BLOCK 4
DATA PROCESSING AND ANALYSIS

BLOCK 4 DATA PROCESSING AND ANALYSIS

Introduction

Block 4, will introduce you to the ways in which data processes and statistical methods were used in rural development. In particular, this block will deal with data processing and analysis, descriptive statistics, inferential statistics and reporting research.

Unit 13, 'Data Processing and Analysis' covered the characteristics of both quantitative and qualitative data, as well as the numerous graphic representations of quantifiable data and techniques for their analysis. The information gathered by using various techniques on the chosen samples is both qualitative and quantitative in nature. The act of tabulating involves arranging data into rows and columns to make comparisons easier and to highlight the relationships that are at play. We created analytical models that depict variables and their relationships diagrammatically while keeping in mind the goals of the research investigation. Tables that provide information about a single variable are referred to as univariate analysis. A bivariate table simultaneously displays data for two variables as percentages in the columns and rows. To determine whether a third variable is influencing the relationships between two variables, trivariate analysis is used. The methods of multivariate analysis are employed by researchers who are interested in evaluating the combined influence of three or more factors.

Unit 14, 'Descriptive Statistics'talked about the numerous ways to express quantitative dataas well as the approaches to quantitative data analysis. The quantitative information is tabulated in a "frequency distribution," which can be graphically represented using an ogive, a frequency polygon, and/or a histogram. The three different categories of descriptive statistical measures are central tendency, variability, and association. The three metrics for measuring central tendency are mean, median, and mode. The mean is the distribution's arithmetic average. It is calculated by dividing the total number of values by the sum of all observational values. The data is organised in a frequency distribution when there are many observations. An array's median is the place at which half of the values or measurements fall.

Unit 15, 'Inferential Statistics' covered the characteristics of both parametric and non-parametric tests as well as the underlying presumptions. Additionally, applications of several non-parametric tests (Chi-square) and parametric tests (Z, t, and F) in the analysis of data have been described. Data accessible in interval or ratio scales of measurement are subjected to parametric tests for analysis. The data must be regularly distributed or nearly so for a parametric test to be valid. The most widely used parametric tests are Z-test, t-test, and F-test. Frequencies are a type of discrete data that can be used with chi-square. When determining if a factor other than "chance" is responsible for the observed relationship between the variables, an independence test is used. If two independent samples have different central

tendencies, the median test is performed to determine whether this is the case. Any time the measures for the two samples are expressed using an ordinal scale, it is especially helpful.

Unit 16, 'Reporting Research' concentrated on research reporting as a professional activity. The reason for doing the research study determines the goal of the report's authoring. It might be required for a degree, a project report that must be turned in to a funding source, etc. Once the research has been completed, the sponsoring organisation and the school administrators may use the conclusions and recommendations to further their goals. Additionally, other researchers may use the research as a resource. A research report contains three sections: the introduction, the body, and the conclusion. The table of contents, the list of tables, and the list of figures are all included in the beginning, together with the cover or title page. The introduction, a review of the pertinent literature, objectives, hypotheses, research design analysis, and data interpretation are often found in the main body.

ignou
THE PEOPLE'S
UNIVERSITY

UNIT 13 DATA PROCESSING AND ANALYSIS

Structure

13.0 Introduction

13.1 Objectives

13.2 Types of Data: Quantitative and Qualitative

13.3 Quantitative Data

13.4 Processing and Analysis of Quantitative Data

13.5 Coding of Data

13.6 Preparing a Master Chart

13.7 Analysis of Quantitative Data

- 13.7.1 Univariate Analysis
- 13.7.2 Bivariate Analysis
- 13.7.3 Trivariate Analysis
- 13.7.4 Multivariate Analysis

13.8 Use of Computer in Data Processing and Tabulation

13.9 Let Us Sum Up

13.10 Glossary

13.11 Check Your Progress: The Key

13.0 INTRODUCTION

In Block 3, we dealt with the nature of various tools used in the collection of data. These data are mostly expressed in quantified terms. However, quantitative data may not be available in certain cases. In such a situation, the researcher has to consider the phenomenon as a whole without breaking it down into measurable or quantifiable components. As such, he/she should be familiar not only with the two types of data–quantitative and qualitative, but also with processing of data and the various methods of data analysis.

The aim of this Unit is to make you understand the nature of quantitative and qualitative data, the procedures for classifying and tabulating quantitative data, presenting them in tabular form, and the various methods used in data analysis.

13.1 OBJECTIVES

After the completion of this Unit, you should be able to:

- define quantitative and qualitative data;
- prepare code book and master chart;

- prepare analytic model for data analysis; and
- prepare various types of tables for presenting data.

13.2 TYPES OF DATA: QUANTITATIVE AND QUALITATIVE

The data collected through the administration of various types of tools on the selected samples are of two types: quantitative and qualitative. In quantitative data, numerical values are assigned to the characteristics or properties of objects or events, according to logically accepted rule. It is a process wherein a number system like figures, ratings or scores is imposed on empirical data. However, when the researcher takes into consideration the phenomenon as a whole and does not attempt to analyse it in measurable or quantifiable terms, the characteristics or properties of objects or events are known as qualitative data. We look at the characteristics of both in the following sections.

13.3 QUANTITATIVE DATA

Quantitative data describe an empirical event or phenomenon in a numerical system with the help of different scales of measurement: nominal, ordinal, interval or ratio. Nominal scales of measurement are used when a set of objects among two or more categories is to be differentiated on the basis of certain clearly known characteristics. For example, we may assign individual's categories such as sex (male and female), nationality (French and Indian), type of education (formal and non-formal), etc. The ordinal scales of measurement correspond to quantitative classification of a set of objects done with the help of ranking on a continuum. For example, we may rank students according to their height from the lowest to the highest. The interval scale of measurement is based on equal units of measurement. It includes how much or how little of a given characteristic or attribute is present. For example, the difference in the amount of an attribute possessed by individuals with test scores of 60 and 61 is assumed to be equivalent to that between individuals with scores of 50 and 51. Ratio scale is the highest level of measurement. Since this scale assumes the existence of absolute zero, this type of measurement is almost non-existent in educational and psychological measurement. Thus, quantitative data are expressed in nominal, ordinal, interval and ratio scales of measurement.

Types of Quantitative Data

Quantitative data may be classified into two categories:

I) Parametric and

II) Non-parametric

Parametric data are obtained by applying interval or ratio scales of measurement. Scores of the tests of ability, achievement, attitude, interest,

values, personality etc. are examples of interval scales of measurement. In the study of reaction time we use ratio scale. In this type of experiment, the zero point in the absolute sense is known and it makes sense to look at the ratio of the time taken to respond in different treatment situations.

Non-parametric data are either counted or ranked. In counted data, we make use of the nominal scale. Each individual can be a member of only one category and all the members of that category have the same, defined characteristics. For example, we may categorise a group as a sample of 'female students' of a particular 'study center' of an open university. The categorisation of teachers at different educational levels–school, college and university –is another example of nominal data.

In ranked data, we apply the ordinal scale of measurement. The sets or classes of objects are ordered on a continuum in a series ranging from the lowest to the highest according to the characteristics we wish to measure. The ranking of students in a class for height, weight or academic achievement are examples of ordinal data.

Check Your Progress 1

Define quantitative data. Describe the various types of quantitative data along with examples.

Note: a) Space is given below for your answer.

b) Compare your answer with the one given at the end of the Unit.

……

……

……

……

……

13.4 PROCESSING AND ANALYSIS OF QUANTITATIVE DATA

Once data are collected, the researcher turns his/her focus of attention on its processing. In this sub-section, we will discuss about one of the most important stages of the research process, i.e. data processing and analysis.

A researcher has to make his/her plan for each and every stage of the research process. As such, a good researcher makes a perfect plan of processing and analysis of data. To some researchers data processing and analysis is not a very serious activity. They feel many a time that data processing is a job of a computer assistant. As a consequence, they have to be contended with the results given by computer assistant which may not help them to achieve their objectives. To avoid suchsituations, it is essential that data processing must

be planned in advance and instructed to assistants accordingly. In this sub-section we will discuss about data processing and data analysis.

Data processing refers to certain operations such as editing, coding, computing of the scores, preparation of master charts, etc. After collection of filled in questionnaires, editing of entries therein are not only necessary but also useful in making subsequent steps simpler. Many a times, a researcher or the assistant either miss entries in the questionnaires or enter responses, which are not legible. This sort of discrepancies can be resolved by editing the schedule meticulously. Another problem comes up at the time of tabulation of data when researcher asks for tabulation of responses from consecutive questions. In cases where data are not cleaned there has to be inconsistency in the tabulations, the researcher has to be very particular about, consecutive questions where category 'not applicable' exists. In the process of editing, the researcher has to be very careful about consecutive questions having 'not applicable' as a response.

13.5 CODING OF DATA

Coding of data involves assigning of symbols (numerical) to each response of the question. The purpose of giving numerical symbols is to translate raw data into numerical data, which may be counted and tabulated. The task of researcher is to give numbers to response carefully. As we have already discussed various types of questions (such as open-end, etc.,) in the previous block, the coding scheme will vary accordingly. For example, a close-end question may be already coded and hence it has to be just included in the code book whereas coding of open-end questions involves operations such as classification of major responses and developing a response category of 'others' for responses which were not given frequently. The classification of responses is primarily based on similarities or differences among the responses. Usually, in the case of open-end questions, to classify responses, researcher looks for major characteristics of the responses and puts it accordingly. In case of attitude scales, researcher has to keep in mind, the direction or weightage of responses. For example, a response 'strongly agree' is coded as 'five'. The subsequent codes would be in order. Therefore, if there are responses like 'agree', 'undecided', 'disagree' and 'strongly disagree' they have to be coded as four, three, two, and one. Alternatively, if strongly agree is coded as minus two, the subsequent responses would be coded as minus one, zero, +1 and +4. The matrix questions have to be coded taking into consideration each cell as one variable. Forexample, if the column of matrix represents employment status, namely, 'permanent' and 'temporary' and row represents employers or type of employer, namely, government and private, the first cell would represent a variable 'government- permanent'. The second cell would represent 'government-temporary' and so on. In order to demonstrate the points discussed above a section of a code book used in a study is reproduced in Table 1.1

Table 1.1: Code Book

Q. No.	Var No.	Information Sought	Responses	Code	Column No.	Remarks
		Respondents Number			1-3	
2	1	Age	Actual Worker	-1	4-5	
3	2	Designation	Supervisor Manager	2 3	6	
4	3	Establishment	Public Private Graduate Intermediate	1 2 1 2	7	
5	4	Level of Education	High School Middle School Primary Illiterate Other Married	3 4 5 6 7 1	8	
6	5	Marital Status	Unmarried Widow Divorce	2 3 4	9	
36	35	Nature of Work	Yes No	1 2	10	
	36	Duration of Work	Yes No	1 2	11	
	37	Wages	Yes No	1 2	12	
	38	Promotion	Yes No	1 2	13	
43	42	Attitude of Employer Preference for male employees	Agree Undecided Disagree	1 2 3	14	

13.6 PREPARING A MASTER CHART

After a code book is prepared, the data can be transferred either to a master chart or directly to computer through a statistical package. Going through master chart to computer is much more advantageous than entering data

directly to computers because one can check the wrong entries in the computer by comparing 'data listing' as a computer output and master chart. Entering data directly to computer is disadvantageous, as there is no way to check wrong entries, which will show inconsistencies in tabulated data at the later stages of tabulation. A sample of master chart prepared in accordance with the code book is presented in Table 1.2.

Table 1.2: Master Chart

Correspondent Number	**Variables Labels**												
	Question/Variable/Column Number												
1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	0	1	1	4	1	2	1	2	3	2	3	4	3
0	0	2	2	1	2	3	4	5	6	7	8	9	5
0	0	3	3	3	3	3	4	2	6	7	2	6	3
0	0	4	4	5	7	8	9	1	3	5	6	1	1
0	0	5	5	4	5	1	4	2	1	1	4	3	5
0	0	6	6	3	1	2	3	4	5	6	3	1	5
0	0	7	7	5	4	5	6	9	7	8	5	2	4
0	0	8	8	1	4	2	5	3	6	3	7	8	9
0	0	9	9	2	2	4	2	6	7	8	2	1	5
0	0	0	0	9	5	6	8	7	9	2	4	4	3
0	0	1	1	2	2	8	4	9	3	4	7	3	8
0	0	2	2	2	5	7	9	5	1	4	2	4	3
0	0	3	3	2	9	4	5	6	7	2	6	9	6
0	0	4	4	2	8	7	9	5	2	4	6	2	3
0	0	5	5	3	5	4	8	7	9	2	4	2	3
0	0	6	6	2	4	8	7	9	5	8	4	5	6
0	0	7	7	8	7	4	9	4	3	4	6	3	4

13.7 ANALYSIS OF QUANTITATIVE DATA

The purpose of data analysis is to prepare data as a model where relationships between the variables can be studied. Analysis of data is made with reference to the objectives of the study and research questions if any. It is also designed

to test the hypothesis. Analysis of data involves re-categorization of variables, tabulation, explanation and casual inferences.

The first step in data analysis is a critical examination of the processed data in the form of frequency distribution. This analysis is made with a view to draw meaningful inferences and generalisation.

The process of categorisation in accordance with the objectives and hypothesis of the study is arrived at with the help of frequency distributions. Re-categorisation is a process to arrange categories with the help of statistical techniques. This helps researcher to justify the tabulation. We have seen earlier that the responses to a statement may be assigned scores or weightage. These scores or weightage are summated and re-categorised on the bases of high, medium and low. The basic principle in the process of categorisation or re-categorisation is that the categories thus obtained must be exhaustive and mutually exclusive. In other words the categories have to be independent and not overlapping.

13.7.1 Univariate Analysis

Univariate analysis refers to tables, which give data relating to one variable. Univariate tables which are more commonly known as frequency distribution tables show how frequently an item repeats. Examples of frequency tables are given below. The distribution may be symmetrical or asymmetrical. The characteristics of the sample while examining the percentages, further properties of a distribution can be found out by various measures of central tendencies. However, researcher is required to decide which is most suited for this analysis. To know how much is the variation, the researcher has to calculate measures of dispersion. (The details about all these measures are given in Unit 2 of this Block.)

Usually, frequency distribution tables are prepared to examine each of the independent and dependent variables. Tables 1.3, 1.4 and 1.5 present two independent variables and one dependent variable.

Table 1.3: Table Showing Awareness of the Respondents

(The Independent Variable)

Level of Awareness	Distribution of Respondents	
	Frequency Percentage	
High	110	39.3
Medium	106	37.9
Low	64	21.8
Total	280	100.0

Table 1.4 : Table Showing the Respondents by Regional Development

(The Independent Variable)

Regional Development	Distribution of Respondents	
	Frequencies	Percentage
High	142	57.7
Medium	86	30.7
Low	52	14.6
Total	280	100.0

Table 1.5 : Table Showing Wage Differentials of the Respondents

(The Dependent Variable)

Wage Differentials	Distribution of Respondents	
	Frequencies	Percentage
High	78	27.9
Medium	134	47.9
Low	68	24.2
Total	280	100.0

A frequency distribution of a single variable is the frequency of observation in each category of a variable. For example, an examination of the pattern of response tovariable 'awareness of the respondents' in Table 1.3 would provide a description of the number of respondents who have high, medium and low level of awareness. In case of nominal variables categories can be listed in any arbitrary order. Thus, the variable "Religion" may be described with the category 'Hindu' or the category 'Christian' listed first. However, the categories of ordinal, interval and ratio variables are arranged in order. Let us consider the frequency distribution (Tables 1.3, 1.4 and 1.5) which describes the awareness, wage differentials and regional development of respondents. The tables have four rows, the first three being the categories of variables, which appear in the left-hand columns and the right hand columns show the number of observation in each category. The last rows are the totals of all frequencies appearing in tables. To analyse the data it is necessary to convert the frequencies into figures that can be interpreted meaningfully. Note, for instance, while distribution of respondents by regional development displayed in Table 1.4 clearly shows the predominance of respondents from 'high' development region whereas, distribution of respondents by wage differential in Table 1.5 indicates that the proportions of respondents with 'high' and 'low' wage differentials are almost equal.

