



Block

5

DATA ANALYSIS

UNIT 1

Overview of Statistical Tools	5
--------------------------------------	----------

UNIT 2

Use of Computer in Data Analysis	44
---	-----------

UNIT 3

Data Processing and Analysis	68
-------------------------------------	-----------

UNIT 4

Report Writing	82
-----------------------	-----------

PROGRAMME DESIGN COMMITTEE

Prof. Amita Shah Gujarat Institute of Development Research, Ahmedabad	Prof. P. Radhakrishnan Madras Institute of Development Studies, Chennai
Prof. S. K. Bhati Jamia Millia Islamia, New Delhi	Prof. Ramashray Roy (Rtd) Centre for Study of Developing Societies, Delhi
Prof. J. S. Gandhi (Rtd) Jawaharlal Nehru University, New Delhi	Prof. R. P. Singh (Rtd) Ex-Vice-Chancellor, MPUAT, Udaipur
Prof. Gopal Krishnan (Rtd) Punjab University, Chandigarh	Prof. K. Vijayaraghavan Indian Agriculture Research Institute, New Delhi
Prof. S. Janakrajan Madras Institute of Development Studies, Chennai	Dr. Nilima Shrivastava IGNOU, New Delhi
Prof. Kumar B. Das Utkal University, Bhubaneswar	Prof. B. K. Pattanaik IGNOU, New Delhi
Prof. Nadeem Mohsin (Rtd) A.N.Sinha Institute of Social Sciences, Patna	Dr. Nehal A. Farooquee IGNOU, New Delhi
	Dr. P.V.K.Sasidhar IGNOU, New Delhi

PROGRAMME DESIGN COMMITTEE (REVISED)

Prof. T.S. Papola Institute for Studies in Industrial Development, New Delhi	Prof. Nadeem Mohsin (Rtd) A.N.Sinha Institute of Social Sciences, Patna
Prof. S. Janakrajan Madras Institute of Development Studies, Chennai.	Prof. Rajesh University of Delhi, New Delhi
Prof. S. K. Bhati Jamia Millia Islamia, New Delhi	Prof. B. K. Pattanaik IGNOU, New Delhi
Prof. Preet Rustagi Institute for Human Development, Delhi	Prof. Nehal A. Farooquee IGNOU, New Delhi
Prof. Gopal Iyer (Rtd) Panjab University, Chandigarh	Prof. P.V.K. Sasidhar IGNOU, New Delhi
Dr. S. Srinivasa Rao Jawaharlal Nehru University, New Delhi	Dr. Pradeep Kumar IGNOU, New Delhi
Dr. S. Rubina Naqvi Hindu College, University of Delhi, Delhi	Dr. Nisha Varghese IGNOU, New Delhi
	Dr. Grace Don Nemching IGNOU, New Delhi

COURSE PREPARATION TEAM

Units Writers:

Prof. V. K. Tiwari
National Institute of Health and
Family Welfare
New Delhi (Unit 1, 3 and 4)

Dr. Nisha Varghese
SOEDS, IGNOU (Unit 2)

Editors:

Prof. V.K.Jain (Rtd)
(Content Editor) NCERT, New Delhi

Mr. Praveer Shukla
(Language Editor), New Delhi

Dr. Nisha Varghese, IGNOU

Prof. B. K. Pattanaik, IGNOU

Prof. Nehal A. Farooquee, IGNOU

Prof. P. V. K. Sasidhar, IGNOU

Course Coordinator : Dr.NishaVarghese,E-mail : nishavarghese@ignou.ac.in

Programme Coordinators: Prof. P.V.K. Sasidhar, Prof. B.K. Pattanaik,
Prof.Nehal A. Farooquee,

PRINT PRODUCTION

Mr. S. Burman Deputy Registrar (Publication) MPDD, IGNOU, New Delhi	Mr. KN Mohanan Asst. Registrar (Publication) MPDD, IGNOU, New Delhi	Mr. Babu Lal Rewadia Section Officer (Publication) MPDD, IGNOU, New Delhi
---	---	---

May, 2018

© Indira Gandhi National Open University, 2018

ISBN: 987-

All rights reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from the Copyright holder.

Further information on the IGNOU courses may be obtained from the University's office at Maidan Garhi, New Delhi or the official website of IGNOU at www.ignou.ac.in

Printed and published on behalf of IGNOU, New Delhi by Registrar, MPDD, IGNOU, New Delhi.

Laser Typeset by Rajshree Computers, V-166A, Bhagwati Vihar, (Near Sec. 2, Dwarka), New Delhi
Printed at:

BLOCK 5 DATA ANALYSIS

Block 5 on '**Data Analysis**' with four units gives an overview of various tools and techniques of data analysis needed for conducting development research.

Unit 1 on '**Overview of Statistical Tools**' provides information about various measures of central tendency and dispersion. It also discusses correlation, regression, as well as hypothesis testing.

Unit 2 on '**Use of Computer in Data Analysis**' discusses the use of MS Excel and SPSS in data analysis. Use of these software for working out mean, standard deviation, correlation, regression and testing of hypothesis has been discussed.

Unit 3 on '**Data Processing and Analysis**' discusses about data processing particularly tabulation and graphical presentation. It also briefs about data coding, editing and feeding.

Units 4 on '**Report Writing**' discuss about various types of research reports and details about the various components of a research report.

UNIT 1 OVERVIEW OF STATISTICAL TOOLS

Structure

- 1.1 Introduction
- 1.2 The Data: Meaning and Types
- 1.3 Frequency Distributions
- 1.4 Measures of Central Tendency
- 1.5 Measures of Dispersion
- 1.6 Hypothesis Testing and Inferential Statistics
- 1.7 Statistical Tests
- 1.8 Chi-Square Test
- 1.9 F-Test
- 1.10 Z-Test
- 1.11 t-Test
- 1.12 Correlation Analysis
- 1.13 Regression Analysis
- 1.14 Let Us Sum Up
- 1.15 Keywords
- 1.16 References and Selected Readings
- 1.17 Check Your Progress – Possible Answers

1.1 INTRODUCTION

Statistical tools are the pillars of the research study on which data analysis for all types of developmental programmes stand. Those who are researchers also need some understanding of statistical analysis to be able to produce a meaningful research report. With the availability of several user friendly software, the performance of statistical analysis has now become a reality, even for non-statisticians, provided they are computer literate, and understand the basic principles of statistical analysis.

There are two types of statistics:

- (i) Descriptive statistics which include techniques for organizing, summarizing, and presenting data using tables, graphs, or single numbers
- (ii) Inferential statistics which consist of statistical methods for making inferences about a population based on information obtained from a sample.

This unit aims to make you conversant with the basic statistical tools applicable in developmental research.

After studying this unit, you should be able to

- describe various measures of central tendencies and dispersion;

- discuss the applicability of various tests involved in hypothesis testing;
- explain the use of correlation in data analysis; and
- describe the use of regression in data analysis.

1.2 THE DATA: MEANING AND TYPES

Statistics is a science that deals with the collection, organization, analysis, interpretation and presentation of information that can be presented numerically and/or graphically to help us answer a question of interest. This helps us, in many ways, to

- (i) establish the problem and its effect
- (ii) approve or disapprove a hypothesis
- (iii) establish association and relationship, etc.

1.2.1 Data: Meaning and Types

In order to conduct a study, we collect raw data using one of the sampling techniques/ methods. This raw data does not provide any meaningful conclusion for the program or in policy-making unless; it is presented in some meaningful format. In order to have meaningful information from a study, we need to analyze the data using appropriate statistical tests and techniques. Before applying any statistical test, it is important to identify the type of data, as the choice of statistical tests depends on the type of data. The theory of measurement consists of a set of separate or distinct theories, each concerning a distinct level of measurement. We have already discussed four levels of measurement: Nominal, Ordinal, Interval, and Ratio and here we will discuss the statistics and statistical tests that are permitted with each level.

Types of Data

There are two types of data (i) Qualitative data, viz., occupation, sex, marital status, religion and (ii) Quantitative data, viz., age, weight, height, income, etc. These may be further categorized in the following two types.

- **Discrete:** data that can be divided into categories or groups, such as male and female, and can take only discrete values, as explained - nominal and ordinal scale.
- **Continuous:** data that can take any value including decimal are called continuous data. A continuous data is, at least in interval or ratio scale, as defined in types of measurement. Measurements in ordinal scale can also be considered under this category though it does not fulfill all conditions of the continuous scale. In social sciences, it is difficult to measure variables in interval or ratio scale and one has to depend on measurements taken in the ordinal scale. It is further emphasized that the statistics based on such measurement may provide 'under' or 'over' estimates of the population parameter.

1.2.2 Variable Types : two types of variables are used in analysis of data.

- **Independent Variable:** the characteristic being observed or measured which is hypothesized to influence an event or outcome (dependent variable) is called independent variable. It is not influenced by the event or outcome, but may cause it, or contribute to its variation.

- **Dependent variable:** a variable, whose value is dependent on the effect of other variables (independent variables) in the relationship, being studied. It is also called a Criterion, Outcome or Response variable.

1.3 FREQUENCY DISTRIBUTIONS

Consider a situation, when a test is administered to a group of 60 students. The marks obtained by each student can be listed against the names of students in a register or mark list. The data in the original form are called ungrouped data. When this is arranged in an ascending or descending order of magnitude, the data are said to be arranged in an array. Counting the number of times each value occurs, we get the frequency table of 'number of marks' and 'number of students'.

Consider an example in which 82 clinics in one district were asked to submit the number of patients treated for malaria in one month. 80 clinics out of the 82 responded to the call. The researchers presented both the frequency distribution and percentages (or relative frequencies).

Table 1.1: Distribution of clinics according to number of patients treated for malaria in one month

Number of patients	Number of clinics	Cumulative frequency	Cumulative Percentages	Range of cumulative frequency
1	2	3	4	5
0 - 19	5	5	6.25	0 - 5
20 - 39	8	13	16.25	6 - 13
40 - 59	10	23	28.75	14 - 23
60 - 79	11	34	42.5	24 - 34
80 - 99	19	53	66.25	35 - 53
100 - 119	10	63	78.75	54 - 63
120 - 139	9	72	90.00	64 - 72
140 - 159	8	80	100.00	73 - 80
Total	80			
Missing value (No response)	2	82		

The frequency table given above is an improvement over the arrangement of figures, which were just listed or entered in a register, or kept in a file, or listed in an array form for each respondent, as it presents a clear idea of the data. This type of representation of frequencies is called a grouped frequency distribution. In the above example, 80 observations were divided into eight groups. The groups are known as class limits or class intervals. The difference between upper and lower limit is called the width of the class. The cumulative frequency corresponding to a class is the total number of observations less than or equal to the upper limit of that class, i.e., it is a total of all frequencies up to and including that class.

Instead of presenting data in frequency tables using absolute numbers it is often better to calculate percentages. Percentages may also be called Relative

Frequencies. Percentages *standardize* the data, which means that they make it easier to compare them with similar data obtained in another sample of different size or origin.

Note

One should be cautious when calculating and interpreting percentages where the total number is small because one unit more or less would make a big difference in terms of percentages. As a general rule, percentages should not be used when the total is less than 30. Therefore, it is recommended that the number of observations, or total cases studied, should always be given together with the percentage.

After having gone through the concept of data, answer the following questions given in *Check Your Progress-1*.

Check Your Progress - 1

Note: (a) Write your answer in about 50 words.

(b) Check your answer with possible answers given at the end of the unit.

1. What are the different types of data?

.....

.....

.....

.....

2. What are the two types of variables?

.....

.....

.....

.....

1.4 MEASURES OF CENTRAL TENDENCY

Frequency distributions and histograms provide useful ways of looking at a set of observations of a variable. In many circumstances, it is essential to present them to understand the pattern or trend in the data. However, if one wants to further summarize a set of observations, it is often helpful to use a measure which can be expressed in a single number. First of all, one would like to have the measures of location or measures of central tendency of the distribution. The three measures used for this purpose are the mean, median, and mode.

1.4.1 Mean

The mean (or arithmetic mean) is also known as the average. It is calculated by adding the values of all the observations and dividing by the total number of observations. Note that the mean can only be calculated for numerical data.

$$\bar{X} = \frac{\sum x}{N}$$

Example 1: The measurements of the heights in centimetres of seven girls are given below.

S.No.	Height (cm)
1	141
2	141
3	143
4	144
5	145
6	146
7	155

A total of seven measurements = 1015 cm. The mean = (1015/ 7), which is 145 cm.

In case of frequency, or grouped data, and in this situation it is bit difficult mean is calculated as follows:

Example 2: Measurements of the heights in centimetres of 15 girls are given below

S.No.	Height X_i	Frequency f_i	$f_i \times X_i$
1	141	2	282
2	143	3	429
3	144	4	576
4	145	3	435
5	146	2	292
6	155	1	155
Total		15	2169

Formula for calculating mean:
$$\bar{X} = \frac{\sum f_i \times X_i}{\sum f_i} = \frac{2169}{15} = 144.6$$

Note

Note 1: In the case of the grouped frequency data given in table 1.1, the midpoint of the interval will become X_i and a similar procedure for computing the mean can be followed as shown above.

Note 2: However, in the case of open interval data, such as ‘less than 19’ (denoted as <19) or ‘more than 159’ (denoted as > 159), it is not possible to fix a midpoint, and, therefore, the mean cannot be calculated. As a consequence, we use the median in place of the mean, which is explained in the next section.

