next step is to find out the relationship 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 equations. 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. 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 equational form?" There is no hard and fast rule as such. The form of equation depends upon the reasoning and assumptions made by the researcher.

However, the researcher 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 equational from. If the points are more or less in a straight line then linear equation is assumed. 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.

The researcher has to make another assumption: viz. identification of independent and dependent variables. The again depends upon the logic of the researcher and purpose of analysis: whether Y depends upon X or X depends upon Y. Thus, there may be two regression lines from the same data (a) when Y is assumed to be dependent upon X, this is termed 'Y on X' line, and (b) when X is assumed to be dependent upon Y, this is termed 'X on Y' line.

Let us take an example of a linear equation with Y as the dependent variable and X as the independent variable.

$$Y=3+2X$$

By taking in different values of X, we can determine the values of Y, e.g., when X=1, Y=5; when X=2, Y=7 and so on. If we plot these pairs of points (1,5) (2,7), etc. on a graph paper we get a straight line.

Generalising the above relationship, it can be said that a linear equation of Y on X take3s the form Y=a+bX, where a and b are constants. Similarly, non-linear equations can be specified in many forms. A simple example is Y=a+bX+cX2

8.5.1 Linear Regression

Let us consider the following data. The number of visitors and the number of books issued during weekdays in a week are given.

No. of visitors to a library (X)	6	2	10	4	8
No. of books issued (Y)	8	4	10	7	8

If we plot the data on a graph paper, the scatter diagram looks something like Fig. 8.6.

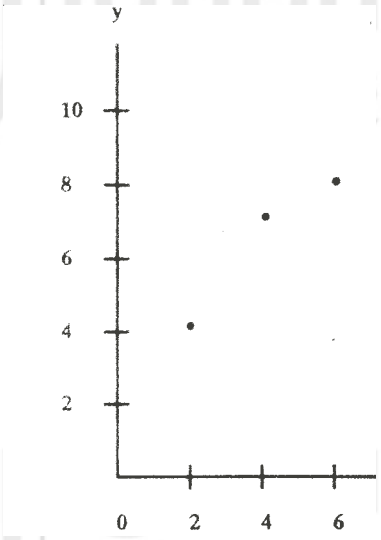


Fig. 8.6: Scatter Diagram for No. of visitors to a library and Books issued

As is obvious from the graph, the points do not strictly lie in a straight line. But they show an upward rising tendency where a straight line can be fitted.

If we plot the straight line along with the scattered points, the diagram looks like Figure 8.7. The difference between the regression line and the observations is the 'error'. For example, against a X value of 2, the Y value is 4. This is called the observed value.

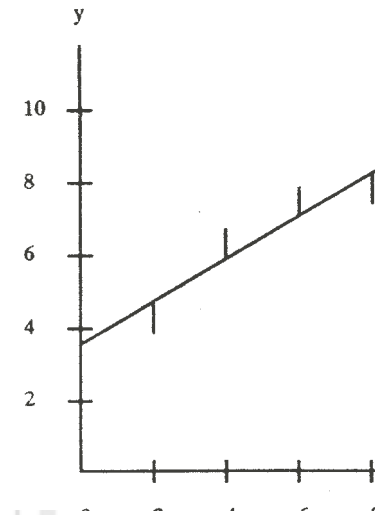


Fig. 8.7

But the regression line shows Y value of 4.8 against X value of 2. This value, which is calculated from the regression line, is the expected value. The difference between the observed value and the *expected value* is termed as the error value. So we see that observed value is the sum of expected value and error and value.

Our objective in fitting a regression line is to minimise the error values. This is usually done by the method of 'least squares'. The method of least squares minimises the value of E, where e is the difference between observed value and expected value. We will not go into the details of the method here. Instead, two equations derived on the basis of least squares method and known as normal equations are given.

These are:

$$Y \quad na \quad b \quad X$$

As a rule of thumb we can say that these normal equations 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 = a + bX. The first normal equation is simply the linear equation Y = a +bX summed over all observations.

SY = a +SbX or SY = na + bSX

The second normal equation is the linear equation multiplied by X and summed over all observations.

$SXY = SaX + SbX^2$ or $SXY = aSX + bSX^2$

It is evident that all the terms in these equations are given numbers, calculated from the data, except a and b.

The values of a and b need to be calculated for getting the estimated value of the dependent variable. This is done in the following.

It can be seen from Table 8.1 that data on X and Y variables are given in the first two columns. The succeeding two columns in the table give the calculations necessary to solve two normal equations given above. The expected value of Y and error value are given in the last two columns.

Table 8.1: Computations of Data for Regression Analysis

(1) X	(2) Y	(3) X2	(4) XY	(5) Expected value of Y	(6) Error
6	8	36	48	7.4	0.6
2	4	4	8	4.8	-0.8
10	10	100	100	10.0	0.0
4	7	16	28	6.1	0.9
8	8	64	64	8.7	-0.7
TOTAL 30	37	220	248	37	0.0

As the normal equations are:

Y = na + b X ……(8,8)

$XY = a \quad X + b \quad X^2$ ……(8.9)

We substitute the respective values from the Table 1.

Thus,

37 = 5a + 30b ……(8.10)

248 = 30a + 220b ……..(8.11)

If we multiply equation (3) by 6 and subtract the product from equation (4) we get :

248	=	30a	+	220b
222	=	30a	+	180b
-		-		-
		40b	=	26
or		b	=	0.65

On substituting the value of b in equation (8.10) we get:

37 = 5a + 30 x 0.65

or 5a = 17.5 or a = 3.5

So the regression line is

$$Y = 3.5 + 0.65 X \quad \ldots\ldots (8.12)$$

If we substitute the values of X in the regression line that is equation (8.12), we get the expected values of Y. For example, when X = 2

$$Y = 3.5 + 0.65 \times 2 = 4.8$$

But our observed value of Y against X = 2 is 4. This difference between the observed value and the expected value (4 - 4.8 = -0.8) is the error 'e'.

The expected values of Y and e are given in the Table 8.1 above in columns (5) and (6).

Notice that the sum of errors for the sample is zero, i.e. Se = 0

For computational purposes we may use the following formulae to find out the value of a and b.

Let us take

$$X = (X - X)$$

$$Y = (Y - Y)$$

$$XY = (X - X)(Y - Y)$$

Where $\bar{X}$ and $\bar{Y}$ are the arithmetic means of X and Y variables respectively. This formula gives

$$b \quad \frac{xy}{x^2} \quad \ldots\ldots (8.13)$$

$$a \quad \bar{Y} \quad b\bar{X} \quad \ldots\ldots (8.14)$$

Since these formulae are derived from the normal equations, we get the same values for 'a' and 'b' in this method also.

The steps in computation are:

1) Find out the values of X and Y

2) Find out x = (X - X) and y = (Y - Y)

3) Find out the values of xy

4) Find out the values of x^2

5) Apply the formulae in equations (8.13) and (8.14) above.

On applying the above formulae in the example given in Table 1, above we get the data given in Table 2.

Table 8.2: Computation of Regression Equation: Short cut Method

X	Y	X= (X - X)	Y = (Y - Y)	xy	X^2
6	8	0	0.6	0	0
2	4	-4	-3.4	13.6	16
10	10	4	2.6	10.4	16
4	7	-2	-0.4	0.8	4
8	8	2	0.6	1.2	4
Total 30	37			26.0	40

$\bar{X} \quad \frac{1}{n} \quad X \quad 30/5 \quad 6 \qquad b \quad \frac{xy}{x^2} \quad 26/40 \quad 0.65$

$\bar{Y} \quad \frac{1}{n} \quad Y \quad 37/5 \quad 7.4 \qquad a \quad \bar{Y} \quad b\bar{X} \quad 3.5$

Needless to mention that values for a and b are same as derived earler.

8.5.2 Properties of Regression Coefficient

Coefficient 'b' is called the regression coefficient. Notice that we can draw two regression lines from the data on X and Y.

(a) Y on X line, $Y = a + bX$

(b) X on Y line, $X = \quad + \quad Y$

The two coefficients, b and ß, demonstrate some interesting properties. First, the product of both regression coefficients is equal to the square of r (correlation coefficient), i.e., $b\ ß = r^2$

So once we know both regression coefficients we can find out the value of r^2. By taking the square root of r^2 we get r. Second if the regression coefficients are negative in sign, then the correlation coefficients also is negative. If the regression coefficients are positive then correlation is positive. Third, you know that

, i.e., r lies between - 1 and +1

Therefore r^2 lies between zero and +1. Regression coefficient can take finite value. But is one regression coefficient is more than 1 the other regression coefficient is less than 1. Both regression coefficients cannot exceed unity. Also it follows that the product of both, which is r^2, cannot exceed unity. The square of correlation coefficient is called the coefficient of determination and implies important characteristics. If r^2, the coefficient of determination, is closer to one we can infer that the independent variable explains the movements in the dependent variable. If the coefficient of determination is closer to zero, the independent variable does not explain the variation in the dependent variable.