13.7.2 Bivariate Analysis

A researcher might be interested in knowing the relationships between the variables. To know the relationship between the variables, the data pertaining to the variables are cross-tabulated. Hence, a bivariate table is also known as cross table. A bivariate table presents data of two variables in column percentages and row percentages simultaneously. An example of a bivariate table is given below:

Table 1.6: Table Showing Levels of Awareness Towards the Act andWage Differentials of the Respondents

Awareness about the Act	Wages Differentials			Total
	High	Medium	Low	
High	96 (66.2)	9 (10.5)	7 (13.5)	110 (39.3)
Medium	37 (26.1)	58 (67.4)	11 (21.2)	106 (37.9)
Low	11 (7.7)	19 (22.1)	34 (65.3)	64 (22.8)
Total	142 (57.70)	86 (30.7)	52 (14.6)	280 (100.0)

The table presents data with regard to two variables namely awareness about the Equal Remuneration Act and the level of wage differential. First row presents data with regard to respondents who were aware of the Act. The second row presents data about who were not. Similarly, the first column gives data pertaining to workers who have low wage differentials. The second column presents data of workers whose differentials were medium and the last column represents the respondents who felt high wage differentials. For example, the first cell (in the left-hand corner) represents 94 respondents who were fully aware of the Act and perceived high wage differentials.

The association between two variables can be explained either by comparing the percentages of respondents column wise or row wise. The relationship between the variables can also be examined by various statistical techniques (The details about all these measures are given in Unit 3 of this Block.) depending upon the level of measurement of the data. Apparently, the two variables are associated therefore with more people who were aware of the Act have perceived low wage differentials than who were not aware of the Act. Alternatively, comparatively smaller percentage of people who were aware of the Act has perceived high wage differentials than people who were not aware of the Act.

13.7.3 Trivariate Analysis

Sometimes researcher might be interested in knowing whether there is a third

variable which is effecting the relationships between two variables. In such cases the researcher has to examine the bivariate relationship by controlling the effects of third/variable. This is performed in two ways. One way of controlling the effects of a third/variable is to prepare partial tables and examine the bi-vairate relationship, and the second method of assessing the effects of a third/variable is to compare the co-efficient of partial correlations. Let us take an example. In the above table, if researcher wants to examine whether there is effect of regional development on the bivariate relationship he may prepare three partial tables giving data relating to awareness of the Act and wage differential for high, medium and low regional development.

Table 1.7 : Regional Development = High (N = 142)

Awareness about the Act	Wages Differentials			Total
	High	Medium	Low	
High	97 (21.9)	10 (21.8)	3 (9.3)	20
Medium	313 (40.6)	26 (33.3)	9 (28.1)	48
Low	12 (37.5)	42 (43.8)	20 (62.5)	74
Total	32	78	32	142

Table 1.8 : Regional Development = Medium (N = 86)

Awareness about the Act	Wages Differentials			Total
	High	Medium	Low	
High	8 (36.4)	11 (25.6)	3 (19.0)	23
Medium	9 (40.9)	17 (39.5)	9 (38.1)	34
Low	5 (34.9)	15 (34.9)	20 (42.9)	29
Total	22	43	21	86

Table 1.9 : Regional Development = Low (N = 52)

Awareness about the Act	Wages Differentials			Total
	High	Medium	Low	
High	14 (58.3)	7 (53.8)	7 (46.7)	28
Medium	8 (33.3)	4 (30.8)	6 (40.0)	18
Low	2 (8.3)	2 (15.4)	2 (13.3)	6
Total	24	3	15	52

On examination of these three partial tables, if the researcher finds out that bivariate relationships do not hold good he/she may infer that it is the third variable, the regional development which is affecting the bivariate relationship. In the partial tables for higher regional development, the proportion of people perceiving high wage differential are those who are having high level of awareness about the Act. The similar trend can be noticed in the remaining two partial tables. Which means regional development does not effect the bivariate relationships between wage differential and awareness about the Act.

In the other method, the researcher tries to find out the effect of third variable on the bivariate relationship with the help of partial correlation. To find out whether the third variable effects the bivariate relationship between awareness (X) and wage differentials (Y), he/she considers work experience of respondents.

To examine this, he/she calculates the correlation coefficients between the 'awareness' and 'wage differential' keeping 'experience' as a constant. The results are presented below.

X1 = Awareness (independent variable)

X2 = Wage differential (independent variable)

X3 = Experience (control variable)

γ1.2 (Correlation coefficient between X1 and X2) = .70

γ1.3 (Correlation coefficient between X1 and X3) = .40

γ2.3 (Correlation coefficient between X2 and X3) = .50

The researcher on examination of the results of the partial correlation finds that even after controlling the third variable, the relation between the first two variables awareness and wage differentials is statistically significant. Thus, he/she concludes that the relationship between awareness and wage differentials is not spurious.

13.7.4 Multivariate Analysis

When a researcher is interested in assessing the joint effect of three or more variables he/she uses the techniques of multivariate analysis. The most common statistical technique used for multivariate analysis is regression analysis (For this statistical technique reader is advised to refer books listed in For Further Reading at the end of this Unit). In the first step of multivariable analysis, the researcher has to obtain the correlation between the variables which are having statistically significant correlation. These variables are put in the regression analysis. One important point in applying correlation and regression analysis is the data must be measured on ratio or interval level. Another point a researcher has to keep in mind is that these two statistical techniques are based on certain assumptions. Hence,

before applying these techniques, the researcher has to see whether the sample selected by him fulfills those conditions.

13.8 USE OF COMPUTER IN DATA PROCESSING AND TABULATION

Research involves large amounts of data, which can be handled manually or by computers. Computers provide the best alternative for more than one reason. Besides its capacity to process large amounts of data, it also analyses data with the help of a number of statistical procedures. Computers carry out processing and analysis of data flawlessly and with a very high speed. The statistical analysis that took months earlier takes now a few seconds or few minutes. Today, availability of statistical software and access to computers has increased substantially over the last few years all over the world.

While there are many specialised software application packages for different types of data analysis, Statistical Package for Social Sciences (SPSS) is one such package that is often used by researchers for data processing and analysis. It is preferred choice for social work research analysis due to its easy-to-use interface and comprehensive range of data manipulation and analytical tools.

Basic Steps in Data Processing and Analysis by Computer

There are four basic steps involved in data processing and analysis using SPSS. They are:

1) Entering of data into SPSS,
2) Selection of a procedure from the menus,
3) Selection of variables for analysis, and
4) Examination of the outputs.

You can enter your data directly into SPSS Data Editor. Before data analysis, it is advised that you should have a detailed plan of analysis so that you are clear as to what analysis is to be performed. Select the procedure to work on the data. All the variables are listed each time a dialog box is opened. Select variables on which you wish to apply a statistical procedure. After completing the selection, execute the SPSS command. Most of the commands are directly executed by clicking 'O.K'. on the dialog box. The processor in the computer will execute the procedures and display the results on the monitor as 'output file'.

Check Your Progress 2

Note: a) Space is given below for your answer.

b) Compare your answe with the one given at the end of the Unit.

1. What is tri-variate analysis? Discuss the process of examining the effect of third variable on the bivariate relationship.

……

……

……

……

……

13.9 LET US SUM UP

In this Unit, we discussed the nature of quantitative and qualitative data, the various methods of representing the quantified data graphically, and the methods used in the analysis of quantitative and qualitative data. The main points are as follows:

1) The data collected through the administration of various tools on the selected samples are of (i) quantitative and (ii) qualitative nature.

2) Quantitative data are expressed in nominal, ordinal, interval or ratio scales of measurement. These data are classified into two categories: (i) parametric and (ii) non-parametric. The parametric data are obtained by applying interval or ratio scales of measurement, whereas non-parametric data are either enumerated or ranked. In the enumerated data we make use of nominal scale and in the ranked one we apply ordinal scale.

3) Data processing refers to certain operations such as editing, coding, computing of the scores, preparation of master charts, etc.

4) Many a times, a researcher or the assistants either miss entries in the questionnaires or enter responses, which are not legible. This sort of discrepancies can be resolved by editing the questionnaires.

5) Coding of data involves assigning of symbols (numerical) to each response of the question. The purpose of giving numerical symbols is to translate raw data into numerical data, which may be counted and tabulated.

6) With the help of a codebook the data can be transferred to a master chart.

7) Tabulation is a process of presenting data in rows and columns in such a way as to facilitate comparisons and show the involved relations.

8) Keeping in view the objectives of research study we set up analytic models which diagrammatically present variables and their interrelationships.

9) Univariate analysis refers to tables, which give data relating to one variable.

10) A bivariate table presents data of two variables in column percentages and row percentages simultaneously.

11) Trivariate analysis is undertaken to know whether there is a third variable which is effecting the relationships between two variables.

12) When a researcher is interested in assessing the joint effect of three or more variables he uses the techniques of multivariate analysis.

13.10 GLOSSARY

Quantitative Data — Data which are expressed in nominal, ordinal, interval or ratio scales of measurement.

Qualitative Data — Data which are available in the form of detailed descriptions of situations, events, people, interactions, and observed behaviour, direct quotations from people about their experiences, attitudes, beliefs, and thoughts, and excerpts from documents, correspondence, records, and case histories.

Parametric Data — There are data which are got by applying interval or ratio scales of measurement.

Non-Parametric Data — These are data which are got by applying nominal or ordinal scales of measurement. These types of data are either countered or ranked.

13.11 CHECK YOUR PROGRESS: THE KEY

1) Quantitative data is the description of an empirical event or phenomenon in a numerical system presented with the help of different scales of measurement such as nominal, ordinal, interval and ratio. The two major types of quantitative data are: parametric (obtained through interval or ratio scales) and non- parametric (counted by a nominal scale or ranked by an ordinal scale).

2) Trivariate analysis is used to know whether there is a third variable which is effecting the relationships between two variables. In such cases the researcher has to examine the bivariate relationship by controlling the effects of third variable. This is performed in two ways. One way of controlling the effects of a third variable is to prepare partial tables and examine the bivariate relationship, and the second method of assessing the effects of a third variable is to compare the co-efficient of partial correlations.

For Further Reading

Bailey, Kenneth, D. (1978), Methods of Social Research, The Free Press, London.

Baker, L. Therese (1988), Doing Social Research, McGraw Hill, New York.

Blalock, Hubert, M. (1972), Social Statistics, Mc Graw Hill, New York.

Doby J.T. (et. al.) (1953), An Introduction to Social Research, Harrisburg: Stackpole Co., London.

Kerlinger Fred R. (1964), Foundations of Behavioural Research, Surjeet Publications, Delhi.

Lal Das, D.K. (2000), Practice of Social Research: A Social Work Perspective, Rawat Publications, Jaipur.

Monete, Duane R. (et. al.) (1986), Applied Social Research, Tool For the Human Services, Holt, Chicago.

Nachmias, D and Nachmias, C. (1981), Research Methods in the Social Sciences, St. Martins Press, New York.

Sellitz. G. et. al. (1973), Research Methods in Social Relations, Rinehart and Winston (3rd edition) Holt, New York.

Sjoberg, G. and Nett., B. (1968), A Methodology of Social Research, Harper &Row, New York.

Wilkinson, T.S. and Bhandarkar, P.L. (1977), Methodology and Techniques of Social Research: Himalaya, Bombay.

UNIT 14 DESCRIPTIVE STATISTICS

Structure

14.0 Introduction

14.1 Objectives

14.2 Tabulation and Organization of Quantitative Data

- 14.2.1 Frequency Distribution
- 14.2.2 Cumulative Frequency Distribution
- 14.2.3 Graphical Presentation of Quantitative Data

14.3 Analysis of Quantitative Data

14.4 Let Us Sum Up

14.5 Glossary

14.6 Check Your Progress: The Key

14.0 INTRODUCTION

In Unit 1 we dealt with the data processing and data analysis. The aim of this Unit is to make you understand the procedures for classifying and tabulating quantitative data, presenting them graphically, and the various statistical methods used in data analysis.

14.1 OBJECTIVES

After the completion of this Unit, you should be able to:

- prepare various types of graphs and tables for presenting data; and
- compute measures of central tendency, variance, standard deviation, measures of relative position and measure of relationships.

14.2 TABULATION AND ORGANIZATION OF QUANTITATIVE DATA

14.2.1 Frequency Distribution

Data collected by using questionnaires/interview schedules or other measuring tools are raw and may have little meaning to the researcher until they are tabulated and organised in a systematic order. One of the ways of doing so is to prepare a frequency distribution. The method for tabulating the quantified data in a frequency distribution can be illustrated by considering the following scores of 40 students of say, Post Graduate Diploma in Rural Development of the Indira Gandhi National Open University.

Table: 14.1Tabulation of Scores in Course PGDRD

57	70	80	82	87
60	72	80	82	88
64	73	80	82	87
67	70	78	80	93
67	76	77	84	95
62	76	78	85	97
61	75	80	85	98
63	70	78	85	90

It is difficult to see from the above list how the scores are distributed. Inspection of scores, however, shows that many scores occur more than once. We observe that there are one 98, one 97, one 95, one 88, two 87s, three 85s, and so on. For our convenience, we can arrange the data in columns as shown in Table 2.1. In one column, we can arrange the marks in class-intervals and in the other, we can record the number of students who have scored these marks by tallies. Inspection of the scores in Table 2.1 shows that the highest score is 98 and the lowest is 57. The range is 41 (i.e. 98-57). Therefore, the distribution of scores can be conveniently arranged by dividing the range of 41 into eight or more class intervals if the classes are taken to be of 5 points each, If we take the starting point at 56, the scores within the range 56 to 60, that is all scores with the values 57 and 60 will be grouped together to form the lowest class-interval. All scores from 61 to 65, that is, 61, 62, 63, 64, and 65 will form the next class-interval. Similarly we shall group all scores within the ranges 66 to70, 71 to 75 and so on. The highest class interval will be 96-100.

Table 14.2 : Frequency Distribution of the Scores of 40 Students

Class Interval	Tallies	Frequency (F)
96-100	II	2
91-95	II	2
86-90	IIII	4
81-85	𝍸 II	7
76-80	𝍸𝍸III𝍸	11
71-75	III	3
66-70	𝍸I	5
61-65	III	4
56-60	II	2
	Total Number of Scores	N = 40

In Table 2.2, the class-intervals have been arranged serially from the smallest at the bottom to the largest at the top, each class-interval covering 5 scores. For each score, we have marked a 'tally' against the corresponding class-interval. The first score, 57, is represented by a tally placed against the class interval 56-60, the second score of 60 by a 'tally' marked against the class interval 56-60, and the third score 64 by a tally against the class interval 61-65. The remaining scores have been tabulated in the same way. When all the 40 scores are listed, the total number of tallies in each class-interval are counted and written in the next column f. The total of 'f' gives the total number of scores (in the present case 40) and is denoted by N. It may be noted that the interval 56-60 takes care of all the scores from 56 upto 60. The score of 56 ordinarily means the interval 55.5 to 56.5 and that the score of 60 means 59.5 to 60.5. The mid-point of the bottom most class-interval is 58. Hence, the distribution represented in Table 2.2 may also be expressed as:

Table 14.3 : Frequency Distribution of the Scores of 40 Students

Score Intervals	*Exact Unit of Class Intervals*	*Mid Point (x)*	*(f)*
96 - 100	95.5 - 100.5	98	2
91 - 95	90.5 - 95.5	93	2
86 - 90	85.5 - 95.5	88	4
81 - 85	80.5 - 85.5	83	7
76 – 80	75.5 - 80.5	78	11
71 - 75	70.5 - 75.5	73	3
66 - 70	65.5 - 70.5	68	5
61 - 65	60.5 - 65.5	63	4
56 - 60	55.5 - 60.5	58	2
N=40			

Check Your Progress-1

Tabulate the following data in a frequency distribution using an interval of 5 units.

185	175	166	177	171
147	176	170	171	180
173	168	181	165	173
175	158	156	162	173
197	151	153	162	188
166	145	191	164	174
178	142	158	167	178

148	187	172	169	184
156	187	172	193	183
181	161	172	179	179

Note:a) Space is given below for writing your answer.

b) See whether you have followed the procedure as shown in Tables 2.2 and 2.3.

c) The answer is incorporated in the one for check your progress 2.

……

……

……

……

……

Table 14.4 : Cumulative Frequency Distribution of the Test Scores of 40 students

Score Intervals	Exact unit of class-intervals	Frequency (f)	Cumulative frequency (f)	Cumulative percentage frequency
96-100	95.5-100.5	2	38+2=40	100.00
91-95	90.5-95.5	2	36+2=38	95.00
86-90	85.5-95.5	4	32+4=36	90.00
81-85	80.5-85.5	7	25+7=32	80.00
76-80	75.5-80.5	11	14+11=25	62.50
71-75	70.5-75.5	3	11+3=14	35.00
66-70	65.5-70.5	5	6+5=11	27.50
61-65	60.5-65.5	4	2+4=6	15.00
56-60	55.5-60.5	2	2	5.00
		N=40		

14.2.3 Graphical Presentation of Quantitative Data

Graphical presentation often facilitates understanding of a set of data. With the help of a well-drawn graph, the data can be read and interpreted very easily. Brief descriptions of the various types of graph which are useful invisualizing the important properties of a frequency distribution are given below.