1.4.2 Median

The median is a value that divides a distribution into two equal halves. The median is useful when the data is in ordinal scale, i.e., some measurements are much bigger or much smaller than the other measurement values. The mean of such data will be biased toward these extreme values. Thus, the mean is not a good measure of distribution, in this case. The median is not influenced by extreme values. The median value, also called the central or halfway value, is obtained in the following way:

- List the observations in order of magnitude (from the lowest to the highest value, or vice versa).
- Count the number of observations = n .
- The median value is the middle value, if n is odd, it will be the $(n+1)/2^{\text{th}}$ term and if n is even, it will be equal to the mean of two middle values, {i.e., $(n/2)$ and the next value}

Example 3:

Case 1: The weights of 7 women are:

S.No.	Weight of women (kg)
1	40
2	41
3	42
4	43
5	44
6	47
7	72

The median value is the value belonging to observation number $(7 + 1)/2$, which is the fourth one value: 43 kg.

Case 2: If there are 8 observations:

S.No.	Weight of women (kg)
1	40
2	41
3	42
4	43
5	44
6	47
7	49
8	72

The median would be the average of ' $(n/2)^{\text{th}}$ or the 4th value i.e. 43' and 'next value, i.e., 44'; the median in this case would be $(43+44)/2 = 43.5$ kg}.

Calculation of Median for Grouped Data

We can use the **grouped data** of the table 1.1 of section 1.3.1, ‘Distribution of clinics according to number of patients treated for malaria in one month’ for **calculation of median**, which is given below

Table 1.2: Distribution of clinics according to number of patients treated for malaria in one month

Number of patients	Number of clinics	Cumulative frequency
0 - 19	5	5
20 - 39	8	13
40 - 59	10	23
60 - 79	11	34
80 - 99	19	53
100 - 119	10	63
120 - 139	9	72
140 - 159	8	80
Total	80	

Step1: The total of frequency is first divided by 2, i.e., $80/2 (=40)$. The cumulative frequency 40 will correspond to the class interval (80-99). This is called the median interval.

Step2: The formula is
$$\text{Median} = L + \left[\left(\frac{N}{2} - F \right) \times d \right] / f$$

Step3: Record all values of symbol variables from the table as given below:

$L (=80)$ is the lower limit of the median interval,

$F (=34)$ is the cumulative frequency of the class, preceding to median class,

$d (=20)$ is the width of class interval,

$f (=19)$ is the frequency of median class.

Step4: Replace the symbol values with numeric values as noted in step3 in the formula,

Therefore, $\text{Median} = 80 + [(40-34) \times 20] / 19 = 80 + 6.32 = 86.32$ patients.

1.4.3 Mode

The mode is the most frequently occurring value in a set of observations. The mode is not very useful for numerical data that are continuous. It is most useful for numerical data that have been grouped. The mode is usually used to find the norm among populations and is calculated when the calculation of mean and median is inappropriate, viz., the average shoe size of the Indian population, standard birth weight, etc. The mode can also be used for categorical data, whether they are nominal or ordinal.

In **Example 1** (height of 7 girls) the mode is 141.

In **Example 2** the mode is 144 as 4 persons are with this observation.

Calculation of Mode for Grouped Data

We will again use the grouped data given in table 1.1 to calculate the mode. The steps for computing the mode are

Step1: Formula for calculating mode, $\text{Mode} = L + \left(\frac{f_m - f_1}{2f_m - f_1 - f_2} \right) \times h$

Step2: Find the class interval with the largest frequency, which is also called the modal class. In this case, the class interval '80-99' has the maximum frequency, equal to 19.

Step3: Record all values of the symbol variables used in the formula.

$L =$ (80) is the lower limit of the modal class,

$f_m =$ (19) is the modal class frequency (i.e. interval having maximum frequency),

$f_1 =$ (11) is the frequency of the class interval preceding the modal class,

$f_2 =$ (10) is the frequency of the class succeeding the modal class, and

$h =$ (20) is the width of the interval.

Step4: Replace the symbol values with the numeric values as noted in step3 in the formula.

The calculated value of mode is:

$$= \text{Mode} = 80 + \left(\frac{19 - 11}{2 \times 19 - 11 - 10} \right) \times 20$$

$$= 80 + 9.41 = 89.41 \text{ patients}$$

1.4.4 Relationship between Mean, Median, and Mode

In normal data, mean, median, and mode are same. However, in a moderately skewed (non normal) distribution, $\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$.

In summary, the mean, the median, and the mode are all measures of central tendency or measures of location. The mean is most widely used. It contains more information because the value of each observation is taken into account in its calculation. However, the mean is strongly affected by values far from the centre of the distribution, while the median and the mode are not. The calculation of the mean forms the beginning of more complex statistical procedures, like, correlation and regression, etc., to describe and analyze data. In general, as the skewness increases, the mean and median move away from the mode. If the mean is less than the median, the data is skewed to the left; and if it is greater than median, the data is skewed to right. The choice of central tendency, therefore, depends upon the type and distribution of data. Nowadays, with the easy availability of scientific

calculators and computers (Excel and other statistical software), the calculation of mean, median, mode, etc., has become very simple.

1.5 MEASURES OF DISPERSION

The mean, median, and mode are measures of the central tendency of a variable, but they do not provide any information of how much the measurements vary or are spread. This section will describe some common measures of variation (or variability), which in statistical text books are often referred to as measures of dispersion. Measures of dispersion or variability of a data give an idea up to which extent the values are clustered or spread out. In other words, it gives an idea of the homogeneity and heterogeneity of data. Two sets of data can have similar measures of central tendency but different measures of dispersion. Therefore, measures of central tendency should be reported along with measures of dispersion. There are various measures of dispersion. They are discussed below:

1.5.1 Range

It is the simplest measure of dispersion. This can be represented as the difference between maximum and minimum values, or simply, as the maximum and minimum values for all observations.

Example 4: If the weights of 7 women were

S.No.	Weight of women (kg)
1	40
2	41
3	42
4	43
5	44
6	47
7	72

The range would be $72 - 40 = 32$ kg.

Although simple to calculate, the range does not tell us anything about the distribution of the values between the two extreme ones.

Example 5: If the weights of 7 other women were

S.No.	Weight of women (kg)
1	40
2	46
3	50
4	55
5	60
6	65
7	72

The range would also be $72 - 40 = 32$ kg, although the values are very much different from those of the previous example.

1.5.2 Percentiles

A second way of describing the variation or dispersion of a set of measurements is to divide the distribution into percentiles (100 parts). As a matter of fact, the concept of percentiles is just an extension of the concept of the median, which may also be called the 50th percentile. Percentiles are points that divide all the measurements into 100 equal parts. The 30th percentile (P30) is the value below which 30% of the measurements lie. The 50th percentile (P50), or the median, is the value below which 50% of the measurements lie. To determine percentiles, the observations should be first listed from the lowest to the highest just like when finding the median. However, in case of grouped data, percentile can be calculated on similar lines of calculating the median.

The concept of percentiles is used by nutritionists to develop standard growth charts for specific countries from a representative sample of children whose weight and height are measured according to their age in months.

1.5.3 Mean Deviation

It is the average of deviation from arithmetic mean.

Mean Deviation = $\frac{\sum |x - \bar{x}|}{N}$ where $| |$ denotes Mod, considering all differences 'as positive' or 'in absolute value'.

1.5.4 Standard Deviation

Standard deviation (sd) is always reported with mean. It denotes (approximately) the extent of variation of values from the mean. Mathematically, it is square root of variance ($sd = \sqrt{\text{Variance}}$). All the deviation values $(X_i - \bar{X})$ in a data set are calculated by subtracting the mean, $\bar{X}$, of the data from each observation X_i . Variance is the mean of the sum of squares of all the deviation scores of a data. Mathematically this is written as follows.

$$\text{Variance} = s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1},$$

$$\text{An easy to calculate formula is: } s^2 = \frac{1}{n-1} \sum X_i^2 - \left(\frac{\sum X_i}{n} \right)^2$$

$$sd = \sqrt{\text{Variance}}$$

$\bar{X}$ = Mean

n = Number of observations

Large values of variance and standard deviation represent higher variability in the data and vice versa. If the value of variance is equal to 'zero', it represents no variability in the data. To obtain the standard deviation of a set of measurements one has to carry out the following steps:

- i) Calculate the mean of all the measurements.
- ii) Calculate the difference between each individual measurement and the mean.

- iii) Square all these differences.
- iv) Take the sum of all squared differences.
- v) Divide this sum by the number of measurements minus one.
- vi) Finally take the square root of the value obtained (in order to get back to the same unit of measurement).

Example 6: Suppose you need to calculate the standard deviation of 2, 4, 6, 8, 10 and 12.

S. No.	X	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
1	2	-5	25
2	4	-3	9
3	6	-1	1
4	8	1	1
5	10	3	9
6	12	5	25
Sum	42		70
$\bar{X}$	7		

$$\text{Variance} = s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1} = \frac{70}{5} = 14$$

(sd): $\sqrt{14} = 3.74$.

Fortunately many pocket calculators can do this calculation for us, but it is still important to understand what it means. In the case of grouped data, the mid value of the interval may be taken as observation value and the above procedure can be followed.

In the above sections you studied about the measures of central tendency and the measures of dispersion. Now try and answer the questions in *Check Your Progress-2*.

Check Your Progress - 2

Note: (a) Write your answer in about 50 words.

(b) Check your answer with possible answers given at the end of the unit.

1. What are the different measures of central tendency?

.....

.....

.....

.....

2. What are the different measures of dispersion?

.....

.....

.....

.....

1.6 HYPOTHESIS TESTING AND INFERENCIAL STATISTICS

1.6.1 Understanding True Difference

The analysis and interpretation of the results of our study must be related to the objectives of study. It is important to tabulate the data in univariate and/ or bi-variate or multivariate tables appropriate to the research objectives. We may find some interesting results. For example, in a study on nutrition, we find that 30% of the women included in the sample are anaemic as compared to only 20% of the men. How should we interpret this result?

- The observed difference of 10% might be a true difference, which also exists in the total population from which the sample was drawn.
- The difference might also be due to the chance; in reality there is no difference between men and women, but the sample of men just happened to differ from the sample of women. One can also say that the observed difference is due to sampling variation.
- A third possibility is that the observed difference of 10% is due to defects in the study design (also referred to as Bias). For example, we only used male interviewers, or omitted a pre-test, so we did not discover that anemia is a very important topic for women which require a female investigator.

If we feel confident that an observed difference between two groups cannot be explained by bias, we would like to find out whether this difference can be considered as a true difference. We can only conclude this only if we can rule out chance (sampling variation) as an explanation. We accomplish this by applying a test of significance. A test of significance estimates the likelihood that the observed result (e.g., a difference between two groups) is due to chance or real. In other words, a significance test is used to find out whether a study result, which is observed in a sample, can be considered as a result which indeed exists in the study population from which the sample was drawn.

1.6.2 Tests of Significance

Different sets of data require different tests of significance. Throughout this section, two major sets of data will be distinguished.

- Two (or more) groups, which will be compared to detect differences. (e.g., men and women, compared to detect differences in anemia.)
- Two (or more) variables, which will be measured in order to detect if there is an association between them. (e.g., between anemia and income.)

(i) How Tests of Significance Work

The reasoning behind significance tests is the same, no matter whether a researcher is comparing two groups for differences or whether s/he is measuring two variables to detect possible associations.

We will first concentrate on the **comparison of groups**.

- Suppose you observed a difference between two groups in your sample.

- You want to know whether this observed difference between the two groups represents a *real* difference in the study population from which the sample was drawn, or whether it just *occurred by chance* (due to sampling variation).
- To find out, you determine how likely this difference could have occurred by chance, if in the population no difference exists between the two groups.

In any study that is looking for differences between groups or associations between variables, the likelihood or probability (p) of observing a certain result by chance has to be calculated by statistical tests.

(ii) How to state Null (H_0) and Alternative (H_1) Hypothesis:

In statistical terms the assumption that no real difference exists between groups in the total study (target) population (or, that no real association exists between variables) is called the Null Hypothesis (H_0). The Alternative Hypothesis (H_1) is that there exists a difference between groups or that a real association exists between variables.

Null Hypothesis is also known as the hypothesis of no difference. Examples of null hypotheses are

- There is no difference in the incidence of measles between vaccinated and non-vaccinated children.
- There is no difference between the alcohol consumption of male and female.
- There is no association between families' income and malnutrition in their children.

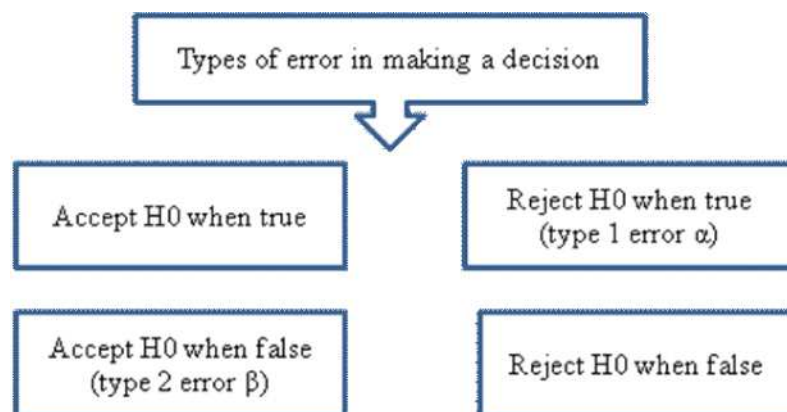
However, if by testing the hypothesis we find that the result is statistically significant, we reject the Null Hypothesis (H_0) and accept the Alternative Hypothesis (H_1) that there is real difference between two groups, or a real association between two variables. Examples of alternative hypotheses (H_1) are:

- There is a difference in the incidence of measles between vaccinated and non-vaccinated children.
- Males drink more alcohol than females.
- There is an association between families' income and malnutrition in their children.