8.5.3 Non-linear Regression

In the previous sub-section we discussed the simple linear regression involving two variables: one dependent and the other independent. Regression can involve one dependent variable and more than one independent variable. Such cases are called multiple regressions.

The equation fitted in regression can be non-linear or curvilinear. 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 :

$$Y = na + b\ X + c\ X^2$$

$$XY = a\ X + b\ X^2 + c\ X^3$$

$$X^2 Y = a\ X^2 + b\ X^3 + c\ X^4$$

Notice again that the normal equations are the regression equation multiplied by the coefficients of a,b and c and summed over all observations.

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 sections. We give below some of the frequently used non-linear equations and the respective transformed linear equations.

1) $Y = a\ e^{bx}$

$\sqrt{\ }$ $\frac{1}{a\ \ bX}$ By taking natural log (that is, ln), it can be written as

$\ln Y = \ln a + b X$

Or $Y = a + ß X$

Where, $Y = \ln Y$, $a = \text{in } a$, $X = X$ and $ß = b$

2) $Y = a X^b$

By taking log, the equation can be transformed into

$\log Y = \log a + b \log X$

Or $Y = a + ß X$

Where, $Y = \log Y$, $= \log a$, $ß = b$ and $X = \log X$

3)

If we take $Y' \quad \frac{1}{Y}$ then

$Y' \quad a \quad bX$

4) $Y = a + b \quad X$

If we take $X = \quad X$ then

$Y = a + b X$

Once the non-linear equation is transformed, the fitting of a regression line is as per the method discussed in the beginning of this section. 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

8.5.4 Prediction

A major interest in studying regression lies in its ability to forecast. In the illustration at the beginning we assumed that the number of books issued depend upon the number of visitors. We fitted a linear equation to the observed data and got the relationship

$$Y = 3.5 + 0.65 X$$

From this equation we can forecast the number of books issued given the number of visitors. For example, if the number of visitors goes up to 30, then the number of books issued will be

$$Y = 3.5 + 0.65 \times 30 = 23$$

The procedure is to substitute the X value in the regression equation and get the expected Y value.

The question that arises here is: Will the predicted value come true? It depends upon the coefficient of determination. If the coefficient of determination is closer to one, there is greater likelihood that the prediction will be realized. However, the predicted value is constrained by elements of randomness involved with human behaviour and other unforeseen factors.

Self Check Exercise

4) Given below is the data on X and Y

X :	15	17	20	22	25	33
Y :	25	22	30	31	37	35

i) Find out the regression line Y on X.

ii) Find out the regression line X on Y.

iii) Find out the coefficient of determination.

iv) Find out the Pearson's product moment correlation.

Note: i) Write your answers in the space given below.

ii) Check your answers with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

.....

.....

8.6 TIME SERIES ANALYSIS

In regression analysis, we discussed the cause and effect relationship between the dependent and the independent variables. The independent variable can be any characteristic based on our reasoning. When the independent variable is 'time', we call it 'time series'. We assume that the dependent variable depends upon time or varies according to time.

In our day-to-day life we encounter several instances where certain characteristics vary according to time. The stock in a library, the expenditure on user facilities, etc., are a few examples.

One of the important tasks before librarians and information managers is to make estimates for the future. For example, a publisher may want to know his probable sales for the next year, so that he can properly plan and take steps to avoid the possibility of unsold stocks or lack of supply of some books published by him. A librarian may wish to study the trend of book issued in order to take appropriate measures while making his future plans. For all these purposes, one needs to consult the data, which have been collected and recorded at successive intervals of time. Such statistical data is referred to as 'time series' data.

An example of time series data could be the number of books on Library and Information Science issued during 1995 to 2004. These data may be recorded as follows:

Table 8.3: Hypothetical Data Recorded on Books Issued Over Time

Year	No.of Books Issued
1995	356
1996	350
1997	391
1998	289
1999	408
2000	412
2001	405
2002	482
2003	497
2004	469

A close look at the data would show that the demand of the books has increased with some fluctuations. There may be several reasons for increase or decrease during a certain period.

8.6.1 Components of Time Series

The fluctuations in time series may be classified into the following types of variations.

a) Secular trend

b) Seasonal variation

c) Cyclical variation

d) Irregular variation

A time series may have the above components in combination as well.

a) **Secular Trend**

Changes which take place as a result of general tendency of the data to increase or decrease are known as secular movement. The general movement persisting over a long period of time is called secular trend. The above time series example for data on books issued is an example of secular trend.

b) **Seasonal Variation**

Changes or variations, which seasonally occur within a period of one year as a result of changes in climate, weather, important happenings etc., are called seasonal variations. It may be possible that in the data on books issued in the above example (If made available on a month to month basis), the year starts with a low figure in the beginning, reaches its peak in the middle and decreases at the end of the year. This type of Fluctuation within a span of a year is called seasonal variation.

c) **Cyclical Variation**

Changes that take place due to cyclic fluctuations like prosperity and depression may be termed as cyclical variations. In every business cycle there are four periods, (i) prosperity (ii) decline, (iii) depression and (iv) improvement. Cyclical variations are of a longer duration than a year.

d) **Irregular Variation**

Changes, which take place due to the factors that, could not be predicted like violent riots, natural calamities, etc., come under irregular variation.

The components of time series data, viz., seasonal, trend, cyclical and irregular, can be separated. In a traditional time series analysis, it is assumed that there is a multiplicative relationship among these four components. This may be represented symbolically as follows:

$$Y = T \times S \times C \times I$$

Where T = Secular Trend

S = Seasonal Variation

C = Cyclical Variation

I = Irregular Variation

This is called the multiplicative model. In another approach, it is assumed that

$$Y = T + S + C + I$$

A model formulation of this category is called 'additive model'.

8.6.2 Measurement of Secular Trends

There are various methods for determining secular trends.

The most frequently used methods are:

1. Moving average method, and
2. Method of least squares.

a) **Method of Moving Average**

It is a method of smoothing out fluctuations by calculating a series of averages by allowing overlapping periods of the time series. Before the moving average is calculated, it is necessary to select a proper period of moving average like three yearly, five-yearly, etc.

If the period chosen is m years, the moving averages are obtained by calculating a series of mean values of m consecutive values covering overlapping periods of the series. The mean of the first m values, given by $(Y_1 + Y_2 + \ldots\ldots + Y_m)$ is placed at the mid-point of the period covering the first m years.

It would be the first moving average value. The second moving average value will be obtained by calculating mean of values covering the period 2nd to (m+1)th years. For example: $\frac{1}{m}(Y_1 + Y_2 + \ldots\ldots + Y_{m\ 1})$ will be the next moving average and so on. This process is repeated till the last observation is covered.

The formula of the three-yearly moving average will be:

1/3 $(Y_1 + Y2 + Y_3)$, 1/3 $(Y_2 + Y_3 + Y_4)$, 1/3 $(Y_3 + Y_4 + Y_5)$ and so on.

And that of five-yearly moving average will be

1/5 $(Y_1 + Y_2 + Y_3 + Y_5)$, 1/5 $(Y2 + Y_3 + Y_4 + Y_5 + Y_6)$ and so on.

$\frac{1}{m}$

The above procedure will be easy to understand from the following illustration.

Illustration

Calculate a three-yearly moving average from the following sales figures of a publisher.

Table 8.4: Hypothetical Data for Computation of Moving Average

Year	Sales (hundred units)
1990	5
1991	7
1992	9
1993	12
1994	11
1995	10
1996	8
1997	12
1998	13
1999	17
2000	19
2001	14
2002	13
2003	12
2004	15

Data on sales of books for 15 years are given above. To calculate three-yearly moving average, we take, to begin with, the first three years total. This is 5+7+9=21 and place this value against the middle observation, i.e., against 1971. In column (4) of Table 8.5 the moving average, i.e., the three-yearly total divided by the period (which is 3 in this case) is given. Thus 21/3=7 for the first entry.

Following a similar procedure other calculations are made and results are presented in the table given below.