The following three types of graph are commonly used for the above mentioned purposes:

i) Histogram or column diagram

ii) Frequency polygon

iii) Cumulative percentage curve or ogive.

i) *Histogram or column diagram*

A histogram or column diagram is a graph in which class-intervals are represented along the horizontal axis and their corresponding frequencies are represented by areas in the form of rectangular vertical bars drawn on the intervals. The following steps are followed in preparing a histogram:

Step 1 : A horizontal line is drawn at the bottom of a graph paper. Units representing class-intervals are marked along this line.

Step 2 : A vertical line is drawn at the left hand extreme of the horizontal axis. Along this vertical axis, units representing individual frequencies of the class-intervals are marked.

Step 3 : Taking class units as bases, rectangles are drawn, such that the areas of rectangles are proportional to the frequencies of the corresponding classes.

Let us consider the following data for drawing a histogram as an illustration of what we have said above.

Table 14.5 : Frequency Distribution of the Scores of 40 Students

Class intervals	Exact units of class-intervals	Mid-point	Frequency (f)	Cumulative frequency (f)	Cumulative percentage frequency
35-39	34.5-39-5	37	4	40	100.00
30-34	29.5-34.5	32	8	36	90.00
25-29	19.5-24.5	27	11	28	70.00
20-24	24.5-29.5	22	8	17	42.50
15-19	14.5-15.5	17	6	9	2.50
10-14	19.5-24.5	12	3	3	7.50
N=40					

The histogram drawn for the above data shown is shown in Figure 2.1.

Fig. 14.1: Histogram plotted from the data of Table 2.5

ii) *Frequency Polygon*

Frequency polygon is drawn by plotting the mid-point of each class-interval at a height proportional to its respective frequency and then joining the points

by straight lines. The first two steps are identical to those used in the construction of a histogram. The next step to be followed is given as under:

Fig. 14.2: Frequency polygon plotted from the data of Table 2.5

Step 3: Directly above the mid-point of each class-interval along the horizontal axis plot the points at a height proportional to the respective frequencies. Join these points by straight lines. The frequency polygon for the distribution of Table 2.5 is shown in the Figure 2.

iii) ***Cumulative Percentage Curve or Ogive***

When the frequencies are expressed as cumulative percentages of N on the vertical axis, the graphic representation is known as a cumulative percentage curve or ogive. After finding the cumulative percentage frequencies, the points are plotted on the exact upper limits of the class-intervals. A curve joining the points thus obtained is called the cumulative percentage curve or ogive. The cumulative percentage curve or ogive of the distribution represented in table 2.5 is illustrated in Figure 3.

Fig. 14.3: Cumulative percentage curve or ogive plotted from the data of Table 2.5.

14.3 ANALYSIS OF QUANTITATIVE DATA

Analysis of quantified data means studying the organised or tabulated data in order to discover the inherent facts. The data are studied from as many angles as possible to explore the new facts. Two types of statistical methods are used in the analysis of the tabulated data measured/expressed in quantified terms. The first category of methods pertain to 'descriptive analysis' and the second, to 'inferential analysis' of data. In this Unit, we will be concerned with 'descriptive analysis'.

Descriptive statistical analysis limits generalisation to the particular observed group of individuals. This analysis describes only one single group. The computed statistical values provide valuable information about the nature of that particular group only.

The following methods are generally used in descriptive statistical analysis of the tabulated data:

i) Measures of central tendency.

ii) Measures of variability.

iii) Measures of relationships.

iv) Measures of relative positions.

A) **Measures of Central Tendency**

The three most commonly used measures of central tendency are the Mean, the Median and the Mode.

I) ***The Mean (M)***

The arithmetic average of a distribution is known as its mean. The mean of a set of observations or measures is obtained by dividing the sum of all values by the total number of values.

B) ***Mean for an upgrouped data:***

The formula for finding the mean for an ungrouped data is:

$$M = \frac{\sum X}{N} \qquad (1)$$

in which

M = Mean

@= Sum of

X = Observations in a distribution

N = Total number of observations.

To illustrate the use of formula (1) let us consider the data:

16, 14, 12, 18, 21, 22, 13, 15, 16, 18

Using the formula (1):

$$M = \frac{\sum X}{N} = \frac{16+14+12+18+21+22+13+15+16+18}{10}$$

$$\frac{\sum X}{N} = 16.5$$

b) ***Mean for grouped data***

When the number of observations or measures is large, the data is grouped in a frequency distribution.

The mean is computed by the formula:

$$M = AM + \frac{\sum fx'}{N} \times xi \qquad (2)$$

where

M = Mean
AM = Assumed Mean
x' = [Mid point score (x) – AM]/(length of the class-interval)
$\Sigma fx'$ = Sum of the products of frequencies and deviation of observations from the assumed mean.
i = Width of the class-interval
N = Total number of observations.

To illustrate the use of formula (2), consider the grouped data given in Table 2.5

Computations

Step 1 : Put the class-intervals in exact limits

Step 2 : Find the mid-point of each class interval and take the assumed mean at the interval which has the maximum frequency.

Step 3 : Find the difference between each mid-point score and the assumed mean and divide it by the length of the class-interval to get the deviation x.

Step 4 : Compute fx for each class-interval (fx is the product of the frequency and deviation of the observationfrom the assumed mean in a particular case).

Step 5 : Find the sum of all fx

Step 6 : Apply formula (2) to compute the mean.

Table 2.6: The Calculation of the Mean from Data Grouped into aFrequency Distribution (Ref. Table)

Class Interval	Mid-point (X)	Frequency (F)	Deviation from the AM x	fx
34.5-39.5	37	4	2*	8
29.5-34.5	32	8	1	8
24.5-29.5	27 AM	11	0	0 (+16)
19.5-24.5	22	8	-1	-8
14.5-19.5	17	6	-2	-12
9.5-14.5	12	3	-3	-9 (-29)
		N=40		

$2^* = \frac{37-27}{5}$

Using the formula (2):

$$M = AM + \frac{\Sigma fx}{N} Xi$$

$$= 27.0 + \frac{-13}{40} x5$$

$$= 24.0 - 1.625$$

$$= 25.375$$

II) ***The Median***

The median is a point in an array, above and below which one half of the observations fall. It is a measure of position rather than magnitude.

a) ***Median for Ungrouped data***

If the observations are ungrouped and their number is small, the observations are arranged in the order of magnitude. The middle score is determined by counting up half the value of N if the number of observation (N) is even. When the number of observations (N) is odd, the mid-observation value is median. For example, 10 is the median of scores: 7, 8, 9, 10, 11, 12, 13. When the number of scores (N) is even, the median is the mid-point between the two middle scores. For example:

$$= \frac{10+11}{2} = 10.5 \text{ is the median of score}$$

7,8,9,10,11,12,13,14

b) ***Median for grouped data***

In the case of grouped data, cumulative frequency distribution is prepared and the median is calculated by the formula:

$$\text{Mdn} = I + \frac{N/2 - F}{F} \times i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Mdn = median

l = Exact lower limit of the class-interval upon which the median lies

N/2 = One half of the total number of observations

F = Sum of all frequencies below 1

f = Frequency within the class-interval in which the median lies

i = Width of the class interval in which the median lies

To illustrate the use of formula (3) consider the data of Table 2.5 once again.

Table 2.7: The Calculation of the Median from Data Grouped into a Frequency Distribution

Class-Interval	Frequency (F)	Cumulative Frequency (f)
34.5-39.5	4	40
29.5-34.5	8	36
Median class	24.5-29.5	11 ⟹ 28
19.5-24.5	8	17
14.5-19.5	6	9
9.5-14.5	3	3
	N=40	

Here N/2 = 20, l = 24.5, F = 17, f = 11 and i = 5

Using formula (3)

$$\text{Mdn} = I + \frac{N/2 - F}{F} \times i$$

$$= 24.5 + \frac{20-17}{11} = 5$$

$$\frac{24.5}{11}$$

$$= 25.86$$

III) ***The Mode***

The mode is defined as the most frequently occurring measure of an observation in a distribution. If only one value occurs a maximum number of times the distribution is said to have one mode; i.e. the distribution is unimodal. In some distributions there may be more than one mode. A two mode distribution is bimodal and it ismultimodal in a distribution, which has more than two modes.

a) *Mode for Ungrouped Data*

In a simple ungrouped series of measures, the crude or empirical mode is that single measure which occurs most frequently. For example, in the series 7, 8, 9, 9, 10, 11, and 12 the most often recurring measure, namely, 9 is the crude or empirical mode.

b) *Mode for Grouped Data*

When data are grouped into a frequency distribution, the crude or empirical mode is usually taken to be the mid-point of that interval which contains the largest frequency. In the example given in Table 2.5, the interval 24-29 contains the largest frequency and hence 26.5, its mid-point, is the crude mode. The true mode, that is, the point of greatest concentration in the distribution, or the point at which more measures fall than at any other point, is calculated by the formula:

$$\text{Mode} = I + \frac{Fm - f1}{2\,fm - f_1 - f_2} \times i \qquad (4)$$

Where

l	=	Lower limit of the modal class i.e., the class-interval having maximum frequency
fm	=	Frequency of the modal class.
f 1	=	Frequency of the class-interval preceding the modal class.
f 2	=	Frequency of the class-interval following the modal class.
i	=	Width of the modal class.

To illustrate, let us make use of formula (4) for the data in Table 2.5. Here the maximum frequency is 11 which lies in class interval 24.5-29.5.

Therefore, the modal class is (24.5 – 29.5). Here *fm* = 11, *f1* = 8, *f2* = 8, *i* = 5 and *l* = 24.5.

Using Formula (4) :

$$\text{Mode} = I + \frac{Fm - f_1}{2\,fm - f_1 - f_2} \times i$$

$$= 24.5 + \frac{11-8}{2x11-8-8} \times 5$$

$$= 24.5 + \frac{3}{6} \times 5$$

$$= 24.5 + 2.5$$

$$= 27.00$$

B) **Measures of Variability**

The measures of central tendency are very useful in describing the nature of a distribution of measures, but they do not give the researcher a complete picture of the data. These measures will not tell the researcher how the scores tend to be distributed. For this, we use a different set of measures which are called measures of 'variability' or measures of 'spread' or 'dispersion'. The most commonly used measures of variability include the range, the variance and standard deviation.

I) ***The Range***

The range is defined as the difference between the two extreme measures or values in a distribution. Suppose the scores of 10 learners are:

50, 40, 39, 35, 29, 28, 24, 27, 19, 18.

The range of this distribution will be (50 – 18) = 32. Although the range has the advantage of being easily calculated, it has the following serious limitations:

1) As the value of range is based on only two extreme values in the total distribution, it does not give any idea of the variation of many other values of the distribution.

2) It is not a stable statistic as its value can differ from sample to sample drawn from the same population.

II*) The Variance and Standard Deviation*

The average of the squared deviations of the measures or values from their mean is known as the variance. The standard deviation is the positive square root of variance.

a) *The Variance and Standard Deviation for the Ungrouped Data*

The variance for the ungrouped data is found by using the formula:

$$\sigma^2 = \frac{\Sigma x^2}{N} \qquad \text{.....(5)}$$

Where

2 = Variance of the sample

x = Deviation of the raw measures or values from the mean

N = Number of values or measures

Let us consider the following data of scores for the application of formula (5):

10, 10, 9, 9, 8, 8, 7, 7, 6, 6.

As the deviation of each score from the mean is required, the first thing to do is to calculate the mean. Using the formula (1)

$$M = \frac{\Sigma x}{N} = \frac{80}{10} = 8$$

Now, from each raw score, the mean is subtracted to get the value of x.

Table 14.8 : Distribution of the Test Scores of Ten Learners

Score (X)	Deviation (X-M) (X)	Deviation Squared (X^2)
10	2	4
10	2	4
9	1	1
9	1	1
8	0	0
8	0	0
7	-1	1
7	-1	1
6	-2	4
6	-2	4
	ΣX^2 =	20

Using formula (5)

$$\sigma^2 = \frac{\Sigma x^2}{N}$$

$$\frac{20}{2}$$

Now to get the standard deviation, we need the positive square root of the variance,

σ^2

Standard Deviation $\sigma = \sqrt{\frac{\Sigma x^2}{N}}$

$= \sqrt{2}$

$= 1.41$

The raw scores instead of deviation scores may also be used. The raw score formula for variance and standard deviation are given as follows.

Variance $= \sigma^2 = \frac{N\Sigma x^2 - (x)^2}{N^2}$ ..... (6)

Standard Deviation $= \sqrt{\frac{N\Sigma x^2 - (\Sigma x)^2}{N^2}}$ (7)

in which

X = Raw score

N = The number of scores in the distribution 31

Data Processing and Analysis: Using the same set of data, we can calculate variance and standard deviation with the help of raw-score formulae:

Table 14.9: The calculation of variance and standard deviation from original (raw) scores when the assumed mean is taken at zero and the data is ungrouped.

Score (X)	X^2
10	100
10	100
9	81
9	81
8	64
8	64
7	49
7	49
6	36
6	36
$\Sigma X = 80$	$\Sigma X^2 = 660$

Using formula (6) $N\Sigma X^2 - d\Sigma Xi2\ N\ 2$

Variance $= \sigma^2 =$

$10 \times 660 - (80)^2$

$= 100$

$6600 - 6400$

$= 200$

100

1

00

= 2

sing formula (7)

Standard Deviation = s

10×660−(80)2

N

=

10

= 6600 − 6400

10

14

= 10 .14

= 1.414

32

c) *Variance and Standard Deviation for Grouped Data* **Descriptive Statistics**

In the case of grouped data in a frequency distribution, the variance and standard deviation are calculated by using the formula:

$$\text{Variance} = \frac{i^2}{N^2}\left[N \sum fx'^2 - (\sum fx')^2\right] \quad \ldots\ldots\ldots (8)$$

$$\text{Standard Deviation} = \frac{i}{N}\sqrt{N \sum fx'^2 - (\sum fx')^2} \quad \ldots\ldots\ldots\ldots\ldots (9)$$

Where

i=Width of the class-interval N =Total number of measures
f=Frequency of class-interval
x' =Deviation of the raw measure from the assumed mean divided by the length of class-interval.
To illustrate the use of these formula let us consider the distribution given in Table 2.10

Table 14.10 : The Calculation of Variance and Standard Deviation from Data Grouped in a FrequencyDistribution

Class-Interval	f	x'	fx'	fx'^2

71-75	3	3	9	27
66-70	8	4	2	8	16
61-65	3	9	1	9	9
56-60	8	15	0	0	0
51-55	3	8	–1	–8	8
46-50	8	6	–2	–12	24
41-45	3	5	–3	–15	45
		N = 50	$\sum fx'$ = □ □9		$\sum fx'^2$ 129

Using formula (8)

$$\text{Variance} = \frac{1}{N^2}\left[N\sum fx'^2 - (\sum fx')^2\right]$$

$$= \frac{1}{(50)^2}\left[50 \times 129 - (-9)^2\right]$$

$$= \sqrt{\frac{1}{(50)^2}}$$

= 63.69

Using Formula (9)

$$\text{Standard Deviation} = i\sqrt{\frac{N\Sigma fx^2 - (\Sigma fx')^2}{N}}$$

$$= \frac{5}{10}\sqrt{50 \times 129 - (-9)^2}$$

$= \frac{1}{10\sqrt{6369}}$

$= \frac{1}{10 \times 79.81}$

$=7.98$

The standard deviation is a very useful device for comparing characteristics that may be different or expressed in different units of measurement. It is also used in describing that status or position of an individual in a group.

Check Your Progress 2

Compute (i) Mean (ii) Variance and (iii) Standard Deviation for the following frequency distribution:

Class Interval	F
195-199	1
190-194	2
185-189	4
180-184	5
175-179	8
170-174	10
165-169	6
160-164	4
155-159	4
150-154	2
145-149	3
140-144	1

Note: a) Space is given below to write your answer.

b) Compare your answer with the one given at the end of this unit.

……

……

……

……

……

C) Measures of Relationship

The data in which we secure measures of two variables for each individual is called bivariate data. The essential feature of bivariate data is that one measure can be compared with another measure for each member of the group. When bivariate data are studied, we may like to know the degree of relationship between the variables of such data. This degree of relationship is known as correlation. It can be quantitatively represented by the co-efficient of correlation. Its value ranges from – 1.00 to + 1.00. A value of – 1.00 describes a perfect negative correlation and + 1.00 describes perfect positive correlation. A zero value describes complete lack of correlation between the two variables. The sign of the co-efficient indicates the direction of relationship and numerical value is its strength/magnitude.