Be aware that 'statistically significant' does not mean that a difference or an association is *of practical importance*. The tiniest and most irrelevant difference will turn out to be statistically significant if a large enough sample is taken. On the other hand, a large and important difference may fail to reach statistical significance if too small a sample is used.

(iii) The Concept of Type I and Type II Error

The testing of hypothesis will lead the investigator to take one of the following decisions:



Of these, two decisions (Accepting H_0 when it is true and Rejecting H_0 when it is false) are the right decisions to make whereas the other two lead to error.

Type I error (α): We reject the null hypothesis when it is true, or false positive error, or type I error ' α ' (called alpha or the level of significance). Eg. A Pharma company introduces a new drug in the market claiming that it is more effective in treating diabetics than the drug already existing in the market.

In the above example, type I error would mean that the effects of two drugs were found to be different by statistical analysis, when, in fact, there was no difference between them.

Type II error (β): We accept the null hypothesis when it is false or false negative error; or simply, type II error ' β ' (called beta). In the above example, type II error would mean that the effects of two drugs were not found different by statistical analysis, when in fact there was difference.

We would like to carry our test, so as to minimize both types of errors simultaneously, but this not possible in a given fixed sample size. In fact decreasing one type of error may very likely increase the other type. In practice, we keep type I error (α) fixed at a specified value (i.e., at 1% or 5%).

Till now you have read about the various concepts of hypothesis testing. Now try and answer the following questions in *Check Your Progress-3*.

Check Your Progress - 3

Note: (a) Write your answer in about 50 words.

(b) Check your answer with possible answers given at the end of the unit.

1. What do you understand by a null and an alternative hypothesis?

.....

.....

.....

2. What are type I and type II errors?

.....

.....

.....

1.7 STATISTICAL TESTS

Depending on the aim of your study and the type of data collected, you have to choose appropriate tests of significance. Before applying any statistical test, state the null hypothesis in relation to the data to which the test is being applied. This will enable you to interpret the results of the test. The following sections will explain how you will choose an appropriate statistical test to determine differences between groups or associations between variables.

1.7.1 Determining Significant Difference between Groups

There are several issues one should consider, while deciding, which test to use to determine differences between groups. It is necessary to determine whether the data are categorical (nominal and ordinal) or continuous (numerical).

Under both of these categories, you also need to decide whether you have paired or unpaired observations. In paired observations, individual observations in one data set (e.g., experimental) are matched with individual observations in another data set (e.g., controls).

For continuous data (in interval, or, ratio scale), the choice of an appropriate significance test depends on whether you are comparing two groups or more than two groups. The z-test, referred to as the standard normal variate, and the t-test, also referred to as the Student's t-test, are used for numerical data of continuous nature, when comparing the means of two groups. The chi-square test is used for categorical data, when comparing proportions of events occurring in two or more groups.

Steps to Follow for testing a Hypothesis

Step1: Formulate the null hypothesis

Step 2: Formulate the alternative hypothesis

Step 3: Choose the level of significance of the test

Step 4: Choose an appropriate test statistic

Step 5: Compute the observed value of the test statistic

Step 6: Compare the calculated value of the statistic with the tabled value

Step 7: Accept or Reject the null hypothesis

Here we will confine our discussion to four types of tests:

- (i) χ^2 test
- (ii) Z- test
- (iii) t-test
- (iv) f-test

1.8 CHI-SQUARE TEST

Suppose that in a cross-sectional study of the factors affecting the utilisation of health care centres, you found that 64% of the women who lived within 10 kilometres of the centres came for health checkups, compared to only

47% of those who lived more than 10 kilometres away. This suggests that health care centre is used more often by women who live close to the centres. The complete results are presented in Table 1.3.

Table 1.3: Utilization of health care centre by women living far from, and near the clinic

Distance from health care centre	Used health care centre	Did not Use health care centre	Total
Less than 10km	51 (64%)	29 (36%)	80 (100%)
10km or more	35 (47%)	40 (53%)	75 (100%)
Total	86	69	155

From the table we conclude that there seems to be a difference in the use of health care centre between those who live close to, and those who live far from, the centre (64% versus 47%). We now want to know if this observed difference is statistically significant or not. The chi-square (χ^2) test, used to test the statistical significance, is given below.

$$\text{Chi-square} = \chi_{df}^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad \sum \text{ (the summation sign, also called sigma)}$$

directs you to add together each value of $(O_i - E_i)^2 / E_i$ for all the 'k' cells of the table. Where, O_i and E_i are the observed and expected frequency of each cell and 'df' is the degree of freedom.

The alternative simplified formula for calculating chi-square is:

$$\chi^2 = \sum \frac{O_i^2}{E_i} - N, \text{ where } N \text{ is the total of observed frequency.}$$

These are the steps to follow:

Formulating the Hypothesis:

- **H₀:** There is no difference in the utilization of health care centre among women living within distance of less than and more than 10 kms
- **H₁:** There is difference in the utilization of health care centre among women living within distance of less than and more than 10 kms

Test Statistic: chi-square (χ^2) Test

Calculating the Chi-square χ^2 value

- (1) Calculate the expected frequency (E_i) for each cell. To find the expected frequency E_i of a cell, you multiply the row total by the column total

$$\text{of the cell and divide by the grand total: } E = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}.$$

- (2) For each cell, subtract the expected frequency E_i from the observed frequency O_i , $(O_i - E_i)$, square the difference of $(O_i - E_i)$ and divide it by the corresponding expected frequency E_i . (You may skip this step and follow the next step 3 as an alternative step).
- (3) Alternatively, for each cell, simply square the observed frequency O_i and divide it by the corresponding expected frequency, E_i .
- (4) Find the sum by adding the values calculated in step (3) for all the cells and subtract N (total of observed frequency) from the sum.

For a 2x2 table the formula is: $\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \frac{(O_4 - E_4)^2}{E_4}$,

The alternative formula $\chi^2 = \left(\frac{O_1^2}{E_1} + \frac{O_2^2}{E_2} + \frac{O_3^2}{E_3} + \frac{O_4^2}{E_4} \right) - N$

Using a Chi-square table

The calculated chi-square value has to be compared with a theoretical chi-square value in order to determine whether the null hypothesis is rejected or not. **Annex I** contains a table of theoretical chi-square values.

- (1) First you must decide what significance, or alpha, level you want to use (α value). We usually take 0.05.
- (2) Then, the degrees of freedom have to be calculated. With the chi-square test the number of degrees of freedom is related to the number of cells, i.e., the number of groups you are comparing. The number of degrees of freedom is found as:

$$\text{degrees of freedom} = (\text{Number of rows} - 1) \times (\text{Number of columns} - 1) = (r - 1) \times (c - 1)$$

For a simple 2 x 2 table, the number of degrees of freedom is 1: d.f.
= $(2-1) \times (2-1) = 1$

- (3) Then the chi-square value belonging to the α -value and the number of degrees of freedom are located in the table.

Interpreting the results

- If the calculated chi-square value is equal to or larger than the chi-square value from the table then we reject the null hypothesis and conclude that there is a statistically significant difference between the groups.
- If the calculated chi-square value is smaller than the chi-square value from the table, then we accept the null hypothesis and conclude that the observed difference is not statistically significant (do not differ significantly).

Let us now apply the chi-square test to the data given in table 1.5 (utilization of health care centre). This gives the following result:

Calculating the Chi-square χ^2 value

First the expected frequencies for each cell are computed for calculating the Chi-square value as follows:

Distance from health care centre	Used health care centre	Did not Use health care centre	Total
Less than 10km	$E_1 = \frac{86 \times 80}{155} = 44.4$	$E_2 = \frac{69 \times 80}{155} = 35.6$	80
10km or more	$E_3 = \frac{86 \times 75}{155} = 41.6$	$E_4 = \frac{69 \times 75}{155} = 33.4$	75
Total	86	69	155

Note that the expected frequencies refer to the values we would have expected, given the total numbers of 80 and 75 women in the two groups, if the null hypothesis, there is **no** difference between the two groups, were true. Now the chi-square value can be calculated:

$$\text{Chi-square} = \frac{(51 - 44.4)^2}{44.4} + \frac{(29 - 35.6)^2}{35.6} + \frac{(35 - 41.6)^2}{41.6} + \frac{(40 - 33.4)^2}{33.4}$$

$$(\chi^2) = 0.98 + 1.22 + 1.05 + 1.30 = 4.55$$

With the alternative formula, the value of chi-square is same and is given below.

$$\chi^2 = \left(\frac{51^2}{44.4} + \frac{29^2}{35.6} + \frac{35^2}{41.6} + \frac{40^2}{33.4} \right) - 155 = 4.56$$

Using a Chi-square table

The number of degrees of freedom (d.f.) for 2x2 table is 1. Use the table of chi-square values in Annex 1. We decided, beforehand, on a level of significance of 5% (α -value = 0.05).

As the number of d.f. is 1, we look along that row in the column where $p = 0.05$. This gives us the tabulated value of 3.84.

Interpreting the result

Our calculated value of chi-square 4.55 is larger than 3.84. Hence we reject the null hypothesis.

We can now conclude that the women living within a distance of 10 km from the clinic utilise antenatal care significantly more often than the women living more than 10 km away.

Table 1.5 indicates that 64% of the women living within a distance of 10 km from the clinic used antenatal care during pregnancy, compared to only 47% of women living 10 km or further away from the nearest clinic. This difference is statistically significant (chi-square = 4.55; $p < 0.05$).

1.9 F-test

The F-test is used either for testing the hypothesis about the equality of two population variances or the equality of two or more population means. The ratio of two sample variances follows F distribution.

i. Test of Equality of Two Population Variances

Let there be two normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$.

Where μ_1 and μ_2 are the population means and σ_1^2 and σ_2^2 are the population variances.

We use F test to test the hypothesis

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Let an independent sample of size n_1 be selected from population $N(\mu_1, \sigma_1^2)$, and of size n_2 from population $N(\mu_2, \sigma_2^2)$.

The test statistic:

$$F_{k1, k2} = \frac{S_1^2}{S_2^2}$$

Where $k1 = (n_1 - 1)$ and $K2 = (n_2 - 1)$.

As a norm, the larger variance is taken in the numerator and the degrees of freedom corresponding to it is denoted as k_1 .

Interpreting the results:

If the calculated value of F is greater than the tabled value, reject H_0 , if not, accept H_0 .

Example 7: Life expectancy in 9 regions of Brazil in 1900, and in 11 regions of Brazil in 1970 was as given in the table below:

Regions	Life Expectancy (Years)	
	1900	1970
1	42.7	54.2
2	43.7	50.4
3	34.0	44.2
4	39.2	49.7
5	46.1	55.4
6	48.7	57.0
7	49.4	58.2
8	45.9	56.6
9	55.3	61.9
10		57.5
11		53.4

We want to confirm whether the variation in life expectancy in various regions in 1900 and in 1970 is the same or not. Let the populations in 1900 and 1970 be considered $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively.

We use F test to test the hypothesis

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Calculating the test statistic

First we calculate S_1^2 and S_2^2 .

$$S_1^2 = \frac{1\{\sum x_{1i}^2 - \frac{(\sum x_{1i})^2}{9}\}}{8}$$

$$= 37.848$$

$$S_2^2 = \frac{1\{\sum x_{2i}^2 - \frac{(\sum x_{2i})^2}{11}\}}{10}$$

$$= 26.606$$

$$F = \frac{37.848}{26.606}$$

$$= 1.603$$

Tabled value

$$F_{\text{tab}} = 3.07$$

Interpreting the results

The tabled value of F is 3.07. Since $F_{\text{cal}} < F_{\text{tab}}$, we accept the null hypothesis. It means that the variation in life expectancy in various regions of Brazil in 1900 and in 1970 is the same.

1.10 Z -TEST

To test the hypothesis about a population mean or two population means when the sample size is large (>30) and population variances are known, we use the Z-test.

1.10.1 Testing of Significance of Difference between Two Proportions (Two Large Samples)

Let there be two large samples drawn from one population or from two populations, having same variance. The populations are distributed normally. The proportions, p_1 and p_2 are the two proportions of an event from the two samples (>30), which are compared for their difference. We may apply

$$\text{the formula: } Z = \frac{(p_1 - p_2)}{\sqrt{P \times Q \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \text{ where, } P = \frac{n_1 \times p_1 + n_2 \times p_2}{n_1 + n_2}, Q$$

$$= 1 - P$$

The P is the mean proportion of success of the two proportions, p_1 and p_2 , and n_1 and n_2 are the respective sample sizes.

The above calculated value of 'z' is compared with the tabulated value of 'z-normal variate' for ' $\alpha = 0.05$ ', i.e., at 5% level of significance, which is 1.96 and the value of z for ' $\alpha = 0.01$ ', i.e., at 1% level significance is 2.58.

Example 8: To test the conjecture of the management that 60 % of employees prefer the new bonus scheme, a sample of 150 employees was drawn, and their opinion was taken, whether they favoured it or not. Only 55 employees out of 150 favoured the new bonus scheme.