Table 8.5: Computation of Three-Yearly Moving Average

Year	Sales	3-yearly Totals	3-yearly Moving Average
(1)	(2)	(3)	(4)
1990	5		
1991	7	21	7.0
1992	9	28	9.33
1993	12	32	10.67
1994	11	33	11.00
1995	10	29	9.67
1996	8	30	10.00
1997	12	33	11.00
1998	13	42	14.00
1999	17	49	16.33
2000	19	50	16.67
2001	14	46	15.33
2002	13	39	13.00
2003	12	40	13.33
2004	15		

From the illustration above it may be noted that we do not get moving average value for the beginning and the end years. In the case of a five-yearly moving average, we lose moving averages for the beginning two years and the two years at the end. This loss of information increases as the 'period' of moving average increases. Secondly, from the moving averages, we cannot predict the figures for the future. It is just an analysis of past behaviour.

b) **Method of Least Squares**

We used this method earlier to obtain the regression lines. The procedure here is very similar to the fitting of regression lines. Here the independent variable is 'time' t. The first step is to form the equation of 'secular trend'. As you know, both straight lines and curves can be fitted by the least squares method. If Y is the dependent variable, the straight line to be fitted is $Y = a + bt$. Again, we have to minimize the error between the observed and the expected values. The method of least squares suggests that the sum

of squares of the error terms should be the minimum. From this method, the relationship between dependent and independent variables is estimated from the normal equations. For a linear equation Y = a +bt, the normal equations are:

$Y = na + b \quad t$

$tY = a \quad t + b \quad t^2$

The constants 'a' and 'b' are determined from these two equations, and 'n' indicates the number of observations in the sample.

We measure the variable 't' by taking the mid-point of time as the origin. Suppose n =5 years. Then taking the origin at the third year of the time, we get t = -2, -1, 0, 1, 2. It may be seen that t = 0 for the third year.

In case the number of years covered is even, say 6 years, the origin at the mid point of the two middle years, i.e., at 6 months past the third year is considered. In such cases t takes values

t = -5, -3, -1, +1, +3, +5

To fit a linear trend line by making use of the data given in Table 8.5 above, the necessary computations are summarised in Table 8.6 below.

Table 8. 6: Data Computation of Time Trend

Year	t	Sales	tY	t2
1990	-7	5	-35	49
1991	-6	7	-42	36
1992	-5	9	-45	25
1993	-4	12	-48	16
1994	-3	11	-33	9
1995	-2	10	-20	4
1996	-1	8	-8	1
1997	0	12	0	0
1998	1	13	13	1
1999	2	17	34	4
2000	3	19	57	9
2001	4	14	56	16
2002	5	13	65	25
2003	6	12	72	36
2004	7	15	105	
Total		0	177	171 280

Recall that the normal equations are:

$Y = na + b \quad t$

$tY = a \quad t + \quad t^2$

Substituting the respective values from the values from the table, we get,

177 = 15 a + b 0

171 = a 0 + b 280

or

15a =177 or, a = 11.8

280b =171 or, b = 0.61

So the trend line is Y = 1.8 +0.61 t

Remember that 't' is the codified time value with 197 as the origin.

The method of least squares enables us to forecast future values for Y. This is done by substituting the 't' value in the equation.

In the illustration given above, the sales books (in hundred units) in the year 2006 will be 17.29.

Y = 11.8 +0.61 9 = 17.29

In place of 't' 9 is substituted since starting with 1997, the year of origin, 2006will be 9 years. The predicted sales of books (in hundred units) in the year 2006 will be 17.29.

As you know, non-linear trends can also be fitted to the observed data. Hence, predictions can also be made on them. In the analysis of trend line, we make the implicit assumption that the past behaviour continues to persist in a future period also. So, a change in the past behaviour would make prediction unreliable.

8.6.3 Measurement of Other Components

Seasonal variation and cyclical fluctuation are periodic and recurring movements in the data. It has been stated above that the seasonal variation is short term in nature and usually the periodicity is less than a year. In contrast to this, the cyclical fluctuation lasts longer than a year.

Just as there are several methods of measuring the seasonal variations viz., ratio to trend method, ratio to moving averages method and link relative method, the cyclical fluctuations can be measured by harmonic analysis, spectrum analysis, etc. These methods involve tedious calculations and hence are not discussed here. If you are interested in these methods, you may look into the books referred to at the end of this Unit. However, attempts to separate the time series of its four components - seasonal, trend, cyclical and erratic, may follow some simple procedure. This is briefly spelt out in the following:

a) The seasonal component described above can be estimated with the help of moving average. This component can be eliminated from the original observations through subtraction if we assume an additive model..

b) The trend of the seasonally adjusted data are then estimated by means of least square straight line or some other function fitted by least squares described above. The trend component can be eliminated from the seasonally adjusted data.

c) The residuals, which remain after the elimination of seasonal and trend components from the original time series can be recorded and potted graphically. This residual variation may be compared visually or through some other method. The remaining variations of the data series are attributed to cyclical and erratic components.

8.7 SUMMARY

In this Unit we discussed various statistical techniques for analyzing data. Often it is necessary to provide a summary figure for a set or series of data. Such figure could be a measure of central tendency such as mean, median and mode, or it could be a measure of dispersion such as variance, standard deviation and coefficient of variation.

There are cases where more than one characteristic of a sampling unit is measured. In such type of data, we can find out the correlation coefficient or we can fit a regression equation. Remember that correlation does not show a cause and effect relationship between variables. It only shows the strength of relationship. In regression analysis variables are divided into two categories: independent and dependent. Regression equation can be a straight line or a curve depending upon the type of equation fitted.

Often we have data at certain intervals for a sufficiently long period of time. Such data re called time series and contains certain components, viz., secular trend, cyclical variation, seasonal variation and irregular movements.

8.8 ANSWERS TO SELF CHECK EXERCISES

1) The standard deviation comes out to be 21.07. Apply the formula given in the text and check the answer.

2) The correlation coefficient r = +0.61.

3) a) It is defined as the positive square root of variance and denoted by .

$$= +$$

^{2}

b) Variance is the most widely used measure of dispersion. It is denoted by the symbol ^{2} (read as 'sigma-squared') and is defined as

Variance = $^2 \quad \frac{1}{N} \quad (X_i \quad \bar{X})^2 \quad \frac{1}{N} \quad X_i \quad \bar{X}^2$

In the case of frequency distribution variance is given by

$$^2 \quad \frac{1}{N} \quad f_i(X_i \quad \bar{X})^2$$

where $N \quad \sum_{i \ 1}^{n} f_i$, the total number of observations.

In order to simplify calculation we use the following formula

$$^2 \quad \frac{1}{N} \quad f_i X_i^2 \quad \bar{X}^2$$

c) $$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sigma_X \sigma_Y}$$

d) $r_s \quad 1 \quad \frac{6 \sum_{i \ 1}^{n} D_i^2}{n(n^2 \quad 1)}$

4) i) You have to find out the normal equations and substitute the values in the equations. The estimated regression line will be Y = 13.94 + 0.73 X

ii) The regression line for X on Y will be X = a + bY

Consequently the normal equations will be

$\Sigma X = n\alpha + \beta \Sigma Y$

$\Sigma XY = \alpha \Sigma Y + \beta \Sigma Y^2$

The estimated regression line will be

X = -5.6 + 0.92 Y

iii) The coefficient of determination is the product of regression coefficients of both the regression lines. So, it is 0.73 x 0.92 = 0.067

iv) Correlation coefficient is the square root of coefficient of determination, i.e., r = Ö 0.067 = 0.82. Since the regression coefficient is positive in sign, the correlation coefficient is also positive.

8.9 KEYWORDS

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.

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.

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 a pure number.

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.

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.

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 minimum.

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.

8.10 REFERENCES AND FURTHER READING

Sanders, D. H. (1980). *Statistics: A Fresh Approach.* New Delhi: McGraw Hill.

Rao, I. K. R. (1983). *Quantitative Methods for Library and Information Science.* New Delhi: Wiley Eastern.

UNIT 9 STATISTICAL PACKAGES

Structure

9.0 OBJECTIVES

After reading this Unit, you would be able to:

- define the word "statistics";
- describe the different data types and data formats;
- illustrate the role of statistical methods;
- describe the features of statistical packages like SPSS and MS-Excel; and
- portray some of the popular statistical packages.

9.1 INTRODUCTION

One of the features of the development of modern world is the development of the capacity to convert observations in numbers. The science, which deals with numbers, is statistics. It crunches the numbers and organises them in a meaningful way so that information is generated. This information builds up knowledge and thus the development goes on. Advances in computing have come handy in this as they help in doing this part of job accurately, timely effectively and convincingly.

Computer can help immensely in the statistical analysis. There exist numerous statistical tools and the need is to identify their actual usage. Even with the use of a statistical package many statistical procedures require a lot of prior knowledge and insight.

In this Unit, we have tried to build-up a case for the usage of statistics by first defining statistics and what it can do to your data. You will further find the definition of the data types. An explanation of the common tasks that are performed in a preliminary analysis is also given. This Unit also presents a description of two popular packages (MS Excel and SPSS) and gives a glimpse of some other statistical packages.