Methods of correlating variables

There are various methods of correlating variables. Their use is relative to the situation and type of data. Product moment correlation and Rank order correlation are mostly used for computing correlation between two variables.

1) *Product-moment correlation*

In some situations, the data for two variables X and Y are expressed in interval or ratio level of measurement and the distribution of these variables have a linear relationship. Moreover, the distribution of variables are uni-modal and their variances are approximately equal. In such situations, product moment method of correlation is used generally. It is also called Pearson's r. i) Calculation of Pearson's r from ungrouped data When the size of sample is small, there is no need of grouping the data and Pearson's r may be calculated with the help of the following formula:

in which

x = deviations of X measures from the assumed mean.

y = deviations of Y measures from the assumed mean.

To illustrate the use of formula (10), let us compute the product moment 'r' from the following data for the two variables X and Y for 10 learners who are enrolled in a Study Centre of an Open University.

X: 45 54 52 58 62 46 55 49 50 54

Y: 42 50 55 46 59 41 46 48 45 48

Using formula (10) for the data in Table 2.11

The distribution of Course X scores for the 50 learners is found in the f(y) column on the right of the scattergram. Assume a mean for the distribution of scores of course X (the mid-point of that interval which contains the largest frequency), and draw double lines to mark off the row in which the assumed mean falls. In the present example, the mean score or course X has been

taken at 46 (mid point of internal 45-47) and y's (deviations from the assumed mean) have been taken from this point. Fill in *fy*and then *fy*2 columns

Step 2

The distribution of the Course Y of 50 learners is found in the f (x) row at the bottom of the scattergram. Assume a mean for this distribution and draw double lines to designate the column under the assumed mean. The mean for the Course Y scores is taken at 26.5 (mid-point of interval 26-27), and the x's (deviations from assumed mean) are taken from this point. Fill in the fx and fx2 rows.

The *fxy* for a cell is computed by multiplying the frequency given in the particular cell with the corresponding x and y. For example, there is a frequency 1 corresponding with Course Y score 24-25 and Course X score 51-53. The corresponding x for this cell frequency is –1 and corresponding y is +2. Thus *fxy* for this cell is (1) (-1) (+2) = -2. Similarly the value for *fxy* is computed for all the cells and their sum Σ fxy is calculated row-wise as well as column-wise. The two sums should equal each other. In the present example, it has come to be 4.

14.4 LET US SUM UP

In this Unit, we discussed the various methods of representing the quantitative data, and the methods used in the analysis of quantitative data. The main points are as follows:

1) The quantitative data is tabulated in 'frequency distribution' and can be represented graphically with the help of a histogram, a frequency polygon, and/or an ogive.

2) Measures of (i) central tendency, (ii) variability, and (iii) relationship are the three types of descriptive statistical measures.

3) Mean, median and mode are the three measures of central tendency.

4) Mean is the arithmetic average of a distribution. It is obtained by dividing the sum of all values of observation by the total number of values. When the number of observation is large, the data is grouped in a frequency distribution. The formula for computing the mean here is:

   M= *AM*

   fx

   N

   □□□*i*

   Σ

5) Median is a point in an array, above and below which one half of the

values or measures fall. If the values are ungrouped and their number is small, the values are arranged in order of magnitude and the middle value is determined by counting up half, the value of N. When the number of values is odd, the midvalue is the median. When the number of values is even, the median is the mid-point between the two middle values. In the case of grouped data, the median is calculated by the formula:

Mdn = *l*

N

F

f

□□*i*

–

2 □

6) Mode is the most frequently occurring value in a distribution. If only one value occurs a maximum number of times, the distribution is said to have one mode (uni-modal). A two mode distribution is bi-modal, and more than a two mode distribution is called multi-modal.

In a simple ungrouped series of measures or values, the crude mode is that single measure or value which occurs most frequently. For a group distribution, the mode is calculated by the formula:

Mode = *l*

fm fi

fm f f

□□*i*

–

–□–

□

2 1 2

7) The range is the difference between the two extreme values or measures in a distribution.

8) The average of the squared deviations of the measures or values from their mean is known as variance. Standard deviation is the positive square root of variance.

UNIT 15 INFERENTIAL STATISTICS

Structure

15.0 Introduction

15.1 Objectives

15.2 Classification of Statistical Tests

15.3 Parametric Tests

15.3.1 Sampling Distribution of Means

15.3.2 Application of Parametric Tests

15.4 Non-parametric Tests

15.5 Let Us Sum Up

15.6 Glossary

15.7 Check Your Progress: The Key

15.8 Appendix

15.0 INTRODUCTION

In Unit 2, we focused on descriptive statistics including the various measures of central tendency, variability, relative positions and relationships. These measures are used to describe the properties of particular samples. In this Unit, we shall introduce inferential statistics. The knowledge of these statistics is useful for testing the hypothesis related to your research problems, and to make generalisations about populations on the basis of data analysis. This requires you to be familiar with certain statistical tests – parametric and non-parametric.

15.1 OBJECTIVES

This Unit aims to provide you with detailed information about the nature and use of parametric and non-parametric tests in general and the application of some of these tests for drawing inferences and generalisations in particular. On the completion of this Unit, you should be able to:

- classify various statistical tests;
- describe the nature of parametric tests along with the assumptions on which they are based;
- work out sampling distribution of means in the context of large samples, and small samples;
- define and illustrate the concept of confidence intervals and levels of significance;
- define and illustrate the concept of degrees of freedom;
- use Z-test and t-test in testing the significance of the difference between

means;

- describe the nature and uses of analysis of variance;
- describe the nature of the non-parametric tests along with their assumptions;
- use of Chi-square test; and
- describe the use of median test and its application.

15.2 CLASSIFICATION OF STATISTICAL TESTS

The descriptive statistics already discussed in Unit 2 are used to explain the properties of samples drawn from a population. The researcher computes certain 'statistics' (sample values) as the basis for inferring the corresponding 'parameters' (population values). Ordinarily, a single sample is drawn from a given population so as to determine how well a researcher can infer or estimate the 'parameter' from a computed sample 'statistics'. For making the inferences about the various parameters, the researcher makes use of parametric and non-parametric tests.

15.3 PARAMETRIC TESTS

Under this section we discuss two sub-themes: sampling distribution of means and application of parametric tests. Sampling distribution of means covers: a) large samples, b) confidence intervals and levels of significance, c) small samples, and d) degree of freedom. Application of parametric tests covers three tests, namely Z-test, t-test, and F-test.

Parametric tests are the most powerful statistical tests for testing the significance of the computed sampling statistics. These tests are based on the following assumptions:

1) the variables described are expressed in interval or ratio scales and not in nominal or ordinal scales of measurement,

2) the population values are normally distributed,

3) the samples have equal or nearly equal variances – this condition is known as 'equality or homogeneity of variances' and is particularly important to determine for small samples,

4) the selection of one case in the sample is not dependent upon the selection of any other.

Z-test, t-test and F-test are the most commonly used parametric tests. Before discussing the application of these tests, it is necessary to describe certain concepts relating to 'sampling distribution of means', 'confidence intervals', 'levels of significance', and 'degree of freedom'.

Check Your Progress 1

Describe the assumptions on which the use of parametric tests is based.

Note: a) Space is given below for writing your answer.

b) Compare your answer with the one given at the end of the unit.

……

……

……

……

……

15.3.1 Sampling Distribution of Means

A. Large samples

An important principle, known as the 'central limit theorem', describes the characteristics of sample means. If a large number of equal-sized samples (greater than 30) are selected at random from an infinite population;

- the distribution of 'sample means' is normal and it possesses all the characteristics of a normal distribution,
- the average value of 'sample means' will be same as the mean of the population,
- the distribution of the sample means around the population mean will have its own standard deviation, known as 'standard error of mean, which is denoted as SEM or □□M . It is computed by the formula:

SEM □ σm = σ

√n ..... (1)

In which σ = Standard deviation of the population and

N = The number of cases in the sample.

Since the value of □□(i.e. standard deviation of population) is usually not known, we make an estimate of this standard error of mean by the formula:

σm = σ

√n ..... (2)

in which σ □= Standard deviation of the sample

N = The number of cases in the sample.

To illustrate the use of formula (2), we assume that the mean of the attitude

scores of a sample of 100 distance learners enrolled with IGNOU towards student support services is 25 and the standard deviation is 5. The standard error of mean can be calculated accordingly:

$$\text{SEM} \quad \sigma m = \frac{5}{\sqrt{100}} = 0.50$$

This standard 'error of mean' may be assumed as the standard deviation of a distribution of sample means, around the fixed population mean of all distance learners. In the case of large randomly selected samples, the sampling distribution of sample means is assumed to be normal.

Fig. 3.1: Sampling distribution of means showing variability of obtained means around the population mean in terms of σM

The normal curve in Figure 1 shows that this sampling distribution is centered around the unknown population mean with standard deviation 0.50. The sample means often fall between the positive and the negative side of the population mean. About 2/3 of our sample means (exactly 68.26 per cent) will lie within ± 1.00 σ□M of the population mean, i.e., within a range of ± 1 □□0.5 = ± 0.50. Furthermore, 95 of our

100 sample means will lie within ± 2.00 σ□M (more exactly ± 1.96 σ□M) of the population mean i.e., 95 of 100 sample means will lie within ± 1.96 □□0.50 or ± 0.98 of the population mean. In other words, the probability that our sample mean of 25 does not miss the population means (M pop.) by more than ± 0.98 is 0.95. also, 99 of our samples means will be within ± 3.00 σ□M (More exactly ± 2.58 σ□M) of the population mean. This indicates that 99 out of 100 sample means fall within ± 2.58 □□0.50 or ± 1.29 of the population mean. The probability (P) that our sample mean of 25 does not miss the M pop. by more than ± 1.29 is .99.

Thus, the value of a population mean, to be inferred from a randomly selected sample mean, can be estimated on a probability basis.

Check Your Progress 2

Given a sample of 100 distance learners with mean = 175.50 and S.D. = 5.82 of the scores of an intelligence test. Compute the standard error of mean.

Note:a)Space is given below for writing your answer.

b) Compare your answer with the one given at the end of the Unit.

……

……

……

……

……

B. Confidence intervals and levels of significance Inferential Statistics

When we draw a large random sample from the population to obtain measures of a variable and compute the mean for the sample, we can use the 'central limit theorem' and 'normal probability curve' to have an estimate of the population mean. We can say that M has a 95 per cent chance of being within 1.96 standard error units of M pop. In other words, a mean for a random sample has a chance of 95 per cent of

being within 1.96 σ□M units from M pop. It may also be said that there is a 99 per cent chance that the sample mean lies within 2.58 σ□□M units of M pop. To be more specific, it may be stated that there is a 95 per cent probability that the limits M ± 1.96 σ M enclose the population mean, and the limits M ± 2.58 σ□□M enclose the population mean with 99 per cent probability. Such limits enclosing the population mean are known as the 'confidence intervals'.

These limits help us to adopt particularly two levels of confidence. One is known as 5 per cent level or 0.05 level, and other is known as 1 per cent level or 0.01 level. The 0.05 level of confidence indicates that the probability M pop that lies within the interval M ± 1.96 σ□□M is 0.95 and that it falls outside of these limits is 0.05. By saying that probability is 0.99, it is meant that M pop. lies within the interval M ± 2.58 σ□□M and that the probability of its falling outside of these limits is .01.

To illustrate, let us apply the concept to the previous problem. Taking as our limits M ± 1.96 σ□□M, we have 25 ± 1.96 □□0.50 or a confidence interval marked off by the limits 24.02 and 25.98. Our confidence that this interval contains M pop.is expressed by a probability of 0.95. If we want a higher degree of confidence, we can take the 0.99 level of confidence for which the limits are M ± 2.58 σ□□M or a confidence interval given by the limits 23.71 and 26.29. We may be quite confident that M pop. is not lower than 23.71 nor higher than 26.29, i.e., the chances are 99 in 100 that the M pop. lies between 23.71 and 26.19.

C. Small samples

When the number of cases in the sample is less than 30, we may estimate the value σ□□M by the formula:

$$SEM \square \frac{S}{\sqrt{n}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

in which

S = Standard deviation of the small sample.

N = The number of cases in the sample.

The formula for computing S is

$$S = \sqrt{\frac{\Sigma x2}{N-1}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

in which

$\Sigma x2$ = Sum of the squares of deviations of individual scores from the sample mean.

N = The number of cases in the sample.

The concept of small size was developed by William Seely Gooset, a consulting statistician for Guinness Breweries of Dublin (Ireland) around 1915. The principle

is that we should not assume that the sampling distribution of means of small samples is normally distributed. He found that the distribution curves of small means were somewhat different from the normal curve. When the size of the sample is small, the t-distribution lies under the normal curve, but the tails or ends of the curve are higher than the corresponding parts of the normal curve. Figure 2 shows that the t-distribution does not differ significantly from the normal distribution unless the sample size is quite small. Further, as the sample size increases in size, the t-distribution approaches more and more closely to the normal curve.

For small samples, it is necessary to make use of selected points in the table of Gosset's t-critical values or Students's t-values, given in Appendix (Table II). As the sample size increases, the Student's t-values approach the Z-values of Normal Probability Table. When small samples are used, the use of t-values involves an important concept known as 'degrees of freedom', which we shall now discuss separately.

D. Degrees of freedom (df)

While finding the standard deviation of small samples we use N-1 in the denominator instead of N in the basic formula for standard deviation. The difference in the two formulae may seem very little, if N is sufficiently large. But there is a very important difference in the 'meaning' in the case of small samples. N-1 is known as the 'number of degrees of freedom', denoted by df. The 'number of degrees of freedom' in a distribution is the number of observations or values that are independent of each other and cannot be deduced from each other. In other words, we may say that the 'degrees of freedom' connote freedom to vary.

To illustrate as to why the df used here is N-1, we take 5 scores, i.e., 5, 6, 7, 8, and 9, the mean of which is 7. This mean score is to be used as an estimate of the population mean. The deviations of the scores from the mean 7 are –2, –1, 0, +1 and +2. A mathematical requirement of the mean is that the sum of these deviations should be zero. Of the five deviations, only 4, i.e., N-1 can be chosen freely (independently) as the condition that the sum is equal to zero restricts the value of the 5th deviate. With this condition, we can arbitrarily change any four of the five deviates and thereby fix the fifth. We could take the first four deviates as –2, –1, 0, and +1, which would mean that for the sum of deviates to be, zero, the fifth deviate has to be +2. Similarly, we can try other changes and if the sum is to remain zero, one of the five deviates is automatically determined. Hence, only 4, i.e., (5-1)'s are free to vary within the restrictions imposed. When a statistics is to be used to estimate a parameter, the number of degrees of freedom depends upon the restrictions imposed. One 'df' is lost for each of the restrictions imposed. Therefore, the number of 'df' varies from one statistics to another. For example, in estimating and computing the population mean (M pop) from the same sample Mean (M), we lose 1 'df.' So, the number of degrees of freedom is (N-1). Let us determine the 0.95 and 0.99 confidence intervals for the population mean (M pop) of the scores 10, 15, 10, 25, 30, 20, 25, 30, 20 and 15, obtained by 10 distance learners on an attitude scale. The mean of the scores is:

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

$$= \frac{200}{10} = 20.0$$

Using formula (4) we compute the standard deviation as:

X	x = X-M	X^2
10	-10	100
15	-5	25
10	-10	100
25	5	25
30	10	100
20	0	0
25	5	25
30	10	100
20	0	0
15	-5	25
	Σ x = 0	**Σ x2 = 500**

$$S=\sqrt{\frac{\Sigma x^2}{N-1}}$$

$$=\sqrt{\frac{500}{10-1}}$$

$$= 7.45$$

From Formula (3) we compute

$$SE_M = \frac{7.45}{\sqrt{10}}$$

$$= 2.36$$

For estimating the M pop from the sample mean of 20.00, we determine the value of t at the selected points using appropriate number of degrees of freedom. The available 'df' for determining t is N-1 or 9. Entering Table II (See Appendix) with 9 'df', we read that t = 2.26 at 0.05 level and 3.25 at 0.01 level. From the first t-value we know that 95 of our 100 sample means will lie within ± 2.26 *SE*M or ± 2.26 □□2.36 of the population mean and 5 out of 100 fall outside of these limits. The probability (P) that our sample mean 20.00 does not miss the M pop by more than ± 2.26 × 2.36 or ± 5.33 is 0.95. From the second t-value, we know that 99 per cent of our sample mean will lie between M pop and 3.25 *SE* M or ± 3.25 □□2.36, and that 1 per cent fall will beyond these limits. So, the probability (P) that our sample mean of 20.00 does not miss the M pop by more than ± 3.25 □□2.36 or ± 7.67 is 0.99.