Thus, we test the hypothesis, $H_0: P = 0.60$

$$H_1: P \neq 0.60$$

Calculating the test statistic:

$$Z = \frac{0.367 - 0.60}{\sqrt{\frac{0.60 \cdot 0.40}{150}}} \quad \text{Since } p = 55/150 = 0.367$$

$$\frac{\sqrt{0.60 \cdot 0.40}}{150}$$

$$Z = - 11.65$$

$$| Z | = 11.65$$

The Table value

At $\alpha = 0.01$, $z_{\text{tab}} = 2.58$.

Interpreting the results

As $Z_{\text{cal}} > Z_{\text{tab}}$, hence H_0 is rejected. It means that 60 percent of the employees do not favour the new bonus scheme.

1.10.2 Testing of Significance of Difference between Means of Two Large Samples (Continuous Data)

In case of large samples where population variances are known, we test the equality of two population means using z – test.

Hypothesis:

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

Calculating the test statistic:

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

where, $\bar{x}$ is the sample mean and σ is the standard deviation.

The Table value

The value of z (α) for comparing the calculated test statistics is taken as 1.96 at 5% level of significance (α); and it is 2.58 at 1% level of significance (α). Here, the sample size is treated as large and, therefore, the degree of freedom plays no role, unlike in the t-test.

Example 9: The table below gives the total income in thousand rupees per year of 36 persons selected randomly from a particular class of people.

Income (thousand Rs)					
6.5	10.5	12.7	13.8	13.2	11.4
5.5	8.0	9.6	9.1	9.0	8.5
4.8	7.3	8.4	8.7	7.3	7.4
5.6	6.8	6.9	6.8	6.1	6.5
4.0	6.4	6.4	8.0	6.6	6.2
4.7	7.4	8.0	8.3	7.6	6.7

On the basis of the sample data, can it be concluded that the mean income of a person in this class of people is Rs. 10,000 per year?

We have to test the hypothesis $H_0 : \mu = 10,000$

$H_1 : \mu \neq 10,000$

Calculating the test statistic:

Since the sample size is 36, we will use a normal test for which the test statistic is

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Now we compute $\bar{x}$ and σ .

$$\bar{x} = 280.7/36 = 7.80$$

$$\begin{aligned}\sigma^2 &= 1/35 \{2368.75 - (2368.75 - (280.7)^2/36)\} \\ &= 5.14\end{aligned}$$

$$\sigma = 2.27$$

$$\begin{aligned}Z &= (7.80 - 10) / (2.27/\sqrt{36}) \\ &= -5.81\end{aligned}$$

$$|Z| = 5.81$$

The Table value: $Z_{\text{tab}} = 1.96$

Interpretation: Since $Z_{\text{cal}} > Z_{\text{tab}}$, reject H_0 . It means that the average annual income is less than ten thousand rupees.

Assumption for use of z statistics: The assumption for using the z statistics is that the parent population, from where samples have been drawn, should be normal. The z statistics presumes that the population variances (σ_1^2 and σ_2^2) of the parent populations are known and, therefore, the z statistics for testing

significance of difference between means is defined as
$$z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Since the population variances are presumed to be known, therefore the samples drawn from a population should also be normally distributed. It

is, therefore, pertinent to verify whether the samples satisfy the criteria of normality or not. If the samples drawn from the population are sufficiently large, then, one can estimate the population variances from the samples itself. As in practice population variances are not known, the researcher has to estimate the population variances from the samples (sd_1^2 and sd_2^2) itself and

use the z statistics $\frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}}}$. In z statistics, we refer to the tables of normal

area curve, where degrees of freedom do not play any role. The computation of z statistics is almost same except for computing of standard error (SE),

which is given as $\sqrt{\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}}$.

1.11 t-Test

1.11.1 Testing the Significance of Independent Samples from Two Groups for Continuous Data

Example 10: It has been observed that in a certain province the proportion of women, joining the army, is very high. A study is, therefore, conducted to discover why this is the case. The height of women is supposed to be the contributory factor; the researcher may want to find out if there is a difference between the mean height of women in this province who preferred joining army and of those who opted for other services. The null hypothesis would be that there is no difference between the mean heights of the two groups of women. Suppose the following results were found:

Table 1.4: Mean heights of women as per type of service

Type of Service	Sample size	Mean height in cm	Standard deviation
Joined army (Group1)	$n_1 = 60$	$\bar{X}_1 = 156$	$sd_1 = 3.1$
Other Services (Group2)	$n_2 = 52$	$\bar{X}_2 = 154$	$sd_2 = 2.8$

The mean height for each of the two samples was calculated and compared, using the t-test, to determine whether there was a difference.

These are the steps to follow, in determining whether the difference is statistically significant:

Hypothesis:

H_0 : There is no significant difference between the heights of the women joining army and other services.

H_1 : There is difference between the heights of the women joining army and other services.

Test Statistic: A t-test would be the appropriate way to determine whether the observed difference of 2 cm can be considered statistically significant.

$$t_{df} = \frac{(\bar{X}_1 - \bar{X}_2)}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{degrees of freedom, } df = n_1 + n_2 - 2$$

$$\text{Pooled Variance } s^2 = \frac{(n_1 - 1)sd_1^2 + (n_2 - 1)sd_2^2}{n_1 + n_2 - 2}$$

To calculate the t-value, you need to follow these steps:

- (1) **Difference between the means.** In the above example the difference is $(156 - 154 = 2 \text{ cm})$.

$$\bar{X}_1 - \bar{X}_2 = 2$$

- (2) **Calculate the standard deviation (square root of variance S^2)** of all observations pooled together for both the samples.

In case the standard deviations for each of the study groups are given or have been calculated, then compute the pooled variance of samples (S^2) as given:

$$S^2 = \frac{59 \times 9.61 + 51 \times 7.84}{59 + 51} = \frac{966.83}{110} = 8.80$$

Standard deviation (of pooled sample) $S = 2.96$

- (3) **Calculate the standard error of the difference between the two means**

$$S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad SE = 2.96 \times 0.01895 = 0.5608$$

- (4) Finally, divide the difference between the means by the standard error of the difference. The value now obtained is called t-

$$\text{value. } t = \frac{2}{0.5608} = 3.57$$

The Table value

Once the t-value has been calculated, you will have to refer to a t-table, from which you can determine whether the null hypothesis is rejected or not. Annex II contains a t-table.

- (1) First, decide which significance level (α value) you want to use. Usually we choose a significance level of 0.05.
- (2) Second, determine the number of degrees of freedom for the test being performed. For student's t-test the number of degrees of freedom is calculated as the sum of the two sample sizes minus 2. Thus, for Example 1.6, it is calculated as follows:

The number of degrees of freedom is: $d.f. = 60 + 52 - 2 = 110$

- (3) Third, the t-value belonging to the ' α ' value (the significance level we chose) and the degrees of freedom are located in the table.

In our example we look up the t-value belonging to $\sigma = 0.05$ and d.f. = 120 (nearest to 110 in the table) and we find it is 1.98.

Interpreting the result

In our example the t-value calculated is 3.6, which is larger than the t-value derived from the table (1.98). We, therefore, reject the null hypothesis and conclude that the observed difference of 2 cm between the mean heights of women who joined army and women who opted for other services is a statistically significant difference.

1.11.2 Paired t Test

This section describes the most commonly used tests for paired observations pertaining to numerical and to nominal data. The concept of pairing or matching subjects is illustrated by the examples that follow.

Example 11: A researcher wanted to find out whether a class of students taught with audio-visual aids (AV) receives on average better grades than those who are taught without audio-visual aids. In other words, whether there is any gain in achievement or not due to use of two teaching methods. To minimize the effect of confounding variables such as *social status* and *previous knowledge of the subjects*, each student in the AV class was paired with another in the non-AV class of similar social status and knowledge level. Paired or matched observations are carried out if researcher wants to ensure through their study design that the relationship between two variables they are interested in is not confounded by another variable. They, therefore, have to sample their cases and controls in such a way that these are similar

Example 12: During a nutritional survey, a quality control exercise was carried out to check the agreement between two observers in measuring the children's weight. In this instance we have paired observations as we have a *set of two observations on the same child*.

When dealing with paired (matched) observations, comparison of sample means is performed by using a modified t-test known as the paired t-test. In the paired t-test, differences between the paired observations (say, Post-test minus Pre-test, or matched observations of 2nd group minus 1st group) are used instead of the observations of two sets of independent samples. The paired t-test calculates the value of 't' as

$$t_{df} = \frac{\text{mean of differences}}{\text{standard error of differences}} \quad , \quad t_{df} = \frac{\bar{d}}{sd(d)/\sqrt{n}}$$

Where, $\bar{d}$ is the mean difference of values obtained from subtracting two sets of paired observation; $sd(d)$ is the standard deviation of values obtained from difference of two sets of paired observations; n is the sample size of paired observations; the degrees of freedom ($df = n - 1$) is the number of paired observations (sample size) minus 1.

The same table of t value is used, as for the t-test for unpaired observations (see Annex II) to interpret result of the study. The use of the paired t-test is illustrated on the results of the nutritional survey referred to in previous example above. The results are given table 1.5.

Table 1.5: Results of quality control exercise during a nutritional survey

Child No.	Weight Measurement (Kg)		Difference A-B (Kg)
	Observer A	Observer B	
1	18.6	17.7	0.9
2	17.1	14.5	2.6
3	14.3	12.4	1.9
4	23.2	20.7	2.5
5	18.4	16.8	1.6
6	14.9	14.4	0.5
7	16.6	14.1	2.5
8	14.8	17.1	-2.3
9	21.5	21.2	0.3
10	24.6	21.9	2.7
11	17.4	16.6	0.8
12	15.7	13.6	2.1
13	16.1	14.5	1.6
14	12.9	11.2	1.7
15	12.3	16.0	-3.7
16	19.4	20.4	-1.0
17	19.3	17.5	1.8
18	24.8	22.2	2.6
19	14.3	15.1	-0.8
20	13.4	10.9	2.5

The null hypothesis in this study is

H : The mean difference of measurements between observers A and B would be zero

H₁: The mean difference of measurements between observers A and B would not be zero

Calculating the test statistic

The paired t-test (significance test) is calculated as follows:

- Calculate the mean difference of the measurements between A and B in the sample. This is the sum of the differences divided by the number of measurements:

$$\text{Mean difference} = \frac{21.1}{20} = 1.05$$

ii) Calculate the standard deviation of the differences and Standard Error,

$$\text{Standard deviation} = 1.77, \text{ and the standard error: } SE = \frac{1.77}{\sqrt{20}} = 0.40$$

iii) The value of 't' is the mean difference divided by the standard error:

$$t = \frac{1.05}{0.40} = 2.62$$

The Table Value

The degrees of freedom are the sample size (the number of pairs of observations) minus 1, which in this case is $(20 - 1 = 19)$. The tabled t-value at 19 degrees of freedom is 2.09

The Interpretation

$$t_{\text{cal}} > t_{\text{tab}}$$

If the calculated t-value (ignoring the sign) is larger than the value indicated in the table, the null hypothesis, stating that there is no difference, is rejected, and it can be concluded that there is a significant difference in the result of your study.

Note: Computers are helpful when dealing with large data sets. A variety of software including Excel and SPSS provides options for various statistical tests.

In this section you have read about the various tests of significance. Now try and answer the questions given in *Check Your Progress-4*.

Check Your Progress - 4

Note: (a) Write your answer in about 50 words.

(b) Check your answer with possible answers given at the end of the unit.

1. What are the important considerations that one should keep in mind while applying the chi-square test?

.....

.....

.....

.....

.....

2. What is a paired t-test and when is it used?

.....

.....

.....

.....

1.12 CORRELATION ANALYSIS

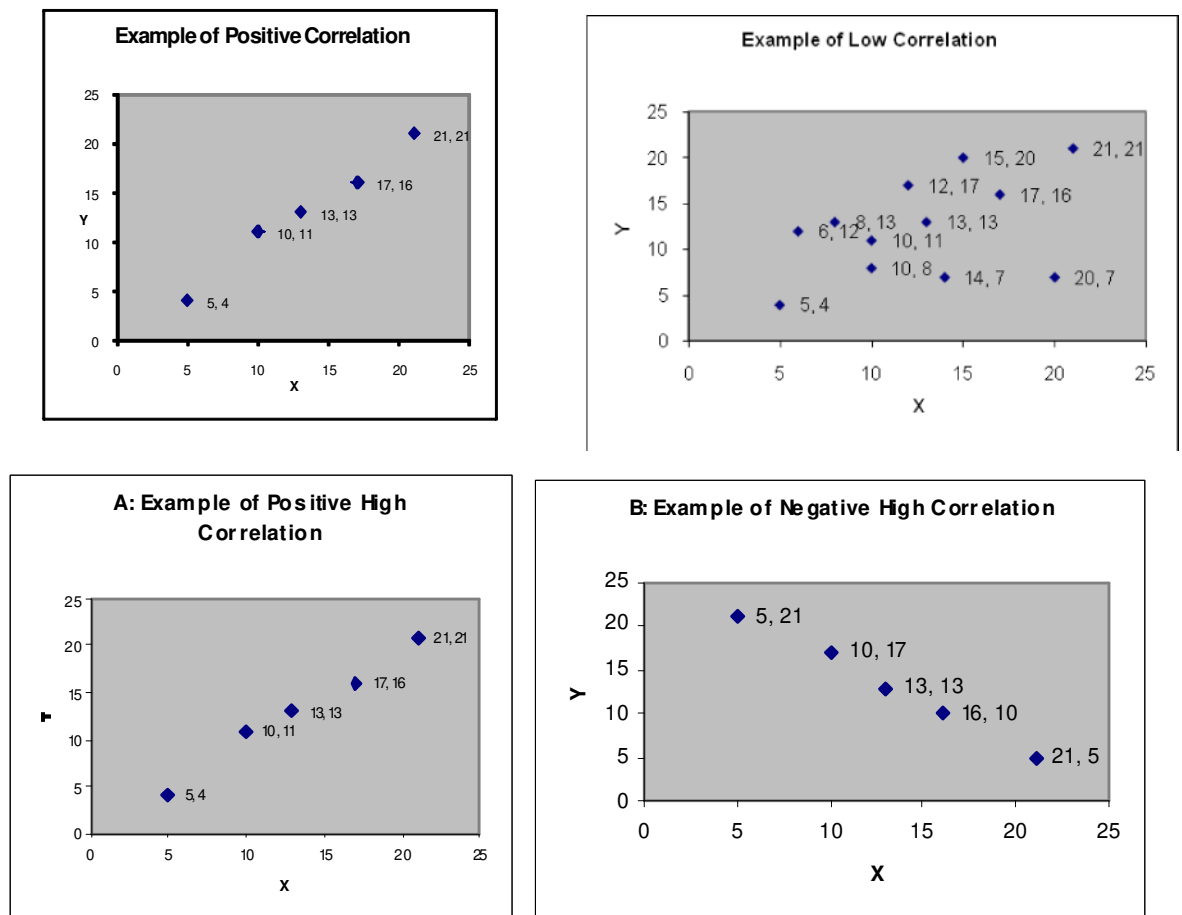
When exploring associations between variables we have to distinguish between nominal, ordinal, and continuous data. In this unit, we will examine associations between continuous data where a linear relationship is suspected.