9.2 STATISTICAL PACKAGES

Before going on to the definition of statistical packages, one needs to revisit the definition of statistics and its functions. In this section, we would highlight the areas/ problems that statistics as a discipline addresses to and the kind of data one gets for the statistical applications.

9.2.1 Definition

The term "Statistics" is used as a "collection of numerical facts or data". It is also used in terms of a "body of methods and techniques for analyzing numerical data". Statistical techniques have many purposes, which include methods and procedures for summarising, simplifying, reducing and presenting raw data. It then makes predictions, tests hypotheses and infers characteristics of a population from the characteristics of a sample. In other words, Statistics is generally thought of as serving two functions. One is to describe sets of data; the other is to help in drawing inferences. When you are studying only a sample, there is possibility that your assumption may not be accurate and you can never be certain that you have drawn the correct inference. For this reason the inferential use of statistics may be thought of as helping you to make decisions under conditions of uncertainty. It is different from guessing, because Statistics also provides you with a method of estimating how reliable your conclusions are. With each statistical statement that you make, you indicate the probability that findings like yours could have been the result of chance factors.

A statistical package is the software for the collection, organisation, interpretation, and presentation of numerical information. The need for a statistical package has arisen because of the complexity of calculations involved in making inferences from the data. The advances in computing technologies have made statistics a yet more powerful field.

According to Ripley (2004), "The most widely used piece of statistical packages/ software for statistics is Excel. SPSS and SAS dominate certain communities, and Minitab is widely used in teaching. Many niche products, e.g. GenStat, Generalised Linear Interactive Modelling Package (GLIM), Stata and S-PLUS dominate the high-end, hence is widely seen in methodology papers."

9.2.2 Data Measurement

Statistical data is generally obtained in many formats such as spreadsheets (e.g. MS Excel) or databases (e.g. MS Access). Data may also be received in various open formats such as typically tab-delimited text (*.dat, *.tab, *.txt), comma-separated text or fixed-width text data (*.dat, *.txt).

The data could be of two types, qualitative and quantitative. Most of the statistical methods are based on the quantitative data. Quantitative variable is a variable whose values are numbers with real numeric meaning. It consists of mainly two types of data viz. discrete and continuous. A set of data is said to be discrete if the values/observations belonging to it are distinct and separate, i.e. they can be counted e.g. number of books in a library. Whereas, a set of data is said to be continuous if the values/observations belonging to it may take on any value within a finite or infinite interval.

There are four well-known levels of measurement scales i.e. nominal, ordinal, interval, and ratio. There is a relationship between the level of measurement and the appropriateness of various statistical procedures. For example, it would be impractical to compute the mean of nominal measurements. Data must be measured on an interval or a ratio scale for the computation of means and other statistics to be valid. Therefore,

if data are measured on an ordinal scale, the median but not the mean can serve as a measure of central tendency. Let us have a brief discussion of what these scales are:

1) **Nominal Scale**

Nominal measurement consists of assigning items to groups or categories. No quantitative information is conveyed and no ordering of the items is implied. Nominal scales are therefore qualitative rather than quantitative. Religious preference, race, and sex are all examples of nominal scales. Frequency distributions are usually used to analyse data measured on a nominal scale. The main statistic computed is the mode. Variables measured on a nominal scale are often referred to as categorical or qualitative variables. Nominal variables allow for only qualitative classification. That is, they can be measured only in terms of whether the individual items belong to some distinctively different categories, but we cannot quantify or even rank order those categories

2) **Ordinal Scale**

Measurements with ordinal scales are ordered in the sense that higher numbers represent higher values. However, the intervals between the numbers are not necessarily equal. Ordinal variables allow us to rank order the items we measure in terms of which has less and which has more of the quality represented by the variable, but still they do not allow us to say "how much more."

3) **Interval Scale (Cardinal Scale)**

On interval measurement scales, one unit on the scale represents the same magnitude on the trait or characteristic being measured across the whole range of the scale. Interval scales do not have a "true" zero point, however, and therefore it is not possible to make statements about how many times higher one score is than another. True interval measurement is somewhere between rare and nonexistent in the behavioral sciences. A good example of an interval scale is the Fahrenheit scale for temperature. Interval variables allow us not only to rank order the items that are measured, but also to quantify and compare the sizes of differences between them.

4) **Ratio Scale**

Ratio scales are like interval scales except they have true zero points. Ratio variables are very similar to interval variables. In addition to all the properties of interval variables, they feature an identifiable absolute zero point, thus they allow for statements such as x is two times more than y. A typical example of ratio scales is measure of time or space. Interval scales do not have the ratio property. Most statistical data analysis procedures do not distinguish between the interval and ratio properties of the measurement scales.

9.2.3 Functions of Statistics

Many people intend to use statistical techniques in their research. It is definitely a good practice to substantiate your claims with the help of data. Statistics has various functions, which can be broadly categorised as follows:

1) *Summarise and Describe data:* One summarises and describes the data in order to view data at a glance. If it is nominal or ordinal data, one makes cross-tabulations and graphs; if it is interval or ratio data then z-scores are calculated.

2) *Variance and distribution of the data:* In order to measure the spread of the data and knowing its distributions one makes tables and charts and graphs for nominal/ordinal data and histograms with normal curve or box plots with inter-quartile range for interval/ratio data.

3) *Compare groups:* When one has to compare two or more populations then one makes cross-tabulations for nominal/ordinal data and employ testing of hypothesis for continuous/numeric data divided into groups.

4) *Identify relationships:* In order to identify relationships in the data, one uses cross-tabulations for nominal/ordinal data; calculate correlation coefficient and scatter plot for Interval/ratio data or go for linear regression/ ANOVA for data with one dependent and 2 or more predictor variables.

5) *Identify groups of similar cases:* Carrying out hierarchical cluster analysis solves the problem of identifying groups of similar cases or k-means cluster analysis. One uses Discriminant analysis for identify characteristics of known groups.

6) *Identify groups of similar variables:* Factor analysis is carried out to identify groups of similar variables.

Self Check Exercise

1) Define a Statistical package. Describe its need and purpose.

Note: i) Write your answer in the space given below.

ii) Check your answer with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

9.3 FEATURES OF SOME STATISTICAL PACKAGES

Advances in computing especially the advent of the personal computer (PC) have made computing a game of the commoners. Today one has the computing power as one can easily load software of his choice or need into his PC. There is a plethora of read-made computer packages available today. Now one can find different statistical packages for applications to different disciplines. We will describe two such packages that are ready available and are popular and user friendly. We will also give a glimpse of some other packages in the subsequent section.

9.3.1 Microsoft Excel

Microsoft Excel is a big worksheet (it can take data rows in thousands across 256 columns). This worksheet can be used for data entry and for performing calculations by click of buttons. It has a "paste function where you can paste any formula from a big list of inbuilt functions. MS Excel can be used to create tables, and graphs and perform statistical calculations. The work done in MS Excel can be easily copied and pasted to many window-based programs for further analysis.

According to Pottel, "Spreadsheets are a useful and popular tool for processing and presenting data. In fact, Microsoft Excel spreadsheets have become somewhat of a standard for data storage, at least for smaller data sets. The fact that the program is often being packaged with new computers, which increases its easy availability, naturally encourages its use for statistical analysis. However, many statisticians find this unfortunate, since Excel is clearly not a statistical package. There is no doubt about that, and Excel has never claimed to be one. But one should face the facts that due to its easy availability many people, including professional statisticians, use Excel, even on a daily basis, for quick and easy statistical calculations. Therefore, it is important to know the flaws in

Excel, which, unfortunately, still exist today! …. "Excel is clearly not an adequate statistics package because many statistical methods are simply not available. This lack of functionality makes it difficult to use it for more than computing summary statistics and simple linear regression and hypothesis testing". However in MS Excel 2003 aspects of the some statistical functions, including rounding results, and precision have been enhanced.

The MS Excel worksheet is a collection of cells. As we have earlier said, there are 65,000 (rows) X 256 (columns) cells in an MS Excel worksheet. Each row or column can be used to enter data belonging to one category. Data entry in MS Excel is as simple as writing on a piece of paper. MS Excel assigns each column a field depending upon the type of data. It supports various data formats; one can choose a data format by formatting the cells.