Taking our limits as M ± 2.26 *SE* M , we have 20.00 ± 2.26 × 2.36 or 14.67 and 25.33 as the limits of the 0.95 confidence interval. The probability (P) that M pop is not less than 14.67 nor greater than 25.33 is 0.95. Taking the limits M ± 3.25 *SE*M, we have 20.00 ± 3.25 × 2.36, or 12.33 and 27.67 as the limits of 0.99 confidence interval, and the probability (P) so that M pop is not less than 12.33 and not greater than 27.67 is 0.99.

The use of small samples to build generalisation in social research should be made cautiously as it is difficult to ensure that a small sample adequately represents the population from which the sample is drawn. Furthermore, conclusions drawn from small samples are usually unsatisfactory because of the great variability from sample to sample. In other words, large samples drawn randomly from the population will provide a more accurate basis than will small samples for inferring population parameters.

3.2.2 Applications of Parametric Tests

In this subsection, we shall discuss the application of parametric tests, namely

Z-test, t-test and F-test.

A. Application of Z-test for testing the significance of difference between means of two independent large samples

It has already been explained in earlier sections that the frequency distribution of large sample means drawn from the same population fall into a normal distribution around M pop as their measure of central tendency. It is also reasonable to expect that the frequency distribution of the difference between the means computed from the two samples will also tend to be normal with a mean of zero and standard deviation of 1. It is termed as standard error of the difference between two means and is denoted by dm. It is computed by the formula:

$$\sigma dm = \sqrt{\sigma m + \sigma m} \quad \text{.....(5)}$$

in which

M 1 = the SE of the mean of the first sample

σ M 2 = the SE of the mean of the second sample

To illustrate, we apply formula (5) to a problem. Suppose a test of creativity was administered on two groups, one of 120 males and the other of 75 females enrolled in Post Graduate Diploma in Rural Development course of IGNOU.

The results are summarized in Table 3.1 below:

Table 3.1: Means and Standard Deviation of Two Independent Large Samples

Statistics	Boys	Girls
N	120	75
Mean (M)	57.50	55.75
Standard Deviation		
σ	8.42	8.13

Assuming that our samples are random, it is to be ascertained whether the difference between the means 57.50 and 55.75 is significant.

Using formula (5) we compute the standard error of the difference between means:

$$\sigma dm = \sqrt{\frac{(8.42)^2}{120} + \frac{(8.13)^2}{75}}$$

$$= 1.21$$

The obtained difference between the means of males and females is 1.75 (i.e.,

57.50 – 55.75); and the SE of this difference (σ□□d M) = 1.21. To determine whether two groups of males and females actually differ in creative thinking ability, we set-up a null hypothesis, i.e., the difference between population means of males and females is zero and that, except for sampling errors, means of males and females is zero and that, except for sampling errors, mean differences from sample will also be zero. In accordance with a null hypothesis, we assume a sampling distribution of differences to be normal with the mean at zero, (or at M pop (males) – M pop **(f**emale) = 0). The deviation of each sample difference, [Mmales – Mfemales] – [M pop(males) – M pop(females) or [Mmales – Mfemales] = 0. The deviation of each sampled difference between two means, given in terms of standard measure, would be the deviation divided by the standard error, which gives us a *z* value in terms of general formula:

$$z = \frac{M_1 - M_2}{d_M} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

Using formula (6)

$$z = \frac{1.75}{1.21} = 1.45$$

For the sake of convenience, we use .05 and .01 levels of significance as two arbitrary standards for accepting or rejecting a hypothesis. From the normal distribution Table 1 (See Appendix) we read that ± 1.96 σ□mark off points along the base line of the normal curve to the left and right of which lie 5 per cent of the cases (i.e., 2.5 per cent at each end of the curve). When a *z* value is 1.96 or more, we reject a null hypothesis at .05 level of significance. The computed *z* value of 1.45 in our problem falls short of 1.96, i.e.; it does not reach the .05 level. Accordingly, we retain the null hypothesis and conclude those two groups of males and females actually do not differ in their mean performance on creative-thinking-test.

Furthermore, from Table I (See Appendix) we know that ± 2.58 σ□□mark off points to the left and right of which lie 1 per cent (0.5 per cent at each end of the curve) of the cases in the normal distribution. If the *z* value is 2.58 or more, we reject the null hypothesis at .01 level and the probability (P) is that not more than once in 100 trials would a difference of this size arise if the true difference (M pop1 – M pop2) was zero.

B. Two-tailed and one-tailed tests of significance

Suppose a null-hypothesis were set up that there was no difference, other than a sampling error difference, between the mean height of two groups, A and B. We would be concerned only with the difference and not with the superiority or inferiority in height of either group. To test this hypothesis, we

apply two-tailed test as the difference between the obtained means of height of two groups may be as often in one direction (plus) as in the other (minus) from the true difference of zero. Moreover, for determining probability, we take both tails of sampling distribution. For a large sample two-tailed test, we make use of a normal distribution curve. The 5 percent area of rejection is divided equally between the upper and lower tails of this curve and we have to go out to ± 1.96 on the base line of the curve to reach the area of rejection as shown in the Figure 3.3.

Similarly, if we have 0.5 per cent area at each end of the normal curve where one percent area of rejection is to be divided equally between its upper and lower tails, it is necessary to go out to ± 2.58 on the baseline to reach the area of rejection as shown in the Figure 3.4.

Fig. 3.4: A Two-tailed Test at 0.01 level (0.5 percent at each level)

In the case of the above example a null hypothesis was set up that there was no difference other than a sampling error difference between the mean creative thinking score of boys and girls. Thus, we were concerned with a difference, and not in superiority or inferiority of either group in the creative thinking ability. To test this hypothesis, we applied 'two-tailed test' as the difference between the two means might have been in one direction (plus) or in other (minus) from the true difference of zero; and we took both tails of sampling distribution in determining probabilities. As is evident from the above example, we make use of a normal distribution curve in the case of a large sample 'two-tailed test'. The five per cent area of rejection is equally divided between the upper and lower tails of the curve and we have to go out to ± 1.96 on the baseline of the curve to reach the area of rejection. Similarly, we have 0.5 per cent area at each end of the normal curve when 1 per cent of rejection is to be divided equally between its upper and lower tails and it is necessary to go out to ± 2.58 on the baseline to reach the area of rejection. In the above problem, if we change the null hypothesis as: boys have significantly higher creative thinking than that of the girls; or boys have significantly lower creative thinking than the girls, than each of these hypotheses indicates a direction of difference. In such situations, the use of 'one-tailed test' is made. For such a test, five per cent area or one per cent area of rejection is either at the upper tail or at the lower tail of the curve, to be read from 0.10 column (instead of 0.05) and 0.02 column (instead of 0.01)

C. Application of t-test for testing the significance of difference between two independent small samples

We have already discussed that the frequency distribution of small sample means drawn from the same population forms a t-distribution and it is reasonable to expect that the sampling distribution of the difference between the means computed from two different populations will also fall under the category of t-distribution. Fisher provided the formula for testing the difference between the means computed fromindependent small samples.

In which

M_1 and M_2	=	Means of two samples
$\Sigma x_1 2$ and $\Sigma x_2 2$	=	sums of squares of the deviations from the means in the two samples
N_1 and N_2	=	number of cares of the two samples
df	=	degrees of freedom = N1 + N2 – 2

To illustrate the use of the formula, let us test the significance of the difference between mean scores of 7 boys and 10 girls on an intelligence test.

Table 3.2: Scores of 7 boys and 10 girls on an intelligence test

N_1=7			N_2=10		
Boys x_1	$X1^2$		Girls X2		$X2^2$
Y_1				Y_2	
13	0	0	10	-4	16
14	1	1	16	2	4
11	-2	4	12	-2	4
12	-1	1	13	-1	1
15	2	4	18	4	16
13	0	0	13	-1	1
13	0	0	19	5	25
$\Sigma X 1$ □□91 $x1$		$\Sigma 2$ □□10	14	0	0
			13	-1	1
			12	-2	4
Σ □□140 $x2$		$\Sigma 2$ □□72			

$M1 = \frac{91}{7} = 13$ $\quad M_2 = \frac{140}{10} = 14$

df $= N_1 + N_2 - 2 = 7 + 10 - 2 = 15$

To test the significance of difference between the two means by making use of two-tailed test (null hypothesis, i.e., no differences between the two groups), we look for the t-critical values for rejection of null hypothesis in Table II (Appendix) for (7 + 10 – 2) or 15 df. These t-values are 2.13 at 0.05 and 2.95 at 0.01 levels of significance. Since the obtained t-value 0.33 is less than the table value necessary for the rejection of the null hypothesis at 0.05 level for df 15, the null hypothesis is accepted and it may be concluded that there is no significant difference in the mean intelligence scores of boys and girls.

If we change the null hypothesis as: boys will have higher intelligence scores than girls, or boys will have lower intelligence scores than girls, then each of these hypotheses indicates a direction of difference rather than simply the existence of the difference. So, we make use of one-tailed test. For given degrees of freedom, i.e., 15, the 0.05 level is read from the 0.10 column (p/2 = 0.05) and the 0.01 level from 0.02 column (p/2 = 0.01) of the t-table. In the one-tailed test, for 15 df tcritical values at 0.05 and 0.01 levels, as read from the 0.10 and the 0.02 columns of Table II (Appendix) are 1.75 and 2.60 respectively. Since the computed t-value of 0.33 does not reach the table value at 0.05 level (i.e., 1.75 for 0.10), we may conclude that the difference in two groups is there merely because of chance factors.

Check Your Progress 3

The following scores were obtained on an Interest test for five boys and eight girls:

Boys: 20, 22, 30, 32, 26.

Girls: 34, 25, 16, 30, 22, 27, 20, 26.

Is the difference between the Mean Interest Scores of the boys and girls significant?

Note: a) Space is given below for writing your answer.

b) Compare your answer with the given at the end of the unit.

……

……

……

……

……

D. Application of F-test for testing the significance of difference between independent means

The use of and t-test is made by a researcher to determine whether there is any significant difference between the means of two random samples. Suppose we have seven randomly drawn samples from a population and we want to determine whether there are any significant differences among their means. This will require

computation of $\frac{7\,(7-1)}{}$ = 21 t-tests to determine the significance of difference between the seven means by taking two means at a time. This procedure is time consuming as well as cumbersome. The technique of analysing of variance is applied to determine if any two of the seven means differ significantly from each other by a single test, known as F-test. The F-test makes it possible to determine whether the sample mean differ from the another (between group variance) to a greater extent than the test scores differ from their own samples means (within group variance). Using the ratio:

$$F = \frac{\text{Variance between the groups}}{\text{Variance within groups}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (8)$$

The values of F-ratio are given in the Appendix (Table III). This table indicates the F-critical values necessary for rejecting the null hypothesis at selected levels of significance, usually 0.05 and 0.01 levels.

E. Factor Analysis

Factor Analysis is a procedure for determining the number and nature of constructs that underlie a set of measures (Wiersman, 1986). Construct is a trait or an attribute that explains some phenomena, e.g., anxiety, intelligence, motivation, attitude etc. In factor analysis artificial variables are generated and called factors. Factor analysis is initiated from the correlation matrix of the variables. The variables which are highly correlated are grouped together.

Suppose we have scores of 100 students on 10 different tests. Question here is – How many different traits or constructs these 10 tests measure. The possibility can be that three or four tests measure the same trait or one test may measure two or more traits. The researcher can determine the correlation co-efficient among the 10 different test. High correlation's between test scores indicate that common constructs are measured. Low or zero correlation's indicate the absence of common constructs.

There are few terms used in factor analysis. If a test measures only one construct, it is labeled as *factorially pure.* A *factorially complex* test is one that measures two or more factors.

Factor loading is the extent to which a test measures a factor. Factor loading is very important in factor analysis because factors which are artificial variables generated from the data must be described and integrated. It is a

correlation coefficient between a test and a factor.

Uses of Factor Analysis in Research

In conducting research, the aim of using factor analysis is to identify the nature and number of constructs that underlie a set of variables. Factor analysis is associated with construct related evidence when validity. Factor analysis is used in confirmatory analysis and exploratory analysis.

Confirmatory factor analysis is used in studies where hypothesized constructs measured by a set of variables are either confirmed or refuted. It is also used to analyze a single test by factor analyzing the item scores.

Suppose 50 test item measures three traits, a confirmatory analysis of the item scores would support or refute this proposition.

On the other side, in the explanatory analysis the number of variables is reduced to a manageable number of explanatory purposes. A set of measures can be factor analyzed to enhance the explanation of what is measured in a more parsimomous manner.

For example, a group of teachers were observed and measured on 2 different competencies. A factor analysis of the competency scores undoubtedly would generate a smaller number of factors, say three or four, that represents the constructs underlying the performance of teacher.

Thus factor analysis in any research analysis provide valuable insights into the nature of phenomena.

Basic Assumptions for the Analysis of Variance

Certain basic assumptions underlying the technique of analysing of variance are:

1) The population distribution should be normal. This assumption is, however, not so important. The study of Norton (Guilford, 1965a; pp. 300-301) also points out that F is rather insensitive to variations in the shape of population distribution.

2) All groups of a certain criterion or of the combination of more than one criterion should be randomly chosen from the sub-population having the same criterion or having the same combination of more than one criterion. For example, if we wish to select two groups from a study centre, one belonging to a rural area and the other to the urban area, we must, choose the groups randomly from the respective sub-populations.

3) The sub-groups under investigation must have the same variability. In other words, there should be homogeneity of variance. It is tested either by applying Bartlett's test of homogeneity or by applying Hartley's test.

To illustrate the use of F-test, let us consider an example of twenty distance learners who have been randomly assigned to 4 groups of 5 each, to be taught

by different methods, i.e., A, B, C, and D. Their performance scores on an achievement test, administered after the completion of experiment are given in Table 3.3

Table 3.3: Achievement Test Scores of the Four Groups Taught through Four Different Methods

Methods of group	A (X_1)	B (X_2)	C (X_3)	D (X_4)
	14	19	12	17
	15	20	16	17
11	19	16	14	
10	16	15	12	
12	16	12	17	
Σx 300	62	90	71	77
Σx^2 4652	786	1632	1025	1207

We may compute the analysis of variance using the following steps:

15.4 NON-PARAMETRIC TESTS AND APPLICATION OF CHI- SQUARE TEST

In the preceding section, we discussed some important parametric tests involving the assumptions based upon the nature of the population distribution. There are some tests which do not make numerous or stringent assumptions about the nature of the population distribution. These tests are known as distribution-free or nonparametric tests. The non-parametric tests are based upon the following assumptions:

1) The nature of the population, from which samples are drawn, is not known to be normal.

2) The variables are expressed in nominal form, that is, classified in categories and represented by frequency counts.

3) The variables are expressed in ordinal form, that is, ranked in order or

expressed in numerical scores which have the strength of ranks.

4) The sample sizes are small.

The most frequently used non-parametric tests are: Chi-square test, the median test, the sign test, the Mann-Whitney U test, the Kolmogorov-Smirnov Two Sample Test, the Wilcoxon-Pairs Signed-Ranks Test, the McNemar Test of Significance of changes, contingency co-efficient, etc. For the present unit, we will discuss the applications of Chi-square test and median test only.

Check Your Progress 4

Describe the assumptions on which non-parametric tests are based.

Note:a) Space is given below for writing your answer.

b) Compare your answer with the one given at the end of the unit.

……

……

……

……

……

A. Application of Chi-Square Test

The Chi-square (pronounced as Ki-square) test is used with discrete data in the form of frequencies. It is a test of independence and is used to estimate the likelihood that some factor other than chance accounts for the observed relationship. Since the null hypothesis states that there is no relationship between the variables under study, the Chi-square test merely evaluates the probability that the observed relationship results from chance. The formula for Chi-square (*X*2) is:

$$X^2 = \frac{\Sigma (F0-Fe^2)}{Fe} \qquad \text{……..(9)}$$

In which

Fo = Frequency of the occurrence of observed or experimentally determined facts

Fe = expected frequency of occurrence

To test the significance of Chi-square, we enter Table IV of the Appendix with the computed value of Chi-square for the appropriate number of degrees of freedom. The number of degrees of freedom df = (r-1) (c-1), in which r is the number or rows and c is the number of columns in which the data are tabulated.

For example consider the following data (in Table 3.5) of 500 distance learners who have been categorised into three groups, elder, middle-aged, and younger on the basis of age and preference for four colours, red, blue, yellow, and green.