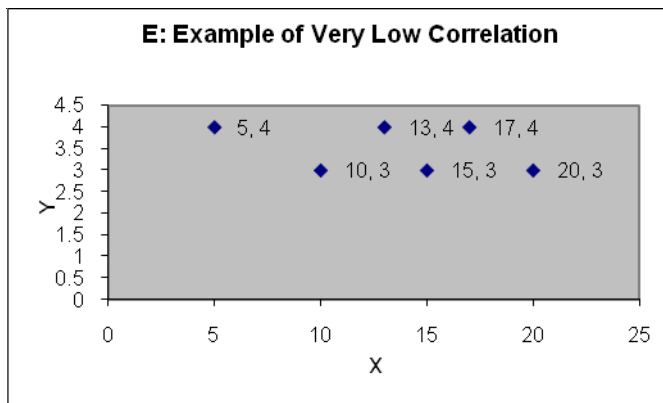
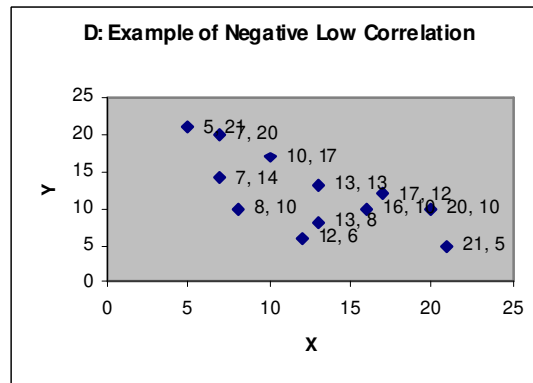
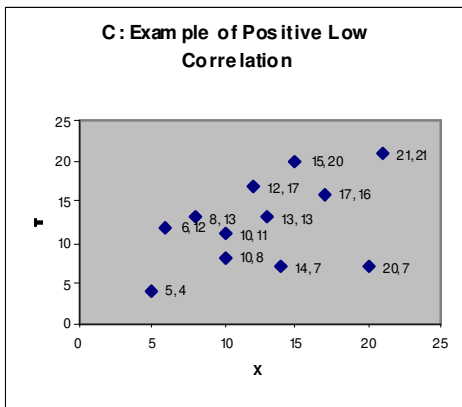
1.12.1 Understanding Correlation

Correlation is relationship between the two sets of continuous data; for example the relationship between height and body weight. Correlation statistics are used to determine the extent to which two independent variables are related and can be expressed by a measure called 'coefficient of correlation'. The correlation coefficient may be positive or negative and therefore it may vary from '-1' to '+1'. Positive correlation means that values of two different variables increase and decrease together. For example, height and weight correlate positively. Negative correlation means that if the value of one variable decreases then the value of the other variable increases (inverse relationship). For example, literacy and number of children in family may correlate negatively.

The strength of a correlation is determined by the absolute value of the correlation coefficient; the closer the value to 1, the stronger the correlation. For example, a correlation of -0.9 indicates an inverse relationship between two variables and shows a stronger relationship than that associated with a correlation of +0.2 or -0.5. Correlation between two variables is shown by scatter plot (Fig. 1.1).

Figure 1.1: Scatter Diagram showing relation between two variables





The slopes of both the lines are identical in these two examples, but the scatter around the line is much greater in the second. Clearly the relationship between variables y and x is much closer in the first diagram.

If we are interested only in measuring the association between the two variables, then Pearson's Correlation Coefficient (r) gives us an estimate of the strength of the linear association between two numerical variables. Pearson's Correlation Coefficient can either be calculated using a calculator or on computer using various analytical software. Note that in case there is curvilinear relationship, the value of r will be shown to be zero. The correlation coefficient has the following properties:

1. For any data set, r lies between '-1' and '+1'.
2. If $r = +1$, or -1 , the linear relationship is perfect, that is, all the points lie exactly on a straight line. If most of the points lie on the line, then it is very strong relationship and r is near to 1. If $r = +1$, variable y increases as x increases (i.e., the line slopes upwards). (See Diagram A.) If $r = -1$, variable y decreases as x increases (i.e., the line slopes downward). (See Diagram B.)
3. If r lies between 0 and +1, the line slopes upwards, but the points are scattered about the line. (See Diagram C.) The same is true of negative values of r , between 0 and -1, but in this case the line slopes downward. (See Diagram D.)
4. If $r = 0$, there is very low linear relationship between y and x . This may mean that there is no relationship at all between the two variables (i.e., knowing x tells us nothing about the value of y). (See Diagram E.).

1.12.2 Calculation of the Pearson's Correlation Coefficient

In a nutrition study in a large rural district, a sample of 20 children 5 years of age were weighed and their family incomes estimated. The results were as follows:

Table 1.6: Weights and family incomes of 20 children 5 years of age

Serial Number	Family Income in \$ Per year	Weight in kg	Serial Number	Family Income in \$ Per year	Weight in kg
1.	130	15.5	11.	225	18.1
2.	200	19.8	12.	95	17.4
3.	345	21.5	13.	130	17.9
4.	245	18.8	14.	330	17.0
5.	155	12.8	15.	295	18.7
6.	300	18.8	16.	170	16.0
7.	360	18.1	17.	250	18.2
8.	105	18.7	18.	355	16.4
9.	80	13.1	19.	220	15.4
10.	275	20.1	20.	175	17.6

The formula for calculating of the correlation coefficient is as follows:

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

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

$$r = \frac{\sum X_i Y_i - (\sum X_i \sum Y_i) / n}{\sqrt{\sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}} \sqrt{\sum_{i=1}^n Y_i^2 - \frac{(\sum_{i=1}^n Y_i)^2}{n}}}$$

It would be more informative to investigate whether the two variables 'family income' and 'weight of five-year-olds' are associated.

In our example of weight and family income, this would mean:

$$r = \frac{79416.5 - \frac{4440 \times 349.9}{20}}{\sqrt{1141550 - \frac{19713600}{20}}} \frac{1}{\sqrt{6210.77 - \frac{122430}{20}}}$$

$$r = 0.466$$

which is positive (indicating an upward sloping line), but away from 1 (indicating that there is plenty of scatter around the line) showing that there is a positive relation between ‘family income’ and ‘weight of five-year-olds’, however the relationship is not strong.

1.12.3 The Significance of the Correlation Coefficient

The value of r was calculated from a sample of just 20 children. The result is, therefore, subject to sampling error, and is unlikely to be equal to the true value of r , which we would obtain if we measured all 5-year-old children in this district. The question arises as to whether there really is any relationship at all between weight and income. Perhaps, in the entire population of 5-year-old children, the scatter diagram would look like diagram ‘E’, above, (no relationship between y and x) and the positive relationship in our sample occurred by chance. To assess whether this is the case we do a significance test on r . The null hypothesis is

H_0 : There is no relationship between X and Y

H_1 : There is significant relationship between X and Y

To do the test we calculate

$$t_{n-2} = r \times \sqrt{\frac{n-2}{1-r^2}}$$

We compare this value of t to tables of the t distribution with $(n - 2)$ degrees of freedom, where n is the number of observations.

$$\text{In our example: } n = 20, \quad r = 0.466, \quad t_{18} = 0.466 \times \sqrt{\frac{18}{1-0.466^2}} = 2.23$$

Therefore, using α value of 0.05, the t -table value for 18 degrees of freedom ($t_{18; 0.05}$) = 2.10. Thus the calculated t -value is more than the table value; therefore we reject the null hypothesis and accept the alternative hypothesis that the linear relationship is statistically significant.

1.13 REGRESSION ANALYSIS

A simple linear regression involves finding the equation of a line in such a way that if the observations are plotted in a scatter diagram, most of the observed points will lie close to the line and the sum of square of the distance between the line and each of the observed points will be minimum.

1.13.1 Fitting the Regression Line

Suppose there are n data points $\{y_i, x_i\}$, where $i = 1, 2, \dots, n$. The goal is to find the equation of the straight line,

$$y = \alpha + \beta x,$$

Where α = the intercept of the regression line

β = the slope of the regression line

which would provide a best fit for the data points. Here the best fit will be understood as a line that minimizes the sum of squared residuals of the linear regression model.

By using **calculus**, it can be shown that the values of α and β that minimize the objective function Q are

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{j=1}^n y_j / n}{\sum_{i=1}^n (x_i^2) - (\sum_{i=1}^n x_i)^2 / n}$$

$$= \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\text{Cov}[x, y]}{\text{Var}[x]} = r_{xy} \frac{s_y}{s_x},$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x},$$

where r_{xy} is the sample correlation coefficient between x and y , s_x is the standard deviation of x , and s_y is correspondingly the standard deviation of y . $\overline{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i$.

Sometimes people consider a simple linear regression model without the intercept term: $y = \beta x$. In such a case the OLS estimator for β will be given by the formula.

In the regression without the intercept

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}.$$

In simple regression analysis, there is one quantitative dependent variable and one independent variable. In multiple regression analysis, there is one quantitative dependent variable and two or more independent variables. For example, one may derive a formula to predict production of wheat (dependent variable) from the quantity of seed, fertilizer, pesticide, number of irrigation etc., (independent variables). Linear regression statistics finds the best fit line (line of regression) that predicts dependent variable from independent variables. Linear regression statistics is applied to data where dependent variable is continuous, but independent variables are continuous and/or categorical variable. It is beyond the scope this study material to give more detailed information about the regression analysis.

1.13.2 How to Find the Regression Equation

In the table below, the x_i column shows scores on the aptitude test. Similarly, the y_i column shows statistics grades. The last two rows show sums and mean scores that we will use to conduct the regression analysis.

Student	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	95	85	17	8	289	64	136
2	85	95	7	18	49	324	126
3	80	70	2	-7	4	49	-14

4	70	65	-8	-12	64	144	96
5	60	70	-18	-7	324	49	126
Sum	390	385			730	630	470
Mean	78	77					

The regression equation is a linear equation of the form: $\bar{y} = \alpha + \beta x$. To conduct a regression analysis, we need to solve for α and β . The computations are shown below.

$$\beta = \frac{\sum [(x_i - \bar{x})(y_i - \bar{y})]}{\sum [(x_i - \bar{x})^2]} \quad \alpha = \bar{y} - \beta \bar{x}$$

$$\beta = 470/730 = 0.644 \quad \alpha = 77 - (0.644)(78) = 26.768$$

Therefore, the regression equation is: $\bar{y} = 26.768 + 0.644x$.

1.13.3 How to Use the Regression Equation

Once you have the regression equation, using it is easy. Choose a value for the independent variable (x), perform the computation, and you have an estimated value ($\bar{y}$) for the dependent variable. In our example, the independent variable is the student's score on the aptitude test. The dependent variable is the student's statistics grade. If a student made an 80 on the aptitude test, the estimated statistics grade would be

$$\bar{y} = 26.768 + 0.644x = 26.768 + 0.644 * 80 = 26.768 + 51.52 = 78.288$$

1.13.4 Difference between Regression and Correlation

S.No.	Correlation	Regression
1	Correlation quantifies the degree to which two variables are related. You simply are computing a correlation coefficient (r) that tells you how much one variable tends to change when the other one does.	Regression finds out the best fit line for a given set of variables.
2	With correlation you don't have to think about cause and effect. You simply quantify how well two variables relate to each other. With regression, you do have to think about cause and effect as the regression line is determined as the best way to predict Y from X.	
3	With correlation, it doesn't matter which of the two variables you call "X" and which you call "Y". You'll get the same correlation coefficient if you swap the two.	With linear regression, the decision of which variable you call "X" and which you call "Y" matters a lot, as you'll get a different best-fit line if you swap the two. The line that best predicts Y from X is not the same as the line that predicts X from Y.