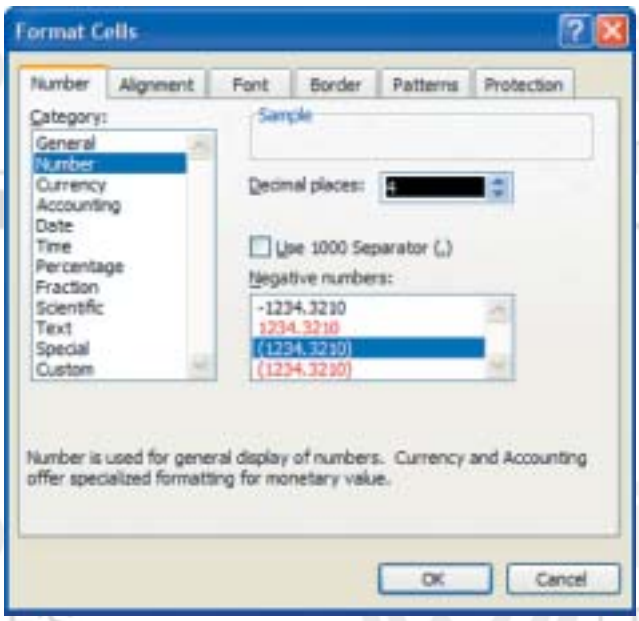


Fig. 9.1: Screen Snapshot of MS-Excel for Choosing Data Format

Once the type of cells is defined it is easy to enter the data without taking care of the format. MS Excel can perform usual calculations on the data so entered. It has an insert function (f_x) icon that contains many inbuilt functions like sum, count, max/min, standard deviation etc. In fact it has a plethora of built-in functions that performs special calculations without even typing the formula. To perform a calculation one has to select a function and specify the range of values on which it has to be applied. These functions are known as paste functions.

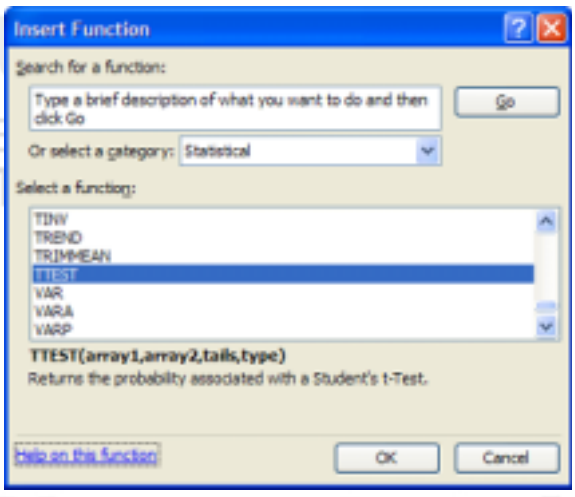


Fig. 9.2: Screen Snapshot of Function Menu Options in MS-Excel

We will concentrate on the statistical functions and see some of the major statistical functions of MS Excel. As you can see in figure 9.2, once you go to the function menu and choose "statistical" category, you will be asked to select a function. Suppose you have chosen t-test. You will be told on the same screen that t-test returns the probability associated with a student's t-test. Now if you are still not comfortable with the description, you may select help on this function, which is at the bottom left of the screen. More help is offered in the following form.

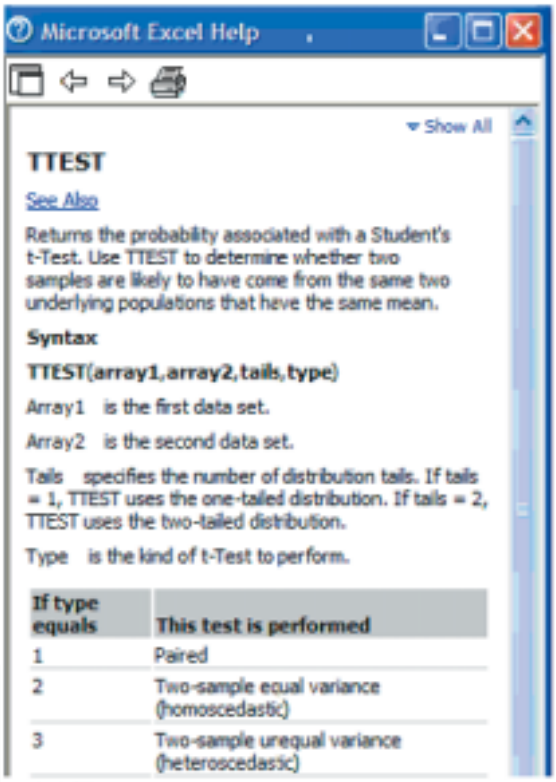


Fig. 9.3: Screen Snapshot of Help for TTest in MS-Excel

MS Excel has a built-in statistical package for taking you in further details of data analysis. It provides a set of data analysis tools called the Analysis ToolPak, which you can use to save steps when you develop complex statistical analyses. You provide the data and parameters for each analysis; the tool uses the appropriate statistical macro functions and then displays the results in an output table. Some tools generate charts in addition to output tables.

To access these tools, click Data Analysis on the Tools menu.

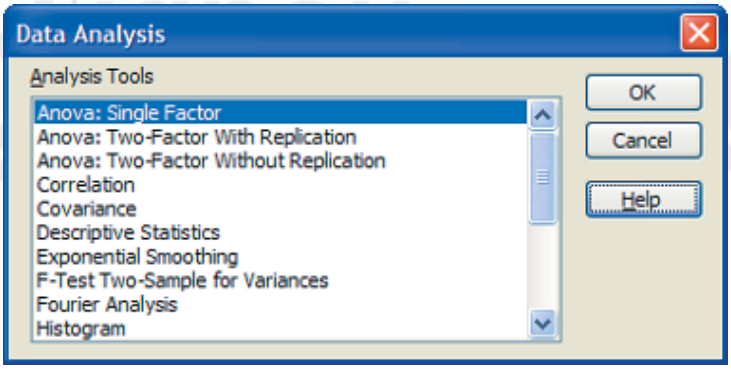


Fig. 9.4: Screen Snapshot of Data Analysis Function in MS-Excel

Let us have a brief description of these tools. The table given below highlights their functions and uses.

Table 9.1: Data Analysis Tools in MS-Excel

Sl. No.	Tool	Function	Use
1	ANOVA	The ANOVA tools provide different types of variance analysis.	Test of the hypothesis that each sample is drawn from the same underlying probability distribution,
2	Correlation Covariance	Calculate the correlation coefficient/ Covariance between two variables when measurements on each variable are observed for each of N subjects.	Examine each pair of measurement variables to determine whether the two measurement variables tend to move together
3	Descriptive Statistics	Generates a report of univariate statistics for data, providing information about the central tendency and variability of your data.	Describes the data in an interpretable format and sow summary statistics like mean, mode, median, std. deviation, skewness, kurtosis and range etc.
4	Exponential Smoothing	Predicts a value based on the forecast for the prior period, adjusted for the error in that prior forecast	Forecast on the basis of a smoothing constant.
5	F-Test	Performs a two-sample F-test to compare two population variances.	Test that these two samples come from distributions with equal variances.
6	Moving Average	Projects values in the forecast period, based on the average value of the variable over a specific number of preceding periods.	Forecast trends on the basis of past figures.
7	Regression analysis	Performs linear regression analysis by using the "least squares" method to fit a line through a set of observations.	Analyse how a single dependent variable is affected by the values of one or more independent variables.
8	Sampling analysis	Creates a sample from a population by treating the input range as a population.	Infer about a population on the basis of a sample.
9	t-Test, z-Test	Determine whether the two samples are likely to have come from distributions with equal population means	Compare two population means when the population variances are known and unknown

Source: Based on the Microsoft Office Excel 2003 help function

You have seen that MS Excel can do virtually most of common statistical calculations. There are two more features that are worth mentioning when one talks about the statistical functions of MS Excel. These two are cross tabulations, pivot tables and the graphical features.

MS Excel can be used to create cross tabulations or two-way frequency tables across categorical variables. In MS Excel there is a pivot table wizard which helps in creating tables in multi-dimensions. Let us explain these concepts with the help of an example. The data given below is the percentage contribution of a country to world research in a particular subject area.

Table 9.2: Percentage Contribution of a Country to World Research in a Particular Subject

Countries	Chem.	Engg.	Clin. Med.	Phy.	Mat. Sc.	Plant. Sc.
Argentina	0.57	0.26	0.31	0.56	0.36	0.92
Australia	1.56	2.18	2.51	1.36	1.69	4.56
France	6.17	4.9	5.7	7.08	5.81	5.16
Germany	9.38	6.56	8.12	9.89	9.07	6.6
Hungary	0.83	0.44	0.25	0.54	0.41	0.61
India	4.02	2.68	0.85	2.51	4.02	3.49
Ireland	0.22	0.27	0.36	0.21	0.27	0.37
Israel	0.79	1.06	1.26	1.37	0.71	1.11
Italy	3.62	4.02	4.4	4.42	2.48	2.33
Japan	11.33	9.75	8.21	11.14	13.77	6.98

Now suppose you want to make a pivot table that would enable you to visualise a country whose contributions differ in the disciplines of physics and chemistry. You can simply drag the subject field in rows and column. This would enable you to see that e.g. the shaded countries Hungary and India have significantly different contributions.