Table 3.5 : The Chi-square Test of Independent in Contingency Table

Colour ⇒ / Age Group ⇓	Red	Blue	Yellow	Green	Total
Elder	40 (38.42)	50 (45.22)	35 (39.10)	45 (47.26)	170
Middle	35 (36.16)	42 (42.56)	44 (36.80)	39 (44.48)	160
Younger	38 (38.42	41 (45.22)	36 (39.10)	55 (47.26)	170
Total	113	133	115	139	500

Across the first row of the table, we find that out of 170 distance learners in the older age-group, 40 have given their preference for red colour, 50 for blue, 35 for yellow, and 45 for green. Reading down the first column, we find that out of 113 distance learners giving preference for red colour, 40 belong to the older age group, 35 to middle and 38 to younger age group. The other columns and rows are interpreted in the same way.

The hypothesis to be tested is the null hypothesis, that is, age and colour preferences are essentially unrelated or independent. To compute Chi-square we must calculate an independent value, i.e., expected frequency for each cell in the contingency table. Independent values are represented by the figures in parentheses within the different cells. They give the number of students whom we should expect to fall in a particular age-group, showing their preference for a particular colour in the absence of any real association. The calculation of expected frequencies (fe) and Chi-square ($X2$) are shown as under:

1) *Calculation of expected frequencies (fe)*

Row-I: $\frac{113 \times 170}{500} = 38.42$; $\frac{133 \times 170}{500} = 45.22$;

$\frac{115 \times 170}{}$ $\frac{139 \times 170}{}$

$\frac{}{500}$ = 39.10; $\frac{}{500}$ = 47.26;

Row-II:

$\frac{113 \times 160}{500}$ = 36.16; $\frac{133 \times 160}{500}$ = 42.56;

$\frac{115 \times 160}{500}$ = 36.80; $\frac{139 \times 160}{500}$ = 44.48;

Row-III:

$\frac{113 \times 170}{500}$ = 38.42; $\frac{133 \times 170}{500}$ = 45.22;

$\frac{115 \times 170}{500}$ = 39.10; $\frac{139 \times 170}{500}$ = 47.26;

2) Computation of the Chi-square value

Using formula (9)

$$\frac{(40-38.42)^2}{38.42} + \frac{(50-45.22)^2}{45.22} + \frac{(35-39.10)^2}{39.10} + \frac{45-47.2}{47.10}$$

$$\frac{(35-36.16)^2}{36.16} + \frac{42-42.56}{42.56} + \frac{44-36.80}{36.80} + \frac{39-44.48}{44.48}$$

$$\frac{(38-38.42)^2}{} + \frac{41-45.22}{} + \frac{36-39.10}{} + \frac{55-47}{}$$

$$\frac{\quad}{38.42} + \frac{\quad}{45.22} + \frac{\quad}{39.10} + \frac{\quad}{47.2}$$

$X^2 = 5.182$

3) df = (r–1) (c–1)

= (3–1) (4–1)

= 8

The X2 critical values for 8 df as given in Table IV (See Appendix) are 15.507 and 20.090 respectively for 0.05 and 0.01 levels of significance and the obtained value, 5.182, is less than the table value even at 0.05 level. This indicates that there is no relationship between the age and the colour preference and thus the hypothesis that age and colour preference are essentially independent may be accepted at 0.05 level of significance. In the case of 2 × 2 contingency table, with (r-1) (c-1) = 1 df, there is no need of computing the expected frequencies (independence values) for each cell. The formula is:

$$X^2 = \frac{N\,(|AD - BC|)^2}{(A+B)\,(C+D)\,(A+C)\,(B+D)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (10)$$

In the above formula A, B, C, and D are the frequencies in the first, second, third and fourth cells respectively and the vertical lines in |AD – BC| mean that the difference is to be taken as positive.

To illustrate the use of formula (10), let us determine whether item 5 of an achievement test differentiates between high and low achievers. The responses to items are given in the following 2 □□2 contingency Table.

Table 3.6 : The Chi-square Test in 2 □□2 Fold Contingency Table

	Passed Item 5	**Failed Item 5**	**Total**
High	(A)	(B)	(A+B)

Achiever	115	35	150
	(C)	(D)	(C+D)
Low Achiever	140	90	130
Total	(A+C) 155	(B+D) 125	280

Using Formula (10)

$$X^2 = \frac{280\,(115 \times 90 - 35 \times 40)^2}{(115+35)\,(40+90)\,(115+40)\,35+90)}$$

$$= \frac{280\,(10350 - 1400)2}{(150)\,(130)\,(155)\,(125)} = 59.36$$

Since the computed X2 value of 59.36 exceeds the critical X2 value of 6.635 to be significant at 0.01 level, we reject the hypothesis that item 5 of the test, does not discriminate significantly between high and low achievers. In other words, it may be concluded that item 5 of the achievement test discriminates significantly between the two groups, namely high and low achieving students.

Further, when entries in 2×2 table are less than 10, Yate's correction for continuity is applied to formula (10). The corrected formula reads:

$$X^2_C = \frac{N\,(AD - (BC - N/2)}{(A+B)\,(C+D)\,(A+C)\,(B+D} \quad \ldots\ldots\ldots\ldots (11)$$

The following example illustrates the use of formula (11).

Fifteen male and twelve female counsellors of a study centre were asked to express their attitude towards population education. Both the groups of counsellors were administered the attitude scale and were classified as having either positive or negative attitude towards population education. The distribution of the sample is shown in table 3.7.

Table 3.7: Distribution of Male and Female Counsellors in Terms of

theirPositive or Negative Attitude towards Population Education

	Positive Attitude	**Negative Attitude**	**Total**
Female Counsellors	(A) 7	(B) 5	(A+B) 12
Male Counsellors	(C) 9	(D) 6	(C+D) 15
Total	(A+C) 16	(B+D) 11	27

$$X^2_C = \frac{27\ (7 \times 6 - 5 \times 9 - 27/2)^2}{(7+5)\ (9+6)\ (7+9)\ (5+6)} \quad \ldots\ldots\ldots\ldots (11)$$

$$= \frac{27\ (42 - 45 - 13.5)^2}{12 \times 15 \times 16 \times 11}$$

$$= 0.23$$

Since the obtained value of X2, 0.23, is less than the table value of 3.842 to be significant at 0.05 level of significance, it may be inferred that there is no true difference in the attitude of male and female counsellors towards population education.

B. Application of Median Test

he median test is used for testing whether two independent samples differ in central tendencies. It gives information as to whether it is likely that two independent samples have been drawn from populations with the same median. It is particularly useful whenever the measurements for two samples are expressed in an ordinal scale.

In the median test, we first compute the combined median for all rank measures in both samples. Then both sets of rank measures at the combined median are dichotomised and the data are set in a 2 × 2 Table as shown in Table 3.8.

Table 3.8: 2 □□2 Table for use of the Median Test

	Group-I	Group-II	Total
No. of measurers above the combined median	(A)	(B)	(A+B)
No. of measurers below the combined median	(C)	(D)	(C+D)
Total	(A+C)	(B+D)	

Under the null-hypothesis, we would expect about half of each group's measures to be above the combined median and about half to be below it, that is, we would expect frequencies A and C to be equal, and frequencies B and D to be nearly equal. In order to test this null-hypothesis, we calculate, *f 2* using the formula (12):

$$f2 = \frac{N\ (AD - (BC - N/2)^2}{(A+B)\ (C+D)\ (A+C)\ (B+D} \quad \ldots\ldots\ldots\ldots (12)$$

Let us illustrate the use of formula (12) in the following example. Eighteen male and female distance learners of a study centre of an open university were asked to express their attitudes towards the functioning of a study centre. Both the groups were administered an attitude scale and a common median attitude measure was worked out. The number of cases from both the groups falling above and below the median point is shown in Table 3.9.

Table 3.9: Distribution of Distance Learners (Male and Female)Below and Above the Common Median

	Below Medium	Above Medium	Total
Female Distance Learner	10	4	14
Male Distance Learner	6	12	18

Total	16	16	32

Using formula (12),

$$f^2 = \frac{32\,(10)\,(12 - (6)\,(4) - 32 \;\Big|\; \Big|}{16 \times 16 \times 18 \times 14} \quad \ldots\ldots\ldots\ldots (12)$$

$$= 3.17$$

Since the obtained value 3.17 of f 2 for 1 df does not exceed the f 2 critical value of 3.84 for a two-tailed test at 0.05 level, the null hypothesis is retained and we may conclude that there is no difference in the attitude of male and female distance learners towards the functioning of the centre.

Check Your Progress 5

The following table shows the number of male and female distance learners who have passed or failed an item of an achievement test. Test whether the item of the test differentiates between the two groups of males and females.

	Number passed the test Item	Number failed the test item
Males	30	20
Females	25	15

Note:a) Space is given below for writing your answer.

b) Compare your answer with the one given at the end of the Unit.

……

……

……

……

15.5 LET US SUM UP

In this Unit we described the nature of parametric and non-parametric tests along with the assumptions they are based on. The applications of some parametric test (Z, t and F) and non-parametric tests (Chi-square) in the analysis of data have also been discussed.

1) Parametric tests are used in the analysis of data available in interval or ratio scales of measurement.

2) Parametric tests assume that the data are normally or nearly distributed.

3) Z-test, t-test and F-test are most commonly used parametric tests.

4) If large-sized samples, i.e., those greater than 30 in size are selected at random from and infinite population, the distribution of sample means is called the 'sampling distribution of means'. This distribution is normal and it possess all the following characteristics of a normal distribution:

   i) The average values of sample means will be the same as the mean of population.

   ii) The distribution of the sample means around the population mean has its own standard deviation which is known as the 'standard error of the mean' (SEM or □□M).

   iii) The sampling distribution is centered at the unknown population mean with standard deviation 0.50.

   iv) The sample means often fall equally on the positive and negative sides of the population mean.

   v) About 2/3 of the sample means (exactly 68.26 percent) will lie within ± 1.00 □□M of population mean, i.e., within a range of ± 1 × 0.50 or ± 0.50 of the population mean.

   vi) 95 of the 100 sample means will lie within ± 1.96 □M of the population mean, i.e., 95 of 100 sample means will lie within ± 1.96 × 0.50 or ± 0.98 of the population mean.

   vii) 99 of the 100 sample means will be within ± 2.58 □□M of the population mean, i.e., 95 of our sample means will lie within ± 2.58 × 0.50 or ± 1.29 of the population mean.

5) If we draw a large sample randomly from a population and compute its mean, the mean has 95 percent chance of being within 1.96 □□M units from the population mean. Also, there is a 99 percent chance that the sample mean lies within 2.58 □□M units from population mean. To be more specific, there is 95 per cent probability that the limits M ± 1.96 □□M encloses the population mean and 99 percent probability that the

limits $M \pm 2.58$ □□M enclose the population mean.

6) The limits $M \pm 1.96$ □□M and $M \pm 2.58$ □□M are called 'confidence intervals' for 0.05 and 0.01 levels of confidence respectively.

7) When a 'statistic' is used to estimate a 'parameter', the number of 'degrees of freedom' depends upon the restrictions placed. Therefore, the number of degrees of freedom (df) will vary from one statistictoanother. In estimating the population mean from the sample mean, for example, 1 df is lost and so the number of degrees of freedom is N-1.

8) The Z-test is used for testing the significance of the difference between the means of two large samples.

9) The t-test is used for testing the significance of the difference between the means of two small samples.

10) Under the null hypothesis, the difference between the sample means may be either plus or minusand as often in one direction as in the other from the true (population) difference of zero, so that in determining probabilities we take both tails of the normal sampling distribution and make use of two-tailed test. But in many situations our primary concern is with the direction of thedifference rather than with its existence in absolute terms. In such situations, we make use of one- tailed test.

11) We use Z-test and t-test to determine whether there is any significant difference between means of two random samples. But when the number of samples is more than two, F-test, based on the technique of analysis of variance, is used for testing the significance of the sample means.

12) Non-parametric tests are used in the analysis of non-parametric data, i.e., when the data are available in nominal or ordinal scales of measurement.

13) Non-parametric tests are distribution-free tests and do not rest upon the more stringent assumptions of normally distributed population.

14) Chi-square test, median test, Sign-test, Mann-Whitney U test, Wilcoxon Matched Pairs Signed Ranks test and Kolmogorov Smirnov two Sample test are examples of some non-parametric tests.

15) Chi-square is used with discrete data in the form of frequencies. It is a test of independence, and is used to estimate the likelihood that some factor other than 'chance' accounts for the observed relationship between the variables.

16) Median test is used for testing whether two independent samples differ in central tendencies. It is particularly useful whenever the measurements for the two samples are expressed in an ordinal scale. Now, use the following checklist and see whether you have learnt to:

- classify various statistical test,

- describe the nature of parametric tests along with the assumptions on which they are based,
- define sampling distribution of means,
- define the standard error of mean,
- define the confidence intervals and levels of confidence,
- compute 0.95 and 0.99 confidence intervals for the true mean from a large sample mean,
- define and illustrate the concept of degrees of freedom,
- compute 0.95 and 0.99 confidence intervals for the true mean from the sample mean,
- apply Z-test for testing the significance of the difference between means of two independent large samples involving: (i) one-tailed and (ii) two tailed tests,
- apply t-test for testing the significance of the difference between means of two independent small samples involving: (i) one-tailed and (ii) two tailed tests,
- describe the nature and uses of the analysis of variance,
- state the basic assumptions of the technique of analysis of variance,
- apply F-test for testing the significance of the difference between means,
- describe the nature of the non-parametric tests along with their assumptions,
- name various non-parametric tests,
- describe the use of Chi-square test,
- illustrate the application of Chi-square test, and
- describe the use of median test.

15.6 GLOSSARY

Parametric Tests — These are statistical tests which are used for analysing parametric data and making inferences about the parameters from the statistics. These tests are based upon certain assumptions about the nature of data distributions and the types of measure used.

Non Parametric Tests — These tests are statistical tests, which are used for analysing non-parametric data and making possible useful inferences without any assumptions about the nature of data distributions.

Standard Error of Mean It is the standard deviation of a distribution of sample means.The number of degrees of freedom in a distribution is the number of observations or values that are independent of each other, and cannot be deduced from each other.

Degrees of Freedom

15.7 CHECK YOUR PROGRESS – THE KEY

1) Parametric data should be used if the following basic assumptions are met. These assumptions arebased on the nature of the population distribution and on the way the scale is used to quantify thedata observations.

   i) The observations are independent. The selection of one case is not dependent upon the selection of any other case.

   ii) The population values are normally distributed.

   iii) The samples have equal or near equal variances. This condition is known as equality or homogeneity of variances and is particularly important to determine when the samples are small.

   iv) The variables described are expressed in interval or ratio scale and not in nominal or ordinal scale of measurement.

2) The Standard Error of Mean is

$$SE_M = \sigma M = \frac{\sigma}{\sqrt{N}}$$

$$= \frac{5.82}{\sqrt{100}}$$

$$= \frac{5.82}{10}$$

$$= 0.582$$

3) i)

Boys ($N_1 = 5$) Girls ($N_2 = 8$
X_1 $X_1X_1^2$ X_2X_2 X_2^2

20 -6 36 34 9 81 22 -4 16 25 0 0 304 16 16-9 81 326 36 30 5 25 260 0 22-3 9 27 2 4 $\Sigma x_1 = 130$ $\Sigma x_1^2 = 104$ 20-5 25 26 11
$\Sigma x_2 = 200$ $\Sigma x_2^2 = 226$
$\Sigma X_1 130$ $\Sigma X_1 130$ 200 Mean = $M_1 = N_1 = 5$ Mean = $M_2 = N_2 = 8$

ii) df = N1 + N2 – 2 = 5 + 8 – 2 = 11

iii) Using formula (7)

iv) We used a two-tailed test as we are not hypothesizing a direction of the difference between the means. The t-values as given in Table II in the Appendix for 11 df for 0.05 and 0.01 columns are 2.20 and 3.1 respectively. Since the obtained value of 0.32 is less than these table values, the difference between the mean interest scores of boys and girls is not significant.

4) Non-parametric tests are distribution-free tests and are based on the following assumptions:

   i) the nature of the population distribution, from which samples are drawn is not known to be normal.

   ii) The variables are expressed in nominal form (classified in categories and represented by frequency counts).

   iii) The variables are expressed in ordinal form (ranked in order).

5) i) The number of the males and females who have passed or failed the test item is given in the following 2x2 table.