4	Correlation is almost always used when you measure both variables. It rarely is appropriate when one variable is something you experimentally manipulate.	With linear regression, the X variable is often something you experimentally manipulate (time, concentration...) and the Y variable is something you measure.
5	In correlation, our focus is on the measurement of the strength of such a relationship.	In regression analysis, we examine the nature of the relationship between the dependent and the independent variables.
6	In correlation, all the variables are implicitly taken to be random in nature.	In regression, at our level, we take the dependent variable as random, or stochastic, and the independent variables as non-random or fixed

Use of Data Reduction Methods in Development Research

There are also some of the data reduction techniques that are extensively used in development studies but are not within the scope of this unit. One of such technique is the Principal Component Analysis which is very commonly used. This technique is generally used when the number of explanatory variables is very high. In Principal Component Analysis, the variables are compressed to get a lesser number of variables called the principal components. There are statistical softwares which can be used to easily work out the Principal Component Analysis. For example a researcher has identified 20 variables that influence a farmer's adaptation to climate change. He may use principal component analysis which will help in grouping similar variables into one component thus reducing the number of variables. The analysis may thus identify just three or four principal components into which all these variables can be grouped. The first principal component will amount for maximum amount of variation out of the existing variables followed by the second, third and the fourth. This kind of analysis helps the researcher to identify those variables which if worked upon would increase the adaption of the farmers to climate change.

In this section you have read about correlation and regression analysis. Now try and answer the questions given in *Check Your Progress-5*.

Check Your Progress 5

Note: a) Write your answer in about 50 words.

b) Check your answer with possible answers given at the end of the unit

1. Differentiate between correlation and regression.

.....

.....

.....

1.14 LET US SUM UP

Statistics is a science that deals with the collection, organization, analysis, interpretation, and presentation data. The information or data collected may be classified as qualitative and quantitative. It may also be classified as discrete or continuous. Frequency distribution is an improved way of presenting a data. For better and more concise presentation of the information contained in a data set, the data is subjected to various calculations. If one wants to further summarize a set of observations, it is often helpful to use a measure which can be expressed in a single number like the measures of location or measures of central tendency of the distribution. The three measures used for this purpose are the mean, median, and mode. Measures of dispersion, on the other hand, give an idea about the extent to which the values are clustered or spread out. In other words, it gives an idea of homogeneity and heterogeneity of data. Two sets of data can have similar measures of central tendency but different measures of dispersion. Therefore, measures of central tendency should be reported along with measures of dispersion. The measures of dispersion include range, percentiles, mean deviation and standard deviation.

The results we obtain by subjecting our data to analysis may actually be true or may be due to chance or sampling variation. In order to rule out chance as an explanation, we use the test of significance. In this unit we have confined our discussion to four tests i.e. χ^2 test, Z- test, t-test and f-test.

Correlation is relationship between the two sets of continuous data; for example relationship between height and body weight. Correlation statistics is used to determine the extent to which two independent variables are related and can be expressed by a measure called the coefficient of correlation. Regression, on the other hand, deals with the cause and effect relation between two sets of data. Simple linear regression fits a straight line through the set of n points in such a way that makes the sum of squared *residuals* of the model (that is, vertical distances between the points of the data set and the fitted line) as small as possible. The regression line, thus, obtained helps us to predict the value of dependent variable for a given value of independent variable.

Annex I: Table of chi-square values

Degrees of freedom	χ^2 value if $\alpha = 0.05$	χ^2 value if $\alpha = 0.01$
1	3.84	6.63
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21
11	19.68	24.72
12	21.03	26.22

Annexure-II: Table of the values

Degrees of freedom	t-value if chosen p (α) = 0.05	t-value if chosen p (α) = 0.01
1	12.71	63.66
2	4.30	9.92
3	3.18	5.84
4	2.78	4.60
5	2.57	4.03
6	2.45	3.71
7	2.36	3.50
8	2.31	3.36
9	2.26	3.25
10	2.23	3.17
11	2.20	3.11
12	2.18	3.05
13	2.16	3.01
14	2.14	2.98
15	2.13	2.95
16	2.12	2.92
17	2.11	2.90
18	2.10	2.88
19	2.09	2.86
20	2.09	2.85
21	2.08	2.83
22	2.07	2.82
23	2.07	2.81
24	2.06	2.80
25	2.06	2.79
30	2.04	2.76
40	2.02	2.70
60	2.00	2.66
120	1.98	2.62
Infinite	1.96	2.58

1.15 KEYWORDS

Independent variable: the characteristic being observed or measured which is hypothesized to influence an event or outcome (dependent variable), and is not influenced by the event or outcome, but may cause it, or contribute to its variation.

Dependent variable: a variable whose value is dependent on the effect of other variables (independent variables) in the relationship being studied.

Mean: the mean (or, arithmetic mean) is also known as the average. It is calculated by totaling the results of all the observations and dividing by the total number of observations.

Median: the median is the value that divides a distribution into two equal halves. The median is useful when some measurements are in ordinal scale, i.e., much bigger or much smaller than the rest.

Mode: the mode is the most frequently occurring value in a set of observations. The mode is not very useful for numerical data that are continuous. It is most useful for numerical data that have been grouped. The mode is usually used to find the norm among populations.

Range: this can be represented as the difference between maximum and minimum value or, simply, as maximum and minimum values.

Percentiles: percentiles are points that divide all the measurements into 100 equal parts. The 30th percentile (P3) is the value below which 30% of the measurements lie. The 50th percentile (P50), or the median, is the value below which 50% of the measurements lie.

Mean Deviation: this is the average of deviation from arithmetic mean.

Standard Deviation: this denotes (approximately) the extent of variation of values from the mean.

Parametric statistical test is a test whose model specifies certain conditions about the parameters of the parent population from which the sample was drawn.

Non-parametric statistical test is a test whose model does not specify conditions about the parameters of the parent population from which sample was drawn.

Normal Distribution: The normal distribution is symmetrical around the mean. The mean, median, and mode assume the same value if observations (data) follows a normal distribution.

Sampling Variation: any value of a variable obtained from the randomly selected sample (e.g., a sample mean) cannot assume the true value in the population. The variation is called a sampling variation.

Test of Significance: a test of significance estimates the likelihood that an observed study result (e.g., a difference between two groups) is due to chance or real.

1.16 REFERENCES AND SELECTED READINGS

Altman, D.G. (1991), *Practical Statistics for Medical Research*, Chapman and Hall, London.

Barker, D.J.P. (1982), *Practical Epidemiology*. (3rd ed.), Churchill Livingstone Edinburgh, UK.

Bradford, H. A. (1984), *A Short Textbook of Medical Statistics* (11th ed.), Hodder and Stoughton London, UK.

Castle, W.M. and North P.M. (1995), *Statistics in Small Doses*. Churchill Livingstone Edinburgh, UK.

Fletcher, R. H., S. W. Fletcher and E. H. Wagner (1996), *Clinical Epidemiology: The Essentials*, Lippincott Williams and Wilkins, 351 West Canadian Street Baltimore, Maryland, USA.

Glaser, A.N. (2000), *High-yield Biostatistics*, Lippincott Williams and Wilkins, 227 East Washington Square, Philadelphia, USA.

Greenhalgh, T.(1998), *How to Read a Paper: The Basics of Evidence Based Medicine*, BMJ publishing group, BMA House, Tavistock Square, London,UK.

Hicks, C.M. (1999), *Research Methods for Clinical Therapists. 3rd Edition*, Churchill Livingstone, Robert Stevenson House, 1-3 Baxter's Place,Leith Walk, Edinburgh, UK.

Kelsey, J.L., W.D. Thompson and A.S. Evans (1986), *Methods in Observational Epidemiology*, Oxford University Press, Oxford, UK.

Kidder, L.H. and C.M. Judd (1986), *Research Methods In Social Relations*, CBS College Publishing, New York, USA.

Kleinbaum, D.G., L.L. Kupper and H. Morgenstern (1982), *Epidemiologic Research - Principles and Quantitative Methods*, Van Nostrald Reinhold, New York, USA.

Riegelman, R.F. (1981), *Studying a Study and Testing a Test*, Little Brown and Company, Boston, MA, USA.

Schlesselman, J.J. (1982), *Case-Control Studies - Design, Conduct, Analysis*, Oxford University Press, Oxford, UK.

Siegel, S. (1956), *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill Book Company.

Swinscow, T.D.V. and M.J. Campbell (1998), *Statistics at Square One* (11th ed.), British Medical Association, London, UK.

1.17 CHECK YOUR PROGRESS – POSSIBLE ANSWERS

Check Your Progress-1

- 1) There are two types of data: (i) qualitative data, viz., occupation, sex, marital status, religion, and; (ii) quantitative data viz., age, weight, height, income, etc. These may be further be categorized in two types viz., Discrete and continous datas.
- 2) The two types of variables are used in analysis of data are (i) independent variables which are not influenced by the event or outcome, but may cause it, or contribute to its variation and (ii) dependent variable whose value is dependent on the effect of other variables (independent variables) in the relationship, being studied.

Check Your Progress-2

- 1) The three measures of central tendency are the mean, median, and mode.
- 2) The measures of dispersion are range, percentiles, mean deviation and standard deviation.

Check Your Progress-3

- 1) In statistical terms, the assumption that no real difference exists between groups in the total study (target) population (or that no real association exists between variables) is called the Null Hypothesis (H_0). The Alternative Hypothesis (H_1) is that there exists a difference between groups or that a real association exists between variables.
- 2) Type I error (α): We reject the null hypothesis when it is true, or a false positive error or type I error α (called alpha). It is the error in detecting true effect.
Type II error (β): We accept the null hypothesis when it is false, or a false negative error; or simply, type II error ' β ' (called beta) can be stated as failure to detect true effect.

Check Your Progress-4

- 1) The chi-square test can only be applied if the sample is large enough. The total sample should be at least 40 and the expected frequencies

in each of the cells should be at least 5. Unlike the t-test, the chi-square test can also be used to compare more than two groups. In that case, a table with three or more rows/columns would be designed, rather than a two-by-two table. The chi-square is always applied on absolute numbers but not on percentage values. The high chi-square value never means high association but it only means high probability of finding such a value and low chance of finding this chi-square value by chance. It may be very misleading to pool dissimilar data.

- 2) When dealing with paired (matched) observations, comparison of sample means is performed by using a modified t-test known as the **paired t-test**. In the **paired t-test**, differences between the paired observations (say, Post-test minus Pre-test, or matched observations of 2nd group minus 1st group) are used instead of the observations of two sets of independent samples.

Check Your Progress 5

- 1) Correlation statistics are used to determine the extent to which two independent variables are related and can be expressed by a measure called the coefficient correlation. Correlation coefficient may be positive or negative and may vary from -1 to +1. In simple regression analysis, there is one quantitative dependent variable and one independent variable. In multiple regression analysis, there is one quantitative dependent variable and two or more independent variables.

UNIT 2 USE OF COMPUTER IN DATA ANALYSIS

Structure

- 2.1 Introduction
- 2.2 Need for the Use of Computer in Data Analysis
- 2.3 Working with Excel
- 2.4 SPSS: An Overview
- 2.5 Data Management
- 2.6 Using Descriptive Statistics to Describe Your Data
- 2.7 Testing of Hypothesis Using SPSS
- 2.8 Multiple Regression Using SPSS
- 2.9 Let US Sum Up
- 2.10 References and Selected Readings
- 2.11 Check Your Progress – Possible Answers

2.1 INTRODUCTION

Data analysis is an integral part of any research and more particularly of development research. It is a process of applying statistical or logical techniques to describe, illustrate, condense or evaluate data. In this unit we will be dealing with data analysis on MS excel and SPSS. MS excel allows the creation, analysis and revision of data. Excel is a spreadsheet programme, used for data analysis, in the Microsoft Office system. Excel offers a variety of functions like tracking data, building models for analyzing data, writing formulas to perform calculations on the data, pivot the data in numerous ways and present the data in a number of chart forms. SPSS or Statistical Package for Social Sciences is a windows based programme that can be used for data entry and analysis and also to create tables and graphs. SPSS is a comprehensive and flexible statistical analysis and data management solution. SPSS can take data from almost any type of file and use them to generate tabulated reports, charts, and plots of distributions and trends, descriptive statistics, and conduct complex statistical analyses. You will find SPSS customers in virtually every industry, including telecommunications, banking, finance, insurance, healthcare, manufacturing, retail, consumer packaged goods, higher education, government, and market research.

After reading this unit you will be able to:

- feed data on MS Excel and work out mean and standard deviation;
- prepare a chart using MS Excel;
- feed your research on SPSS;
- manage and do basic transformations needed for your research on SPSS;
- calculate descriptive for your data on SPSS;
- perform testing of hypothesis on SPSS;
- work out correlation on SPSS; and
- work out regression analysis on SPSS.