Count of Countries	Chem.										Grand
Phy.	0.22	0.57	0.79	0.83	1.56	3.62	4.02	6.17	9.38	11.33	Total
0.21	1										1
0.54											1
0.56		1									1
1.36					1						1
1.37			1								1
2.51											1
4.42						1					1
7.08								1			1
9.89									1		1
11.14										1	1
Grand Total	1	1	1	1	1	1	1	1	1	1	10

There could be many more such cross-tabulations depending upon the need of the researcher. The most pleasant part of working with MS Excel is the ease with which you can drag these fields and have a customised layout as per your wish. It is said that in MS Excel, the most demanding work is to input the data, the analysis being the easiest. You can use pivot table effectively to present data where two-dimensional tables are important. One of the major advantages of this feature is that once the table is prepared, we can change the summary from one characteristic to another.

Similarly, if you want to make a graphical presentation of the data then you can go to the chart wizard and choose the chart that you want to make. MS Excel has a built-in facility to create graphs and charts. There are several types of charts and graphs supported by MS Excel like bar charts, line charts, pie charts and scatter diagrams etc. The chart wizard menu can be summoned by clicking on the graph icon from the menu bar. The chart wizard looks like the following

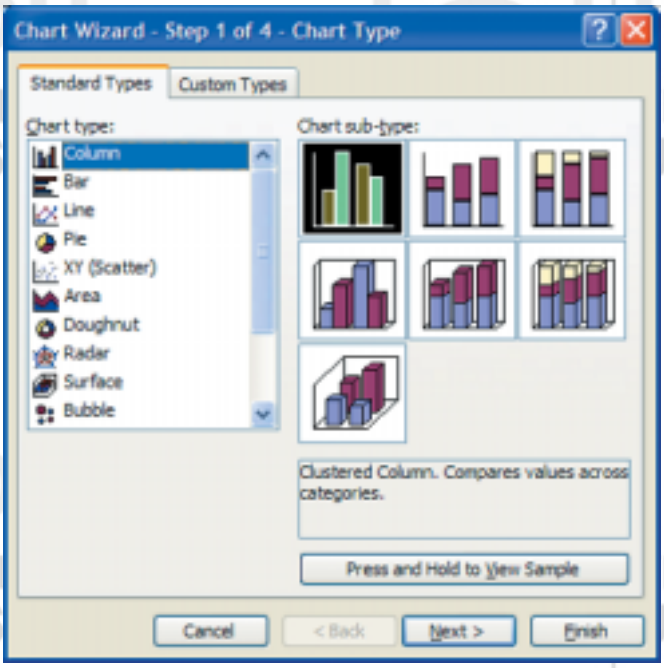


Fig. 9.5 : Screen Snapshot of Chart Wizard in MS-Excel

Now suppose you want to make a pivot table that would enable you to visualize a country whose contributions differ in the disciplines of physics and chemistry. You can simply drag the subject field in rows and column. This would enable you to see that e.g. the shaded countries Hungary and India have significantly different contributions.

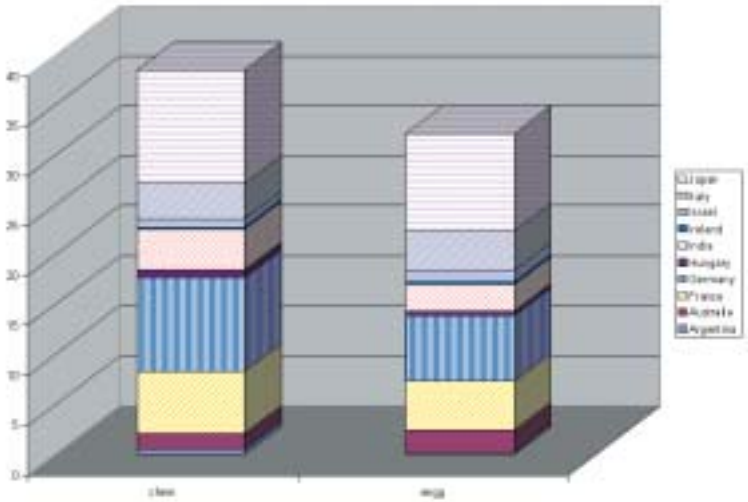


Fig. 9.6 : Screen Snapshot of Chart Companies of the Contribution of Countries to Chemistry & Engineering

So you have seen that it is pretty simple to create a graph and visualize your data once you have some data in your worksheet. These graphs are very useful for common user requirements. We can cut copy and paste these graphs to any document of MS Word or PowerPoint.

Self-Check Exercise

2) Enumerate the data analysis tools in MS-Excel.

Note: i) Write your answer in the space given below.

ii) Check your answer with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

9.3.2 SPSS

SPSS is the short form for Statistical Package for Social Sciences (SPSS). It is a very popular package due to its features and compatibility with other window-based programs. In the late 1960s, three Stanford University graduate students developed the SPSS statistical software system.

SPSS can take data input from many packages like dBase (*.dbf), Excel (*.xls), Lotus 123 (*.w*) and others like *.dat and *.txt. It can filter the data and perform analysis only in selected cases.

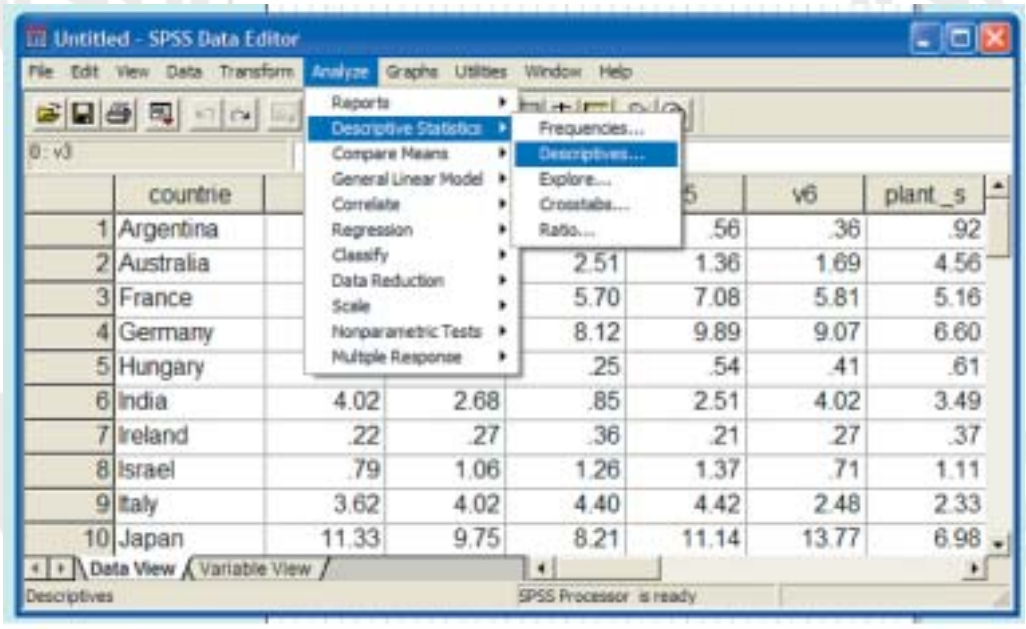


Fig. 9.7 : Screen Snapshot of SPSS Data Editor

Once you open a data file you can go to the analyze menu and start working on the statistical aspects of the data. Figure 9.7 shows you the menu for the descriptive statistics. You can find the frequencies, cross-tabulations, and ratios etc. You can see that there is a long list of statistical analysis in the analyze menu. We will give you a glimpse of what these functions do.