Number Passed	Number Failed	Total
Males	30	20 50
	(A)	(B) (A+B)
Females	25	15 40
	(C)	(D) (C+D)

Total 55 35 90 (A+C)(B+D)	

ii) Using formula (10)

$$X^2_C = \frac{27\ (7 \times 6 - 5 \times 9 - 27/2)2}{(7+5\ (9+6)\ (7+9)\ (5+6)} \quad \ldots\ldots\ldots\ldots (11)$$

$$= \frac{27\ (42 - 45 - 13.5)2}{12x15x16x11}$$

$$= 0.23$$

Since the obtained 0.058 of X2 does not exceed the Table value 3.841 of X2 at 0.05 level of significance, we may conclude that the test item does not differentiate between the two groups of males and females.

References

Fisher, R. A. (1950), *Statistical Methods for Research Workers,* Hagner Publishing, New York.

Guiford, J.P. (1954), *Psychometric Methods,* McGraw Hill, New York.

Wiersmar, William (1986), *Research Methods in Education: An Introduction*, Allyn and Bacon, Inc. Massachusetts.

UNIT 16 REPORTING RESEARCH

Structure

16.0 Introduction

16.1 Objectives

16.2 Why and How to Write a Research Report

16.3 The Beginning

16.4 The Main Body

- 16.4.1 Chapters and their Functions
- 16.4.2 Writing Style
- 16.4.3 References
- 16.4.4 Typing and Production
- 16.4.5 Tables and Figures

16.5 The End

- 16.5.1 Bibliography and References
- 16.5.2 Appendices

16.6 Let Us Sum Up

16.7 Check Your Progress: The Key

16.0 REPORTING RESEARCH

Every research activity is concluded by presenting the results and discussions. The reporting of a research study depends on the purpose with which it was undertaken. One might have conducted a study as a personal research, as an institutional project, as a project funded by an outside agency, or towards fulfilling the requirement for the award of a degree.

Research studies, when reported, follow certain standard patterns, styles and formats for maintaining parity in reporting and for easy grasp by others who are concerned with those studies. The present Unit is devoted to this aspect of research: How to write a research report? It starts with the purposes of writing a research report,followed by the components of the report itself (the beginning, the main body, and the end).

16.1 OBJECTIVES

After the completion of this Unit you should be able to:

- state the reasons for writing a research report;
- list the three main components of a research report;
- describe each component of a research report; and
- write the final report of any research study, viz., its beginning, the main body, and the end of the report.

16.2 WHY AND HOW TO WRITE A RESEARCH REPORT

Once you complete your research project, you are expected to write the report. A research report is a precise presentation of the work done by a researcher while investigating a particular problem. Whether the study is conducted by an individual researcher or by an institution, the findings of the study should be reported for several reasons. There reasons are:

- People learn more about the area of study.
- The discipline gets enriched with new knowledge and theories.
- Researchers and practitioners in the field can apply, test and retest the findings already arrived at.
- Other researchers can refer to the findings and utilise the findings for further research.
- Findings can be utilized and implemented by the policy makers or those who had sponsored the project.

The final report of a research exercise takes a variety of forms.

- A research report funded by an educational institution may be in the form of a written document.
- A research report may also take the form of an article in a professional journal.
- The research reports of students of M.A., M.S.W., M.Sc., M.Ed., M Phil. Or Doctoral programmes take the form of a thesis or dissertation.

In the following sections we shall discuss the main components of a research report. The entire research report is mainly divided into three major divisions: the beginning, the main body and the end (please see

Beginning	Main Body	End
• Cover/Title Page • Second cover • Acknowledgement • Content • List of Tables • List of Figures	• Introduction • Review of Literature • Design of the Study • Analysis and Interpretation of Data • Main Findings and Conclusion • Summary	• Bibliography and references • Appendices

16.3 THE BEGINNING

The beginning of a report is crucial to the entire work. The beginning or the preliminary section of the research report contains the following items, more or less in the order given below:

- Cover or Title Page
- Acknowledgements
- Table of Contents
- List of Tables
- List of Figures and Illustrations
- Glossary

Let us describe in brief each of the above six items of the preliminary section of a report.

i) *Cover or Title page*

The cover page is the beginning of the report, though different colleges, universities and sponsoring institutions prescribe their own format for the title page of their project report of thesis, generally, it indicates the downward vertical order:

- title of the topic,
- relationship of the report to a degree, course, or organisational requirement,
- name of the researcher,
- name of the supervisor,
- name of the institution where the report is to be submitted, and
- the date of submission.

The title page should carry a concise and adequately descriptive title of the research study. The title should briefly convey what the study is about. Researchers tend to make errors in giving the title by using too many redundant and unimportant words.

Here, we have drawn a list of a few titles of research reports and doctoral theses:

a) A Critical analysis of Textual Material for Principles of Accounting and its Translation for Distance Education

b) Developing Self-Instructional

c) Planning, Design and Development of one Self-Instructional Unit in print

In title (b), it is not clear at which level the researcher is developing self-instructional material.

The title should be written either in bold letters or upper-lower case and be

placed in the central portion of the top of the cover page. Here, we have reproduced the cover page of a research report in Box 1.

Evaluation Scheme of Assistance To Organisations For the Disabled

Sponsored By:

Ministry of Social Welfare Government of India

Director

Dr. D.K. Lal Das

Principal

College of Social Work (ICSW)

Red Hills, Hyderabad - 500004

March, 2002

Box 1 : Example of the title page of a research report

Note the other points mentioned on the cover page. Also observe the placement of these points.

ii) ***Preface or/and Acknowledgement***

Preface is not a synonym for either a Acknowledgement or a Foreword. A preface should include the reasons why the topic was selected by the researcher. It may explain the history, scope, methodology and the researcher's opinion about the study. The preface and acknowledgements can be in continuation or written separately. This page follows the inner title page. It records acknowledgement with sincerity for the unusual help received from others to conduct the study. The acknowledgement should be non-emotional and simple.

iii) ***Table of Contents***

A table of contents indicates the logical division of the report into various sections and subsections. In other words, the table of contents presents in itemized form, the beginning, the main body and the end of the report. It should also indicate the page reference for each chapter or section and sub-section on the right hand side of the table.

Sample table of contents is given below:

Contents

Page		
Acknowledgement		i
Table of Contents		ii
List of Tables		iii
Chapter I	Introduction	1-16

Chapter II Methodology 7-16

Chapter III The Profile of the Respondents 17-37

i) Beneficiaries 18

ii) Non-Beneficiaries 28

Chapter IV The Organisations for the Disabled 40-95

i) Physical Setting 40

ii) Administration and Organisation 43

iii) Finance 49

iv) Opinion of Community People 94

Chapter V Evaluation of the services under the Scheme 96-107

i) The Services 97

ii) Opinion of Beneficiaries 97

ii) Opinion of Non-Beneficiaries 103

iv) Grant-in-Aid System 106

Chapter VI Major Findings 107

Chapter VII Conclusions and Suggestions 108-134

Bibliography 135

Appendixes

I Voluntary Organisation 138

II Interview Schedule for Administrator 142

III Interview Schedule for Beneficiaries 145

IV Interview Schedule for Non-Beneficiaries 150

V Interview Guide for Authorities/Community People 153

VI List of Voluntary Organisations 154

Box 2 : Table of Contents

iv) ***List of Tables***

The table of contents page is followed by the page containing a list of tables. The list contains the exact title of each table, table number and the page number on which the table has appeared. We provide you in Box 3 an example of a list of tables.

Table	Title	Page
1	Population of Disabled	2
2	Age of the Beneficiaries	18
3	Level of Education	20
4	Sex of the Beneficiaries	21
5	Occupation of the Beneficiaries	22
6	Caste of the Beneficiaries	23
7	Marital Status	24
8	Size of the Family	25
9	Income of the Family	26
10	Nature of the Disability	27
11	Age of Non-Beneficiaries	28
12	Level of Education of Resident Beneficiaries	30
13	Sex	31
14	Caste	32
15	Marital Status	33
16	Size of the Family	34
17	Income of the Family	35
18	Nature of Disability	37
19	Staffing Pattern	46
20	Year wise Distribution of grants-in-aid	51
21	Age and Opinion of Beneficiaries on Utility of the Services Provided under the Scheme	59

v) ***List of Figures and Illustrations***

The page' List of Figures' comes immediately after the 'List of Tables' page. You will observe in the following example that the list of figures is written in the same way as the list of tables.

List of Figures

1.	Existing NGOs in A.P	7
2.	Conceptual Framework of Capacity Building	9
3.	Network of Training Institutions	31
4.	Nodal Organisations in India	38
5.	Interactive Rehabilitation	40
6.	Aids for Disabled	41

7. Communication Network 43

8. Implementation of the Scheme 48

Box 4: Example of list of figures

vi) ***Glossary***

A glossary is a short dictionary, explaining the technical terms and phrases which are used with special connotation by the author. Entries of the technical termed are made in alphabetical order. A glossary may appear in the introductory pages although it usually comes after the bibliography. An exemplar glossary is given below.

Algorithm	A step-by-procedure consisting of mathematical and/or logical operations for solving a problem.
Artificial Intelligence (AI)	The study of computer techniques that mimic certain functions typically associated with human intelligence.
Back-up	Duplication of a program or file on to a separate storage medium so that a copy will be preserved against possible loss or damage to the original.
Benchmark	A measured point of reference from which comparisons of any kind may be made, often used in evaluating hardware and software, in comparing them against one another.
Command	An instruction to the computer which is not a part of a program.
Cybernetics	The field of science involved in comparative study of the automatic control or regulation of, and communication between machine and man. These studies include comparisons between information-handling machines and the brainsand nervous systems of animals and humans.
Data	Input to a computer that is processed by mathematical and logical operations so that it can ultimately be output in a sensible form.
Data Processing	The input, storage, manipulation and

	dissemination of information using sequences of mathematical and logical operation.
Electronic Spreadsheet	Software that simulates a worksheet in which the user can indicate data relationships. When data are changed, the program has the ability to instantly recalculate any related factors and to save all the information in memory.
Graphics Package	A programme that helps draw graphs.
Hard Copy	Output of information in permanent form, usually on paper, as opposed to temporary display on a CRT screen.
Ink Jet Printer	A type of printer in which dot matrix characters are formed by ink droplets electrostatically aimed at the paper surface.
Laser Printer	A printer that uses a laser beam to form images on photo-sensitive drums. Laser printers are now used as output devices for computers.
Megabyte (MB or M-Byte):	1024 kilobytes or 1024 □□1024 bytes.
Personal Computer	A moderately priced, general use computer designed principally for a single user in a home or small-office environment.

Source: *Balagurusamy E: Selecting and Managing Small Computers.*

vii) ***List of Abbreviations italics***

To avoid repeating long names again and again, a researcher uses abbreviations. Since abbreviations are not universal, it is necessary to provide the full form of the abbreviations in the beginning. An exemplar list of abbreviation is given below.

ABBREVIATIONS

AIMA All India Management Association
AIR All India Radio
APPEP Andhra Pradesh Primary Education Project
AVRC Audio Visual Resource Centre
ATI Administrative Training Institutes

BEL	Bharat Electronics Limited
BEO	Block Education Officer
BRC	Block Resource Center
BSE	Board of Secondary/Senior Secondary Education
CABE	Central Advisory Board of Education
CBT	Computer Based Training
CEO	Circle Education Officer
CIET	Central Institute of Educational Technology
CRC	Cluster Resource Center
CSS	Centrally Sponsored Scheme
DIET	District Institute of Educational Technology
DIT	District Institute of Training
DD	Doordarshan
DOE	Department of Electronics
DoSpace	Department of Space
DOT	Department of Telecommunication
DPEP	District Primary Education Program
DPEPII	District Primary Education Program: Phase II
EMRC	Education Media Resource Centre

Activity

Take any report which has been prepared by your institution and check whether the title page contains all the essential information. If not, try to fill in the gaps.

……

……

……

……

……

Check Your Progress 1

List the major parts of 'the beginning' of a research report. Describe briefly the importance of each part.

Note: a) Space is given below for writing your answer.

b) Compare your answer with the one given at the end of the unit.

……

……

……

.....

.....

16.4 THE MAIN BODY

The main body of the report presents the actual work done by an investigator or a researcher. It tells us precisely and clearly about the investigation/study from the beginning to the end. The methodology section of the final report should be written in the past tense because the study has been completed. The report categorically avoids unnecessary details and loose language–we shall examine this point in detail in this section. At this stage, you may again look at the Box on page 4. You will find that the table of contents for the report outlined of six section/chapters in the main body. These are:

- Introduction
- Review of Literature
- Design of the Study
- Analysis and Interpretation of Data
- Major Findings
- Conclusions and Discussions
- Summary

Besides the logicality of sections/chapters in the main body there are certain other important aspects which need our attention. They are the style of writing, the design and placement of references and footnotes, the typing of the report, and the tables and figures. Let us elaborate these points in the following sub-sections.

Let us elaborate these points in the following sub-sections.

16.4.1 Chapters and their Functions

We will discuss the chapterisation of a thesis or a research report under six heads as noted above. Let us begin with introduction which is usually the first chapter.

Introduction

This is the first chapter of a research report. It introduces the topic or problem under investigation and its importance. The introductory chapter:

- Gives the theoretical background to the specific area of investigation,
- States the problem under investigation with specific reference to its placement in the broader area under study,
- Describes the significance of the present problem,
- Defines the important terms used in the investigation and its reporting,

- States precisely the objective(s) of the study,
- States precisely the important terms used in the study that would be tested through statistical analysis of data, and
- Defines the scope and limitations of the investigations.

Although these sub-sections are common, it is not necessary to follow the given order strictly; there may be variation in the order of the sub-sections. Sometimes the review of literature related to the area under investigation is also presented in the first chapter and is placed immediately after providing the theoretical background to the problem. Many researchers use review to argue the case for their own investigation. In experimental research it becomes essential to review related studies to formulate the hypotheses.

Review of Literature

The second chapter of a research report usually consists of the review of the important literature related to the problem under study. This includes the abstraction of earlier research studies and the theoretical articles and papers of important authorities in the field. This chapter has two functions. Firstly while selecting a problem area or simply a topic for investigation, the researcher goes through many books, journals, research abstracts encyclopaedia, etc. to finally formulate a problem for researcher in order to decide on a specific problem for investigation. The review of literature is the first task for a researcher in order to decide on a specific problem for investigation. It also helps in formulating the theoretical framework for the entire study. Secondly, such a review helps the researcher to formulate the broader assumptions about the factors/variables involved in the problem and later develop the hypothesis/hypotheses for the study. Besides these, the review also indicates the understanding of the researcher in relation to the area under investigation, and thus his/her efficiency to carry out the study. While reviewing literature in the area concerned, you have to keep in mind that (reviewed) literature has to be critically analysed and summarised in terms of agreements and disagreements among the authors and researchers in order to justify the necessity for conducting your investigation. Researchers may make two types of errors in their review exercises. Many simply report the findings of one study after another in sequential order without showing how the findings are connected with one another. Others report on studies that are at best only marginally related to their own hypotheses.

Design of the Study

The design of a study is usually described in the third chapter of the report. Broadly speaking, this chapter provides a detailed overview of "how" the study was conducted. The various sub-section include:

i) description of the research methodology,

ii) variables: the dependent and intervening variables with their operational definitions;

iii) sample: defining the population, and the sampling procedure followed to select the sample for the present study;

iv) listing and describing various tools and techniques used in the study, like questionnaires, attitude scales, etc., whether these have been adopted or developed by the investigator, their reliability, validity, item description, administration and scoring, etc.;

v) describing the statistical technique used in the analysis of data including the rationale of the use and method of data analysis.

Analysis and interpretation of data

This is fourth chapter of the research report. It is the heart of the whole report, for it includes the outcome of the research. The collected data are presented in tabular form and analysed with the help of statistical techniques–parametric and nonparametric. The tables are interpreted and if necessary, the findings are also presented graphically. The figures do not necessarily, repeat the tables, but present data visually for easy understanding and easy comparisons. Data may be presented in parts under relevant sections. The analysis of the data not only includes the actual calculations but also the final results. It is essential that at each stage of analysis the objective(s) of the study and their coverage is taken care of. This chapter also presents the details about the testing of each hypothesis and the conclusions arrived at. This gives the reader a clear idea regarding the status of the analysis and coverage of objectives from point to point.

Main findings and conclusion

This is usually the fifth chapter in a research report. The major findings of the study analysed and interpreted in the preceding chapter are precisely and objectively stated in this chapter. The fourth chapter contains such presentations as only a specialist or a trained researcher can understand because of the complexities involved; but in the fifth chapter the major findings are presented in a non-technical language so that even a non-specialist such as a planner or an administrator in the field can make sense out of them.