2.2 NEED FOR THE USE OF COMPUTER IN DATA ANALYSIS

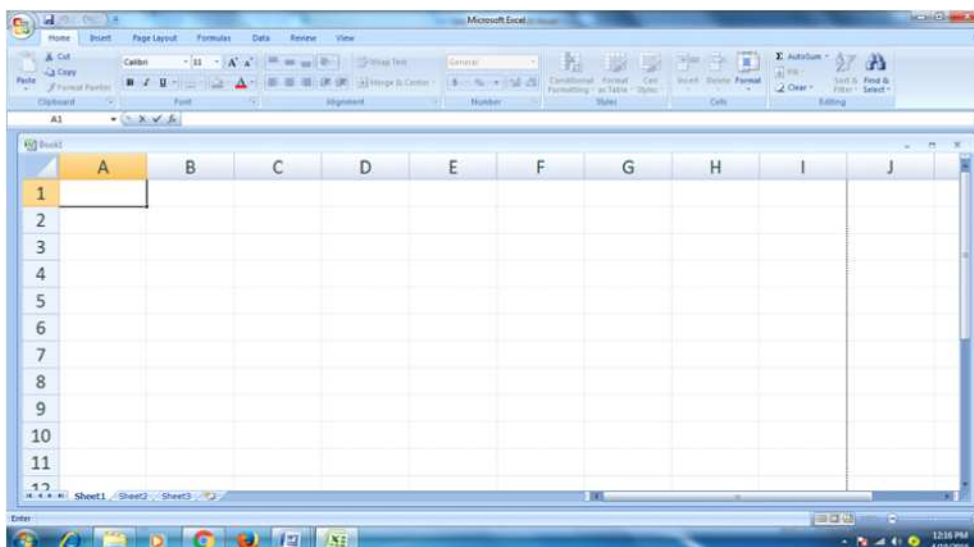
In research, you are required to handle a lot of data. Handling a small data on a piece of paper with the use of a calculator is easy but imagine calculating mean of a variable for 200 or more data. Using a calculator in this case would not just be time consuming but you are also in the danger of making mistakes. You may just punch in a wrong number and not notice it. Such mistakes can lead to erroneous results. Wrong results can render your research meaningless and hence you need to be very cautious in your data analysis and interpretation. Using computer to analyse your data will not just speed up the process of data analysis but also help in reducing the errors. Some of the advantages of use of computers in data analysis is given below:

1. It helps in reducing/eliminating errors in calculation
2. It helps in better data management as you can easily add or delete variables and observations, code and recode values etc.
3. It is very easy to depict the results in graphs and charts
4. Several users can work on the same set of data simultaneously by sharing the data file
5. It is a faster and more efficient way of handling data

2.3 WORKING WITH EXCEL

2.3.1 Excel Worksheet

An excel worksheet is a two dimensional grid comprising of rows and columns. The columns are named alphabetically and the rows are numbered. The cells in the first row will be referred to as A1, B1, C1, and so on and the cells in the first *column will be referred to as* A1, A2, A3, and so on. Each excel file is a workbook which is a collection of worksheets. By default, a workbook contains three worksheets which are named as sheet1, sheet 2 and sheet 3. You can use the arrow key to move up, down, right or left from the current cell. Enter key can be used to move one cell down from the current position and the tab key can be used to move one cell to the right.



In order to select a cell, click in the cell and in order to select a group of cells, click in a corner cell and, with the left mouse button depressed, drag the cursor horizontally and/or vertically until all of the cells you want selected are outlined in black. To select multiple cells that are not contiguous, press and hold the Ctrl key while clicking in the desired cells. To select every cell in the worksheet, click in the upper right corner of the worksheet to the left of "A."

2.3.2 Entering Data in a Worksheet

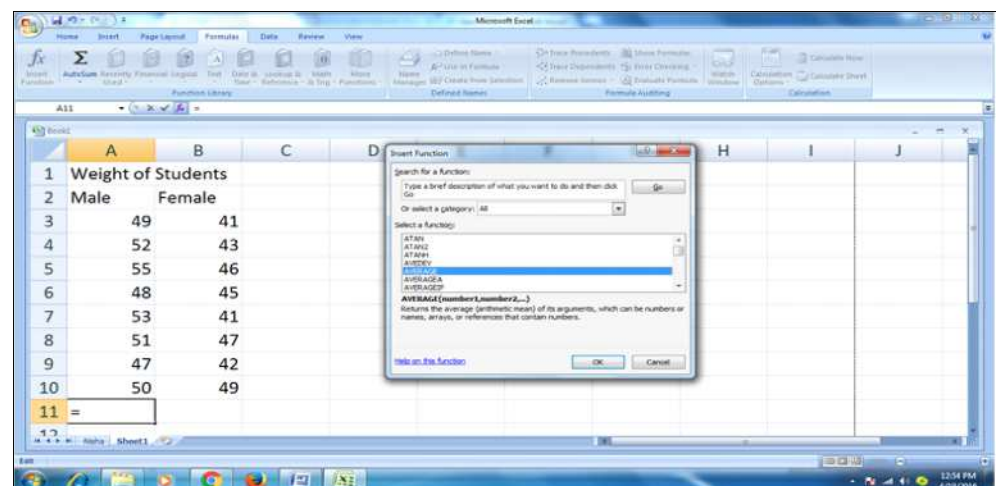
To enter data in a cell, you just need to click in that cell and type. The contents in a cell can be edited by double clicking in the desired cell and then making the needed correction. To insert a new row in a worksheet, right click on the row number and click insert. A new row will be added above the row that was clicked. To delete a row, right click on the row number and click delete. In the same way, to insert a new column, right click on a column head and click insert. A new column will be inserted to the left of the column that was clicked. To delete a column, right click on the column letter and click delete.

Worksheet tabs are found at the bottom of the workbook, towards the left. If you want to open any particular worksheet within a workbook, you just need to click on that particular worksheet tab. It is also possible to rename the worksheet. In order to rename a worksheet, you need to right click on the worksheet tab and click on the rename option. The new name can then be entered. In order to insert an additional worksheet, right click on the worksheet tab and click insert. A new worksheet will be inserted towards the left. Similarly, to delete an existing worksheet, right click on that worksheet tab and click delete. The positions of the worksheet can also be changed by left clicking on the worksheet tab and then dragging it to the desired position.

2.3.3 Basic Calculations on Excel

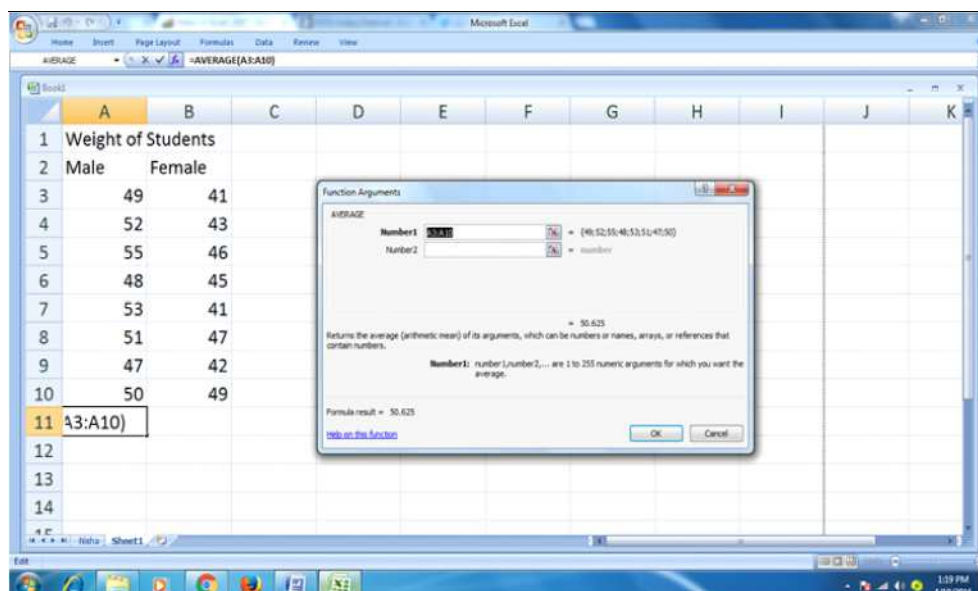
Calculating Mean in Three Easy Steps

We take a simple example in which we have the weights of 8 male and 8 female students. We need to calculate the average weight of male and female students.



Step 1: Click on the 'formula' tab on the top and then click 'insert function'.

Step 2: Select the option 'All' in the *Select a category* option so that all the available functions are displayed in the box under the heading *Select a function*. Select 'AVERAGE' function in the box and click **OK**.



Step 3: Another box titled *Function Arguments* opens. Within this box are options 'Number 1' and 'Number 2'. In 'Number 1' you need to select the numbers for which you want the average. This can be done by left clicking on the first number 40 and then dragging it upto the last number which is 50 in case of male students. Then press the **OK** button. It will give the average weight of eight male students as 50.625.

Calculating Standard Deviation in Three Easy Steps

Step 1: Click on the 'formula' tab on the top and then click 'insert function'.

Step 2: Select the option 'All' in the *Select a category* option so that all the available functions are displayed in the box under the heading *Select a function*. Select 'STDEV' function in the box and click **OK**.

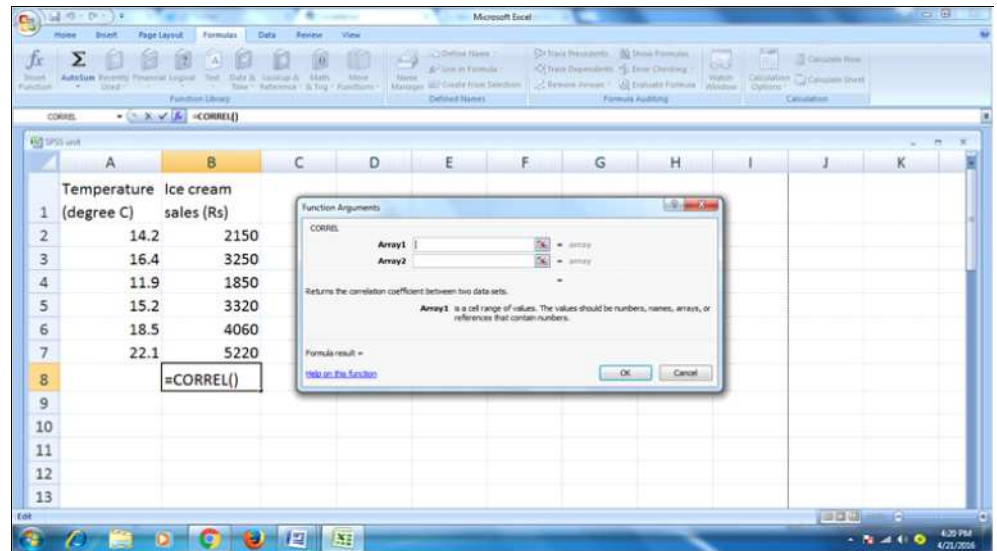
Step 3: Step 3: Another box titled *Function Arguments* opens. Within this box are options 'Number 1' and 'Number 2'. In 'Number 1' you need to select the numbers for which you want the average. This can be done by left clicking on the first number 40 and then dragging it upto the last number which is 50 in case of male students. Then press the **OK** button. It will give the standard deviation of the weight of eight male students which is 2.67. Similarly the standard deviation of female students works out to be 2.96.

Calculating Correlation

Step 1: Click on the 'formula' tab on the top and then click 'insert function'.

Step 2: Select the option 'All' in the *Select a category* option so that all the available functions are displayed in the box under the heading *Select a function*. Select 'CORREL' function in the box and click **OK**.

Step 3: Another box titled 'Function Arguments' opens which has two boxes 'Array1' and 'Array2'. Select the data on temperature for 'Array1' and the data on ice cream sales for 'Array2'. Then press the **OK** button. It will give the correlation between temperature and the corresponding ice cream sales as .976 indicating that there is a very high correlation between the temperature levels and sale of ice creams.



Similarly, you can do several other statistical and mathematical calculations using excel by selecting the desired function in the formula → insert function option.

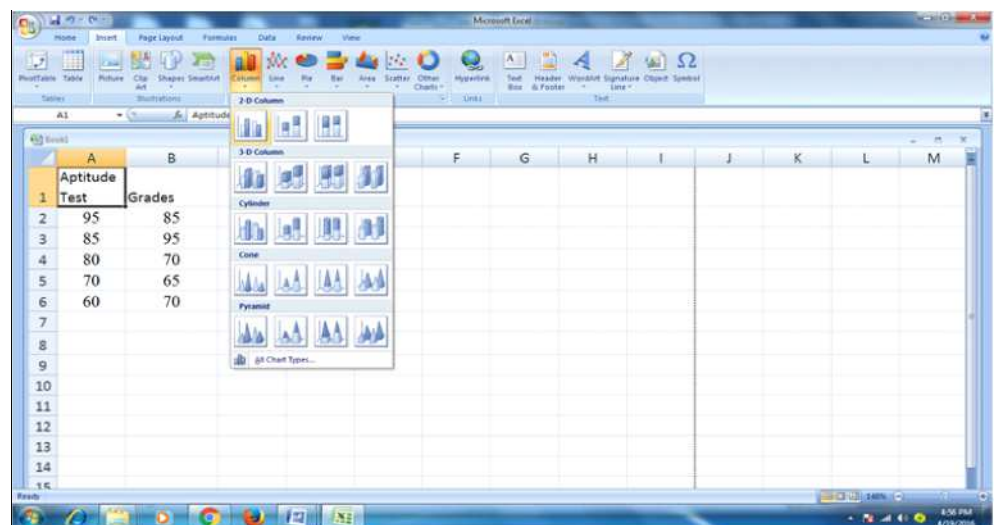
Activity

Feed the data of the weight of Female Students in the example shown in the picture above in Excel and work out the mean and standard deviation.