Fig. 9.8: Screen Snapshot of Help Topics in SPSS

Table 9.3 : Functions of Statistical Analysis in SPSS

Tool/ Module	Statistical Procedures
Report	OLAP (Online Analytical Processing) Cubes, Summarize procedure, Report Summaries in rows, Report Summaries in Columns
Descriptive Statistics	Frequencies procedure, Descriptive procedure, Explore procedure, Cross-tabs procedure, Ratio Statistics procedure
Compare means	Means procedure, One-Sample T Test procedure, independent-Samples T Test procedure, Paired-Samples T Test procedure, One-Way ANOVA procedure
General Linear Model	GLM Univariate procedure
Correlate	Bivariate Correlations procedure, Partial Correlations procedure, calculates statistics measuring either similarities or dissimilarities (distances)
Regression	Linear Regression procedure, Curve Estimation procedure
Classify	K-Means Cluster Analysis procedure, Hierarchical Cluster Analysis procedure, Discriminant Analysis procedure
Data Reduction	Factor analysis procedure
Scale	Reliability analysis procedure, Multidimensional scaling procedure
Non Parametric Tests	Chi-Square Test procedure, Binomial Test procedure, Runs Test procedure, One-Sample Kolmogorov-Smirnov Test procedure, Two-Independent-Samples Tests procedure, Tests for Several Independent Samples procedure, Two-Related-Samples Tests procedure, Tests for Several Related Samples procedure
Multiple Response	Define Multiple Response Sets procedure, Multiple Response Frequencies procedure, Multiple Response Cross-tabs procedure.

Source: SPSS for Windows Release 11.0 Help

One can see that many of these tools are available in MS Excel also, but the difference is that the output given by SPSS contains many more details regarding the statistical aspects of the findings. For example the cross-tabs procedure forms two-way and multi-way tables in both MS Excel and SPSS, but in SPSS it also provides a variety of tests and measures of association for two-way tables. The supporting statistics provided in SPSS is Pearson chi-square, likelihood-ratio chi-square, linear-by-linear association test, Fisher's exact test, Yates' corrected chi-square, Pearson's r, Spearman's rho, contingency coefficient, phi, Cramér's V, symmetric and asymmetric lambdas, Goodman and Kruskal's tau, uncertainty coefficient, gamma, Somers' d, Kendall's tau-b, Kendall's tau-c, eta coefficient, Cohen's kappa, relative risk estimate, odds ratio, McNemar test, and Cochran's and Mantel-Haenszel statistics. SPSS is thus more comprehensive.

SPSS also supports several statistical graphs. It displays many statistics on the graph itself. It has a feature that helps you to find a chart that is most suitable for your data, which is called "Chart Galleries by Data Structure".

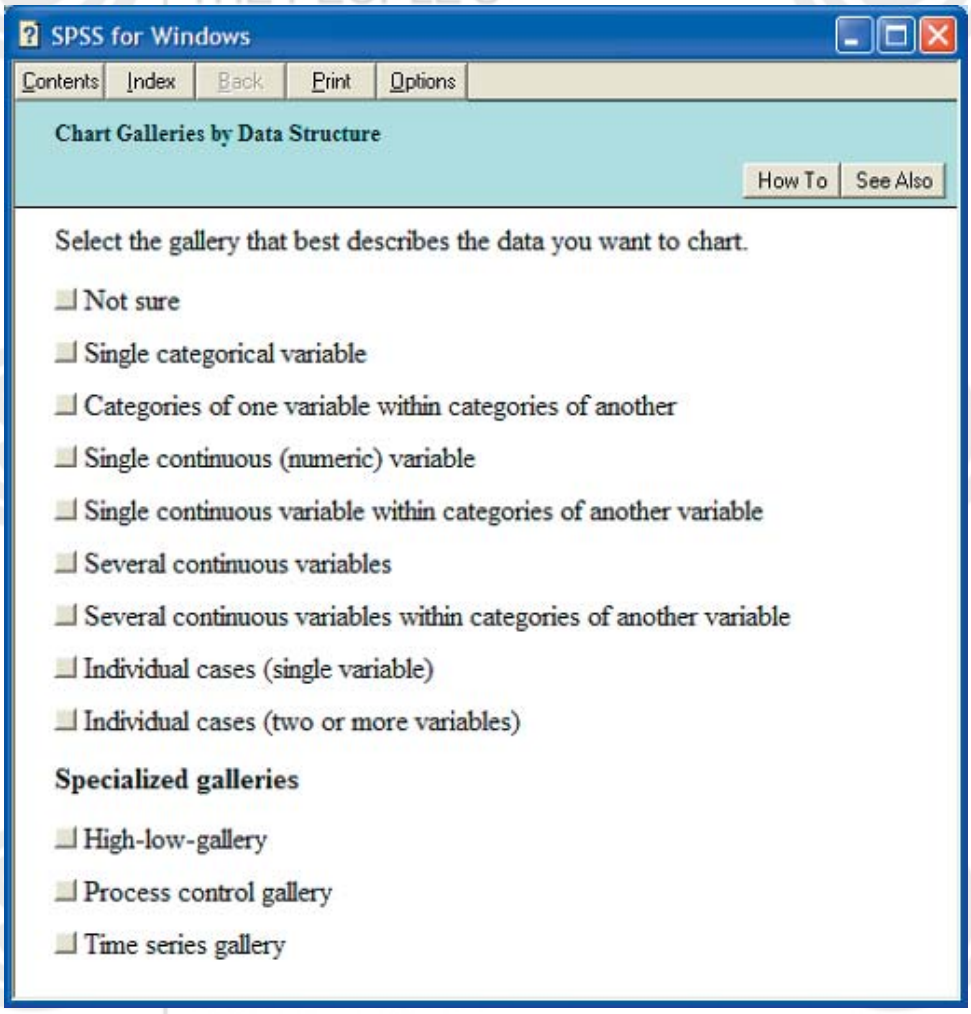


Fig. 9.9: ScreenSnapshot of Chart Galleries by Data Structure in SPSS

Now suppose you have chosen single categorical variable as the gallery that best describes your data. Then the next screen that will appear would be like the following figure. Suppose here you choose Simple Pareto Counts or Sums for Groups of Cases, then SPSS will describe what this graph does like, "Creates a bar chart summarizing categories of a single variable, sorted in descending order. A line shows the cumulative sum."

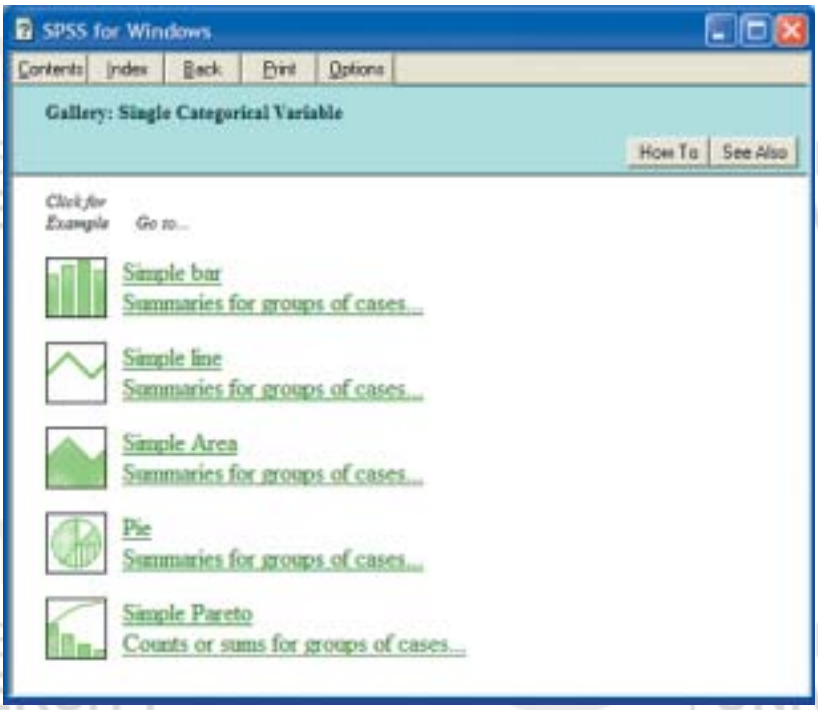


Fig. 9.10: Screen Snapshot of Gallery : Single Categorical Variable

So in this way, SPSS also acts as a mentor also. Probably, this is the reason for its success also. SPSS has a menu called "Statistics Coach", which asks questions about your data like "What do you want to do with your data?"