The main findings are followed by a discussion of the results/findings. The major findings are matched against the findings of other related research works which have already been reviewed in the second chapter of the report. Accordingly, the hypotheses formulated in the first chapter are either confirmed or discarded. In case the null-hypotheses are rejected, alternative hypotheses are accepted. If the findings do have any discrepancy in comparison with those of other researches, or if the findings do not explain sufficiently the situation or problem under study, or if they are inadequate for generalisations, explanations with proper justification and explanation have to be provided.

The next task in this chapter is to provide implications of, the findings and

their generalisations. The implications should suggest activities for and provide some direction to the practitioners in the field. Unless these implications are clearly and categorically noted, it becomes difficult for the practitioners to implement them on the one hand, and on the other research findings do not get utilised at all even if they have been recorded in a report. The implications follow a presentation/listing of the limitations of the study on the basis of which suggestions are made to carry out further investigation or extend the study from where it has reached.

Summary

Some researchers include a summary along with the research report (as the last chapter) or as a pull-out to the report itself. If sums up precisely the whole of the research report right from the theoretical background to the suggestions for further study. Sometimes researchers get tempted to report more than what the data say. It is advisable to check this tendency and be always careful to report within the framework provided by the analysis and interpretation of data, i.e., within the limits of the findings of the study.

Check Your Progress 2

Comment briefly on the uses of (a) review of literature, and (b) conclusion in a research report.

Note: a) Space is given below for writing your answer.

b) Compare your answer with the one given at the end of the unit.

……

……

……

……

……

16.4.2 Writing Style

The style of writing a research report is different from other writings. The report should be very concise, unambiguous, and creatively presented. The presentation should be simple, direct and in short sentences. Special care should be taken to see that it is not dull and demotivating.

Statements made should be as precise as possible – they should be objective and there should be no room on for subjectivity, personal bias and persuasion. Similarly, over generalisation must be avoided. There is no place for hackneyed, slang and flippant phrases and folk expressions. The writing style should be such that the sentences describe and explain the data, but do not try to convince or pursuade the reader. Since the report describes what has already been completed, the writing should be in the past tense.

In the case of citations, only the last name of the author is used, and in all

cases academic and allied titles like, Dr., Prof., Mr., Mrs., etc. should the avoided. Some authors recommend that the use of personal pronouns like "I", "We" etc., should be avoided. There is however no hard and fast rule in this case. Similarly, a large number of research reports use passive voice which is strongly discouraged by the linguists. Similarly, abbreviations of words and phrases – like IGNOU, DDE, NIRD, etc. – should be used to avoid long names repeatedly inside the text, as well as in figures, tables, and footnotes.

Special care should be taken while using quantitative terms in a report, such as *few* for number, *less* for quantity etc. No sentence should begin with numericals like "40 students", instead it should start as "Forty students". Commas should be used when numbers exceed three digits –1,556 or 523,489, etc.

Language, grammar and usage are very important in a research report, the *Roget's Thesaurus Handbook of Style* by Campbell and Ballon (1974), and a good dictionary would be of much help. MS-Word software provides good support to

- Spelling and Grammar
- Thesaurus
- Auto Correct
- Auto Summarise

A researcher is advised to use these features on the MS-Word to make the report error free. It is always advisable to show the report to learned friends or language experts for correction before it is finally typed. Revision is an important feature of good report writing – even experienced researchers with many publications revise their reports many times before giving them for final typing.

16.4.3 References

Articles, papers, books, monographs, etc. quoted inside the text should always accompany relevant references, i.e., the author and the year of publication e.g., (Mukherjee, 1998). If a few lines or sentences are actually quoted from a source, the page number too should be noted e.g., (Mukherjee, 1998:120-124). Besides, full reference should be placed in the Reference section of the report (see sub-section 4.5.1 below).

In preparing the references, another factor to be considered is the abbreviations of words and expressions and their right placement. While writing a research report, abbreviations may be used to conserve space in references. If a researcher is not familiar with the abbreviations, he/she should consult the relevant literature as and when required. In the following table (table 1) a comprehensive list of abbreviations has been given for ready reference (the Latin abbreviations have been italicised).

Table 1: List of some important Abbreviations used in Footnotes and Bibliographies

Words	**Abbreviation**
About (Approximate)	c. (*circeca*)
Above	*supra*.
And the Following	*et seg*.
And the following	*f.,ff*.
And others	*et. al*.
Article, articles	art., arts.
Article, articles	*infra*.
Book, books	bk., bks.
Chapter, chapter	chap., chaps.
Colum, colums	Col., Cols.
Compare	*cf.*
Division, division	div., divs.
Editor, editors	ed., eds.
Edition, editions	ed., eds.
For example	*e.g.*
Figure, figures	fig., figs.
Here and there (scattered)	*Passim*
Illustrated	Ill
Line, lines	l. ll.
Manuscript	ms.
Mimeographed	mimeo.
No date given	n.d.
No name given	n.n.
No place given	n.p.
Number, numbers	no., nos.
Page, pages	p., pp.
Part, parts	pt., pts.
Paragraph in lenth	(….)
Paragraph, paragraphs	par., pars
Previously cited	*op. cit*.

Revised	rev.
Same person	*idem*.
Same reference	*ibid.*
Section, sections	sec., secs.
See	*Vide*
The place cited	*loc. Cit*
Thus	*sic*.
Translated	trans.

16.4.4 Typing and Production

Typing of dissertations, research reports, project reports etc. needs greater care then other typed documents. In a research report, one does not expect overwriting, strikeovers, erasures and insertions.

Before typing the report, it is necessary to check whether the handwritten report, i.e., the manuscript is in a proper shape. Whether the manuscript of the report is typed by a typist or by the researcher himself /herself, a clear and comprehensible manuscript makes typing easy. Too many additions and corrections make the manuscript crammed., and a crammed manuscript makes typing difficult and time consuming. Only one side of the paper should be typed and typing should be double spaced. Space should be left on each side of the paper as follows:

- left side margin
- right side margin
- top margin
- bottom margin

If there is a lengthy quotation, it should be indented and typed in single space. At the end of each line, words should be divided as per convention. A dictionary which shows syllabification should be consulted if words are to be broken at all. Unlike the lengthy quotations, short quotation of three/four lines may be included in the text within quotation marks. Subject to access to a computer and word processing software, 8 it is better to prepare the report on a computer. It has several advantages, for example, you can

- edit time and again without incorporating new errors which is what happens when you use a manual typewriter,
- define your margin – top, bottom, left and right easily,
- define pages in landscape or portrait size, particularly for tables and diagrams,
- choose out of about 70+ fonts – shapes of letters and type sizes from the smallest 8 point to the large 72 point,

- check spelling, grammar, synonyms and antonyms,
- choose illustrations from the clip-art file, and
- can index (alphabetical order) the references automatically.

16.4.5 Tables and Figures

Tables: Preparation and appropriate placement of tables in the text are equally important. They need careful attention from the researcher. Tables help the readers to get a quick view of the data and comprehend vast data at one go. However, tables should be presented only when they are necessary. Too many tables may confuse the reader, instead of facilitating his/her reading. As such you need to be selective in placing tables in the report. If data are too complicated to be presented in one table, several tables, may be used to give a clear picture of the data in proper sequential order. Tables, if small, may accompany the textual material, and if large, should be put on one full page without mixing them with the text. All the tables should be numbered serially in the text, so that they may be quoted or referred to with the help of those numbers conveniently.

If a table is large, it should continue on the next page with the table title repeated on the top of the next page; otherwise, tables can be typed in smaller fonts like 8 point or 9 point. The table itself is centred between the two margins of the page, and its title typed in capital letters and is placed in pyramid size. The title of the table should be brief but self-explanatory.

Figures: Figures are necessary when the data is to be presented in the graphic form. They include charts, maps, photographs, drawings, graphs, diagrams, etc. The important function of a figure is to represent the data in a visual form for clear and easy understanding. Textual materials should not be repeated through figures unless very necessary.

Figures should be as simple as possible and the title of each figure should precisely explain the data that has been presented. Usually, a figure is accompanied by a table of numerical data. Again, figures are presented only after textual discussion and not the other way round. The title design of figures should be followed consistently throughout the report. Every first letter of a word of the title should be in capitals, and figures should be numbered in Indian numerals like 1, 2, 3 etc. And the title, unlike for tables, is presented below the figure.

16.5 THE END

The end of the report consists of references and appendix/appendices. References come at the end after the last chapter of the report. The last section labelled as references appears at the top of a new sheet of paper. The reference section is a list of the works that have been cited in the report/thesis. All references quoted in the text are listed alphabetically according to the last name of the authors. The works of the same author

should be listed according to the date of publication with the earliest appearing first.

16.5.1 Bibliography and References

Research reports present both bibliographies and references. Although many researchers use these terms interchangeably, the two terms have definite and distinct meanings. A bibliography is a list of titles – books, research reports, articles, etc. that may or may not have been referred to in the text of the research report. References include only such studies, books or papers that have been actually referred to in the text of the research report. Whereas research reports should present references, books meant for larger circulation may be listed in bibliographies that should include all such titles as have been referred to.

There are mainly two style manuals detailing general form and style for research reports. These are:

- American Psychological Association, *Publication Manual,* 3rd ed. Washington, DC: American Psychological Association, 1983.
- *The Chicago Manual of Style, 13th* rev.ed., Chicago University of Chicago Press, 1982.

Style of Referencing

There are mainly two types of referencing:

1) arranging references in alphabetical order where the researcher has cited the name of the author and year of publication/completion of the work in the text.
2) arranging references in a sequence as they appear in the text of the research report. In this case, related statement in the body of the text is numbered.

However, most research reports use alphabetical listing of references.

For example, entries in a reference section may look like the following:

Gannicott, K. and Throsby, D. (1994), *Educational Quality and Effective Schooling*, UNESCO, (Book), Paris.

Koul, B.N., Singh, B. and Ansari, M.M. (1988), *Studies in Distance Education*, IGNOU & AIU, New Delhi.

Kumar, K.L. (1995), *Educational Technology*, New Age Publishers, New Delhi.

Ministry of Human Resource Development, DPEP: *Guidelines* (1995), Department of Education MHRD, Government of India, New Delhi.

Mukhopadhyay, M. (ed.) (1990), *Educational Technology: Challenging*

Issues, Sterling Publishers, (Edited Book), New Delhi.

Mukhopadyay, M. (1998), "Teacher Education and Distance Education: The Artificial Controversy", in Buch, Piloo M., (ed.) *Contemporary Thoughts on Education*, SERD, (Chapter in Book), Baroda.

Parhar, M. (1993), Impact of Media on Student Learning, Unpublished Doctoral Dissertation, Jamia Millia Islamia, (Thesis), New Delhi.

Sachidananda, Tribal Education: New Perspectives and Challenges, *Journal of Indian Education*, New Delhi: NCERT, 1994. (Article in a Journal)

Selltiz, Claire et. al. (1959), *Research Methods in Social Relations*, Rinehart & Winston, Holt, New York.

Dhanarajan, Gajaraj, "Access to Learning and Asian Open Universities: In Context" in the 12th Annual Conference of Asian Association of Open Universities, (1998) "*The Distance Learner*" The Open University of Hong Kong, Hong Kong SAR, China, 4-6 Nov., (Conference Paper).

You would notice the following:

- All studies are arranged in alphabetical order
- The names of the authors are recorded by title and initials (not full name).
- To indicate two or three authors, 'and' is used between the first and the second, ',' between first and 'and' between second and third author.
- In case of more than three authors, only the name of the first author is mentioned followed by *et al.* (*et allibi*) or others.
- In case of a chapter in a book, after the author and chapter title and the name of the author or editor of the book.
- Titles of printed books, names of journals are highlighted by using 'italics' or by underlining (in case of manually typed material).
- Place of publication of a book precedes the name of the publisher separated by a ':'(colon).
- Names of journals are following by the relevant volume and issue numbers usually in the form 10(3) –Volume 10, Number 2 and page numbers.
- Unpublished thesis or dissertation titles are not highlighted and the word 'unpublished' is mentioned.

Referring Web Based Documents

Computers have brought revolution in all sectors of development including education. Computers were conventionally used for data storage, processing and retrieval. Through internet, information can be accessed from any part of the world. As researchers, reviewing the relevant literature related to the problem under study is almost magnum opus. These days, internet is a rich

academic and professional resource. World Wide Web (WWW) is the easiest and most popularly used browsing mechanism on the Internet. Here we will very briefly explain as how to write the references when we quote from any Web Site.

Citing E-Mail

E-Mail communications should be citied as personal communications as noted in APA's publication Manualhttp://www.apa.org/journals/webref.html. Personal Communications are hot cited in the reference list. The format in the text should be as:

Citing a Web Site

When you access the entire Web site (not a specific document on the site), you just give the address of the site in the text. It is not necessary to enter in the reference section.

For example,

http://www.ignou.ac.in (IGNOU's website)

http://www.webct.com/ (This site provides tools for development of web based courses)

Citation of specific document on a web site has a similar format to that for print. Here, we give few examples of how to cite documents. The Web information is given at the end of the reference section. The date of retrieval of the site should be given because documents on the Web can change in content or they may be removed from a site.

Example

Duchier, D. (1996), Hypertext, New York: Intelligent Software Group. [Online] http://www.isg.sfu.ca/duchier/misc/hypertext - review/chapter4.htm]Accessed on 25/1/99].

Flinn, S. (1996), Exploiting information structure to guide visual browsing and exploratory search in distributed information systems [Online] http://www.cs.ubc.ca/ reading-room / [Accessed June 1998]

If you have to cite some specific parts of a web document, indicate the chapter, figure, table as required.

16.5.2 Appendices

Usually, the appendices present the raw data, the true copy of the tools used in the study, important statistical calculations, photographs and charts not used inside the text. These are ordered serially like Appendix-1, Appendix-2, or they can be serialized with capital letters (Appendix A, Appendix B) etc. to facilitate referencing within the text. The appendices provide reference facilities to readers and others interested in that particular field of

investigation.

Activity

1) Take any report and check whether the references are written in the standard form. If not, try to rewrite them properly.

……

……

……

……

……

2) Examine the appendices in the same report. Are all of them essential for the report? Comment.

……

……

……

……

……

16.6 LET US SUM UP

In this Unit, we focused on research reporting as a professional activity. The purpose of writing the report depends on the reason behind undertaking the research study. It could be for obtaining a degree, or as a project report to be submitted to the funding agency, etc. Once submitted, the funding agency and the educational managers could utilise the findings and recommendations to achieve their objectives; other researchers may seek guidance from it; and lastly, the findings may be used for developing new theories in the discipline concerned. A research report has three parts: the beginning, the main body and the end. The beginning includes: cover or the title page, acknowledgements, table of contents, the list of tables, and the list of figures. The main body normally contains an introduction, review of the relevant literature, objectives, hypotheses, research design (research methodology, population and sample, tools, procedure of collecting data), analysis and interpretation of data, the main findings and conclusion (that also includes its educational implications and suggestions for further studies). While discussing the main body, we have talked about the style of writing the report, style and placement of footnotes and reference, the typing process and the format and placement of tables and figures. We closed the discussion with notes on the style, arrangement and placement of references and appendices which constitute the end of a research report.

16.7 CHECK YOUR PROGRESS: THE KEY

1) The major parts of the beginning of a research report are: cover/title page, acknowledgements, table of contents, list of tables, list of figures and list of abbreviations.

   The cover page gives us clear information about the subject/theme, author and the year of the research study as well as the organisation for which or where the study has been conducted.

   Acknowledgements are words of appreciation from the researcher for those who have helped him/her while conducting the study. Table of contents indicates the main themes/areas studied, the methodology followed and the outcome of the study.

   List of tables, figures and abbreviations are useful as reference tools.

2) a) Review of literature helps the researcher to specifically define the problem for investigation, decide about the usefulness of the study and formulate his/her hypothesis.

   b) The conclusion of a research report sums up the findings, states what is new in the report concerned and indicates the indicates the direction for future studies as well as implications for implementation of recommendations, if any.

Further Readings

Baker, L., Therese (1988), *Doing Social Research,* McGraw Hill, New York.

Black, James A. and Champion, Dean J. (1976), *Methods and Issues in Social Research,* John Wiley, New York.

Kerlinger, Fred R. (1964), *Foundations of Behavioral Research,* Surjeet Publications, Delhi.

Kidder, Louise H. (1981), *Research Methods in Social Relations,* Holt, New York.

Lal Das, D.K. (2000), *Practice of Social Research : A Social Work Perspective,* Rawat Publications, Jaipur.

Monette, Duane R. et. al. (1986), *Applied Social Research: Tool For the Human Services,* Holt, Chicago.

Nachmias D and Nachmias C. (1981), *Research Methods in the Social Sciences*, St. Martins Press, New York.

Rubin, Allen & Babbie E. (1989), *Research Methodology for Social Work*, Belmont, Wadsworth, California.