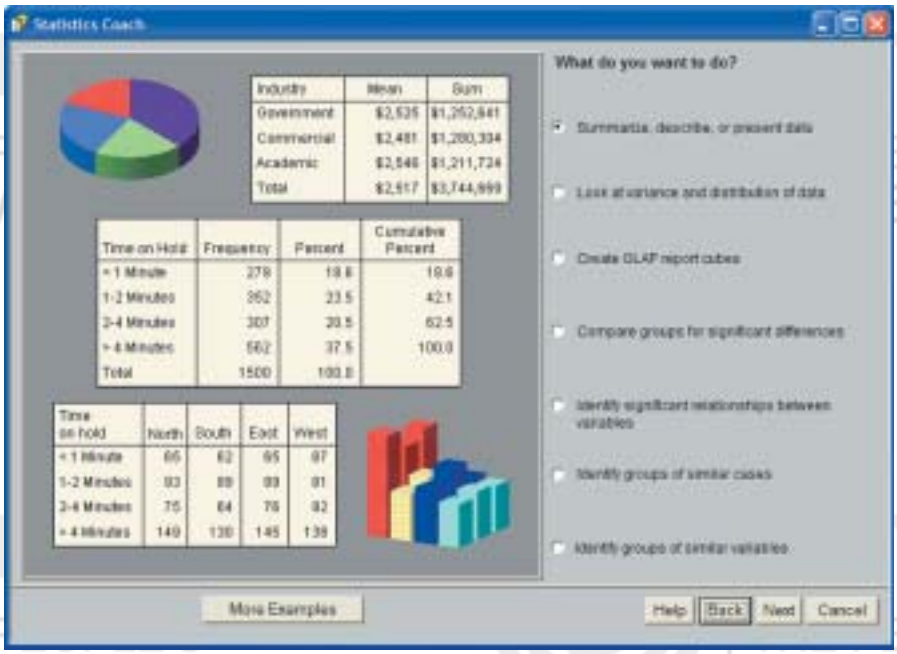


Fig. 9.11: Screen Snapshot of Gallery : Single Categorical Variable

It asks you further questions about your data in four steps and then suggests the right kind of analysis for your dataset. The output of SPSS appears as pivot table, which can be cut and pasted to Word documents, Excel worksheets and PowerPoint presentations. According to Wegman and Solka (2005), "SPSS supports numerous add-on modules including one for regression, advanced models, classification trees, table creation, exact tests, categorical analysis, trend analysis, conjoint analysis, missing value analysis, map-based analysis, and complex samples analysis".

Self Check Exercise

3) Enumerate the statistical analysis provisions in SPSS.

Note: i) Write your answer in the space given below.

ii) Check your answer with the answers given at the end of the Unit.

.....

.....

.....

.....

9.4 OTHER SOFTWARES FOR STATISTICAL ANALYSIS

Modern statistics can perform very large and complex calculations with the help of computers. There are a lot of softwares available in the market. Many of them are shareware, freeware and online pages that perform statistical calculations. Many of the universities offer statistics online computational resources e.g. http://www.socr.ucla.edu/ (University of California at LA), Statlib: Data, Software and News from the Statistics Community http://lib.stat.cmu.edu/ (Carnegie Melon University) and free statistical tools on the web http://www.cbs.nl/isi/ (International Statistical Institute). The list is endless. We therefore restrict ourselves by giving brief introduction to some of the popular statistical softwares. The table below presents this:

Table 9.4: Brief Description of Some Statistical Softwares

S.No.	Software and its URL	Brief Description
1.	The SAS System http://www.sas.com/	SAS evolved in the late 1960's at North Carolina State University. It has now become a system for complete data management and analysis. SAS represents the Microsoft of the statistical software companies. The SAS website claims that their software resides at 40,000 sites worldwide including 90 percent of those companies on the Fortune 500 list. SAS has given due importance to recent "statistical-like" advances like data mining. It has integrated mathematical/statistical methodologies, database technology, and business applications in an effective manner to remain at the top of the commercial statistical software arena.
2.	BMDP http://www.statsol.ie/bmdp/bmdp.htm	BMDP originated during the 1960s as a bio-medical analysis package. It still remains a clear favorite of biomedical field. BMDP has aligned itself with a number of other currently popular statistical products including StatXact 5.0, LogXact 5.0, SOLAS, EquivTest, SigmaPlot, Meta Analysis, and SYSTAT.
3.	S-PLUS http://www.insightful.com/	S-Plus is based on the statistical analysis language S. S-PLUS provides the user with a fully extensible environment by supporting the user to develop their own functions using the S-PLUS language. S-PLUS contains over 4,200 data analysis functions, which implement modern and robust statistical procedures. S-PLUS has superb graphical capabilities. It receives a strong support from the academic and commercial sectors.

S.No.	Software and its URL	Brief Description
4.	MINITAB http://www.minitab.com/	MINITAB combines an array of statistical methods, graphics tools, and project organization features in a user-friendly package. It combines the user-friendliness of MS Excel with the ability to perform complex statistical analysis. Thousands of successful companies worldwide, including GE, 3M, Ford Motor Company, and the leading Six Sigma consultants, use MINITAB to make data-driven decisions and achieve world-class quality.
5.	GLIM http://www.nag.co.uk/stats/GDGE_soft.asp	Generalized Linear Interactive Modeling package (GLIM) is a flexible, interactive statistical analysis program developed by the Royal Statistical Society. GLIM is not just a modeling package; it also contains many of the standard statistical procedures and has high-resolution graphics. Professionals, scientists and statisticians worldwide respect GLIM.

Source : **Wegman & Solka (2005) and websites of the softwares.**

So you have seen that there are many statistical packages that provide the state of the art facilities for performing statistical calculations. All these software are extremely good and it is for the user to work on the one on which he feel most comfortable. There is a competition among these software for providing enhanced statistical functions, enhanced user-friendliness, better graphics and sound technical support. Also there is trend to move towards "statistics-like" disciplines e.g. data mining.

9.5 SUMMARY

The science, which deals with numbers, is statistics. It crunches the numbers and organises them in a meaningful way so that information is generated. Computer can help immensely in the statistical analysis. There exists numerous statistical tools available and the need is to identify their actual usage. Most of the Statistical methods are based on the quantitative data. One can find different statistical packages for applications to different disciplines. In this Unit you have read about two such packages MS Excel and SPSS. The Unit has discussed some of their applications in details. You have also gone through a brief introduction of some of the popular statistical softwares. This Unit was intended to make you familiar with the basic statistical functions that can be performed with the help of computer and to arouse your interest in the beautiful and huge world of statistical computing.

9.6 ANSWERS TO SELF CHECK EXERCISES

1) A statistical package is defined as the software used to collect, organise, interpret and present numerical information. The need of a statistical package arises due to the complexity of calculations involved therein for analysis and inference. It helps to bring accuracy in results.

2) The data analysis tools in MS-Excel are : ANOVA, corrilation covariance, Descriptive Statistics, Expowertial Smoothing, F-Test, Moving Average, Regression Analysis, sampling Analysis, t-Test and z-Test.

3) The statistical analysis tools in SPSS are : Report, Descriptive Statistics, Compare Means, General Linear Model, Correlate, Regression, Classify, Data Reduction, Scale, Non-Pavansetric Tests and Multiple Response.

9.7 KEYWORDS

Statistics : It is a broad mathematical discipline which studies ways to collect, summarize and draw conclusions from data. In other words, it is used as a "collection of numerical facts or data".

Statistical Package : It is software for the collection, organization, interpretation, and presentation of numerical information.

Nominal Scale : Nominal measurement consists of assigning items to groups or categories.

Ordinal scale : Measurements with ordinal scales are ordered in the sense that higher numbers represent higher values.

Cardinal Scale : On interval measurement scales, one unit on the scale represents the same magnitude on the trait or characteristic being measured across the whole range of the scale.

Ratio scale : Ratio scales are like interval scales except they have true zero points. Ratio variables are very similar to interval variables.

Spreadsheet : It is a big worksheet (in many rows and columns). This worksheet can be used for data entry and for performing calculations by click of buttons.

9.8 REFERENCES AND FURTHER READING

Box, George E.P., William G. Hunter, and J. Stuart Hunter (1978). *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. New York: John Wiley and Sons.

Johnson R.A., Wichern D.W. (2003). *Business Statistics: Decision making with Data*, John Wiley & Sons (Asia) Pvt. Ltd. : Singapore, 2003.

Levin R.I., Rubin, D.S. (2001). *Statistics for Management*. Prentice Hall of India Private Limited : New Delhi.

Microsoft Office Excel 2003. *Help Documentation.*

Pottel Hans. *Statistical flaws in Excel.* Innogenetics NV, Technologiepark 6, 9052 Zwijnaarde, Belgium, http://www.mis.coventry.ac.uk/~nhunt/pottel.pdf. Downloaded 08/11/05

Ripley Brian D.(2004). *Statistical Methods Need Software.* Netherlands Statistical Society, 27 April 2004. http://www.stats.ox.ac.uk/pub/bdr/NethStatSoc.pdf Downloaded 08/11/05

Sarma K.V.S. (2001). *Statistics made simple- Do it yourself on PC*. Prentice-Hall of India: New Delhi.

SPSS for Windows Release 11.0. *Help Documentation.*

Wegman Edward J., Solka Jeffrey L. (2005). *Statistical Software for Today and Tomorrow*. Center for Computational Statistics, George Mason University Fairfax, VA 22030. Downloaded 16/11/05 http://www.galaxy.gmu.edu/papers/guide.pdf