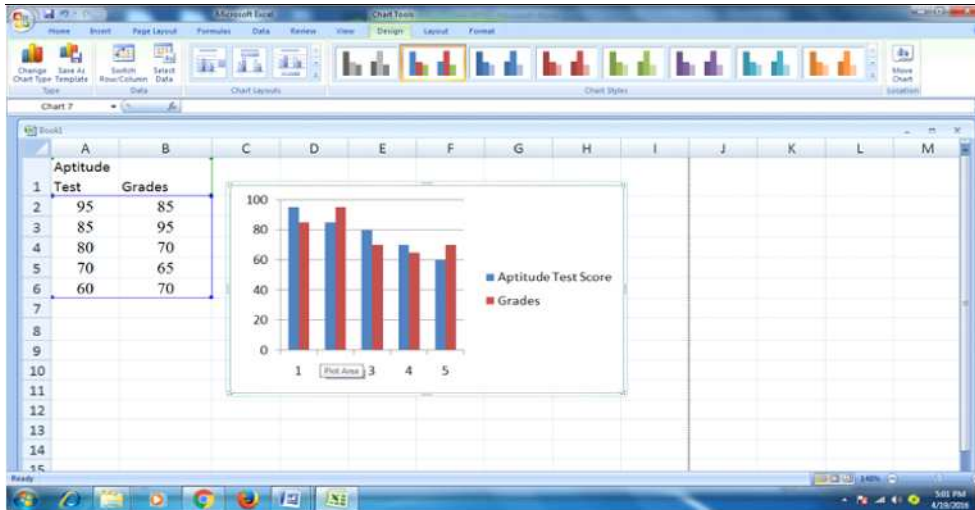
2.3.4 Charts in Excel

One of the most important and appealing features of excel is the charts. Charts make data visual. It is easy to present your results in charts using excel. We will see how to construct a simple 2D column chart using excel. You can explore other options like line, pie, bar, scatter and many other options available on excel in similar fashion.

In our example, we take the results of aptitude test scores and grades of five students and plot it in a simple 2D column chart.



Step 1: Go to 'Insert' and then on 'column'. A number of chart options within column chart will be displayed and you can click on the chart you wish to construct. We click on the first option shown which is the 2D column chart. We can get a chart at the click of a button.



Note: you can modify this chart as you desire.

1. To remove gridlines, right click on the grid line and click delete.
2. You can include the data labels (value of data on each bar) by selecting the bars by a left click on any of the bar. For example if you left click on any one of the bar representing grades, all the bars will be selected. Then right click and then click on the option 'insert data labels'. By default, the data labels appear on the top of the bar which may be repositioned by first selecting the data labels by a left click and then right click and click on the option 'format data labels'.
3. You can change the colour of the bars by selecting the bars with a left click. Then right click and select 'format data series', 'fill' and then 'solid fill'. Then choose the desired colour and click on the **close** button.

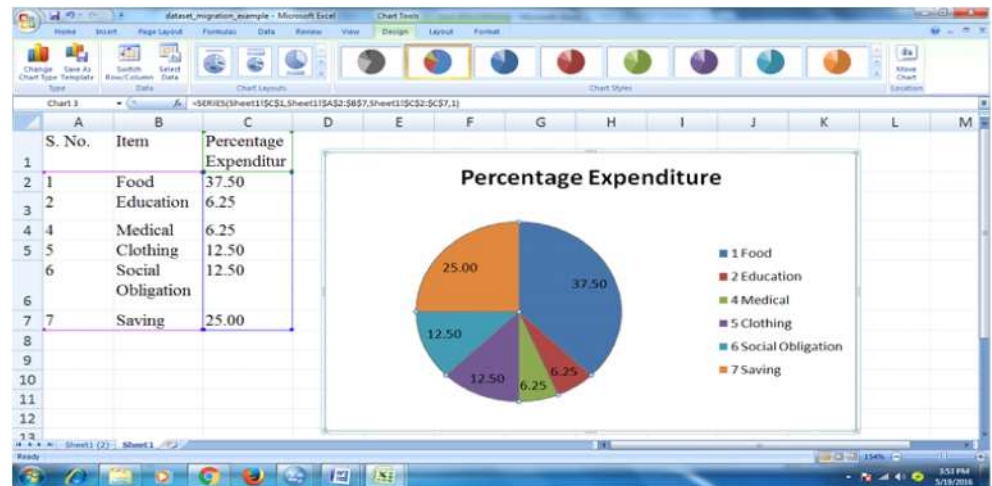
There are several other options available on excel to modify your chart to suit your style and liking.

Let us also see how you can use excel to draw a pie chart. For example, A survey of 100 vegetable vendors showed an average monthly consumption expenditure as follows:

S. No.	Item	Expenditure	Percentage
1.	Food	3000	37.50
2.	Education	500	6.25
4.	Medical	500	6.25
5.	Clothing	1000	12.50
6.	Social Obligation	1000	12.50
7.	Saving	2000	25.00

We can use excel to represent this in a pie chart by following steps.

Step 1: Go to 'Insert' and then on 'Pie'. You then have an option to choose a 2D or a 3D pie. Choose your option and click ok. You get a pie chart for the expenditure percentage as shown in the figure below:



Excel can be used to work out correlations, regression, carrying out tests of significance and several other purposes. In this section you have seen how to work out the basic calculations and charts. You can explore all the other options on Excel as the more you use it the more you learn.

2.4 SPSS: AN OVERVIEW

SPSS files:

SPSS covers a broad range of statistical procedures used in social sciences like data summarization, test of significance, determining relationship between variables, time series analysis and also graphs. SPSS has three main windows which allow you to i) see your data(data editor), ii) see your output (output viewer), and iii) see any programming commands that you have written(syntax editor). Each of these windows have a separate type of SPSS file. Let us look at each of these files and their functions.

1. Data Editor (.sav files):

Data editor is used to enter data in a spreadsheet format. Data editor lets you see and manipulate your data. Data files are saved as a file type .sav. SPSS lets you to open many datasets simultaneously. Creating a data file will be dealt with in subsequent sessions.

2. Output Viewer (.spv files):

As you command SPSS to perform various computations, the statistical results will show up in output viewer. The output viewer shows the tables and graphs that you create. Output viewer files are saved with they file type .spv.

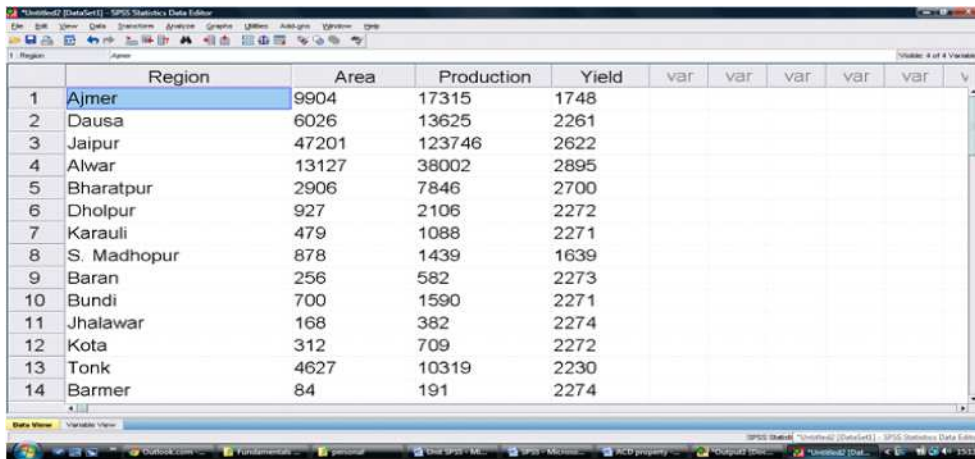
3. Syntax editor (.spx files):

If you want to work on SPSS programming language directly, you will have to open the SPSS syntax editor. The syntax editor can be used to write, edit and run commands. If you are using menus and dialog boxes, the paste button automatically pastes the syntax for the command you have specified in the active syntax editor. The syntax files are saved as plain text and most of the text editors can open them with a file extension of .sps. More than one syntax editor can be opened at a time and in such cases when syntax is pasted from a dialogue box, it gets pasted in the active syntax editor. When all the syntax editors are closed, the paste command leads to opening of a fresh syntax editor.

1. Working with Data Editor

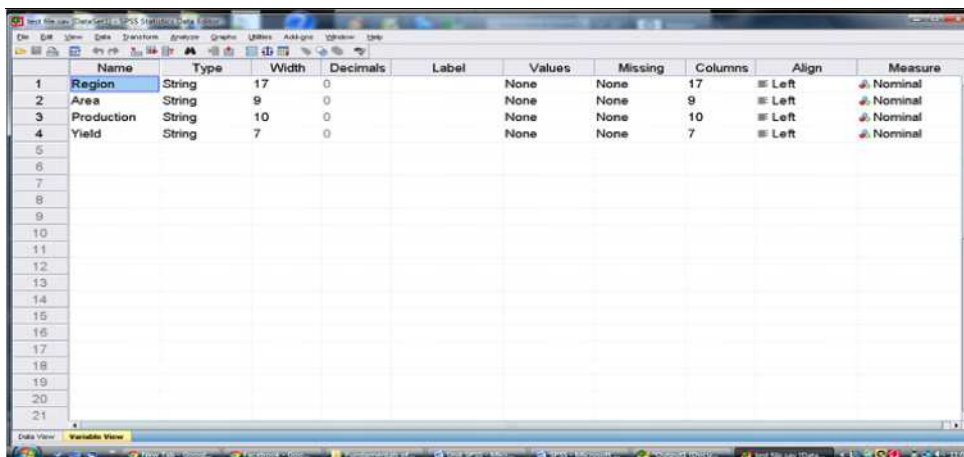
Data editor is used to define variables, enter data and define or redefine the characteristics of the variables. The data editor gives you two views

of your data: **the data view** and **the variable view**. The relevant view can be selected from the tabs given at the lower left corner of the window.



	Region	Area	Production	Yield	var	var	var	var	var
1	Ajmer	9904	17315	1748					
2	Dausa	6026	13625	2261					
3	Jaipur	47201	123746	2622					
4	Alwar	13127	38002	2895					
5	Bharatpur	2906	7846	2700					
6	Dholpur	927	2106	2272					
7	Karauli	479	1088	2271					
8	S. Madhopur	878	1439	1639					
9	Baran	256	582	2273					
10	Bundi	700	1590	2271					
11	Jhalawar	168	382	2274					
12	Kota	312	709	2272					
13	Tonk	4627	10319	2230					
14	Barmer	84	191	2274					

- i. **The Data View:** In the data view, the data are laid in a rectangular format of rows and columns in which the columns represent the variables and each row represents a unit of observation. The data in a column should be of the same type ie. either numeric (numbers) or string (characters). Data values can be added by typing the values directly in the data view.

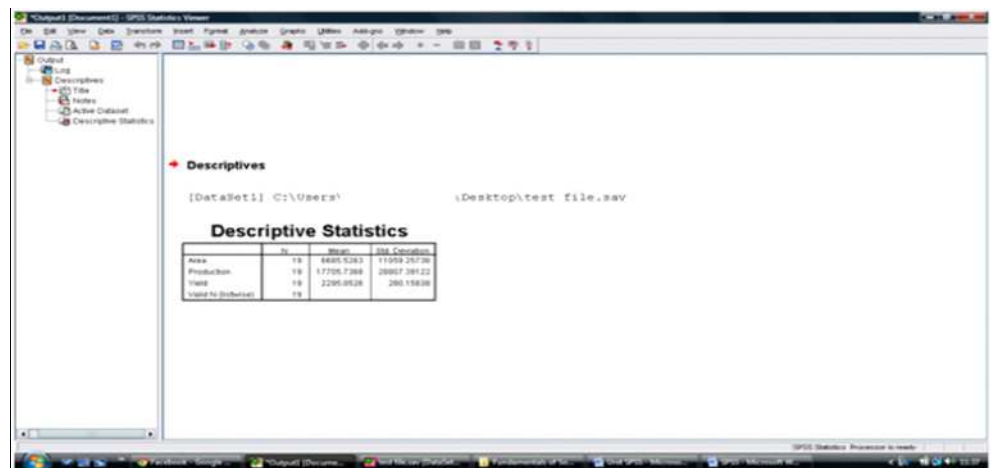


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	Region	String	17	0		None	None	17	Left	Nominal
2	Area	String	9	0		None	None	9	Left	Nominal
3	Production	String	10	0		None	None	10	Left	Nominal
4	Yield	String	7	0		None	None	7	Left	Nominal
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										
21										

- ii. **The Variable View:** You can easily switch between data view and variable view by clicking on the relevant button on the lower left corner. In the variable view you can edit the information that defines a variable. Each row of the variable view describes a column of the data view. The first attribute or the **Name** is how the data column is named or identified. The Name may contain characters, numerals, punctuation marks etc. However, *space* is considered as an illegal character while naming a variable. The **Type** of data is generally a numeric or a string but SPSS also provides with other options (comma, dot, scientific notation, currency etc.) by which you may define the type of your data. **Label** may be considered as a longer description of the variable name. **Values** allow you to create a list of value labels for example you may use 1 and 2 for male and female gender. If a researcher wants SPSS to designate certain data values as missing values, it can be done conveniently on SPSS using the missing attribute. For instance the researcher may give a number, say '8' to all the responses in which respondent says "I don't know" in response to a question, the researcher can have SPSS treat all 8s in a variable as missing data.

2. Working with Output Viewer:

All the statistical tables and graphs analysed in the data editor can be viewed in the output viewer. The output viewer consists of two sections: the one on the left is the outline pane and the one on the right is the tables pane. Whenever the output for any analysis is generated, it is added on the tables pane as object. Selection of an object can be done by clicking on it in the tables pane or clicking the corresponding entry in the outline pane. It is also possible to edit the objects on the table plane by double clicking on the object. To delete the object, simply select the object and press the **Delete** key. Double clicking an object in the outline pane can be used to hide an object from the tables pane. To rearrange or reposition any object on either pane, you just need to select the object and drag and drop the object at the place where you want it to be. You can also export your output by clicking **File à Export** in the output viewer. A new dialogue box will open and you can select the options regarding the portion of output and the type of file in which you want it to be saved.

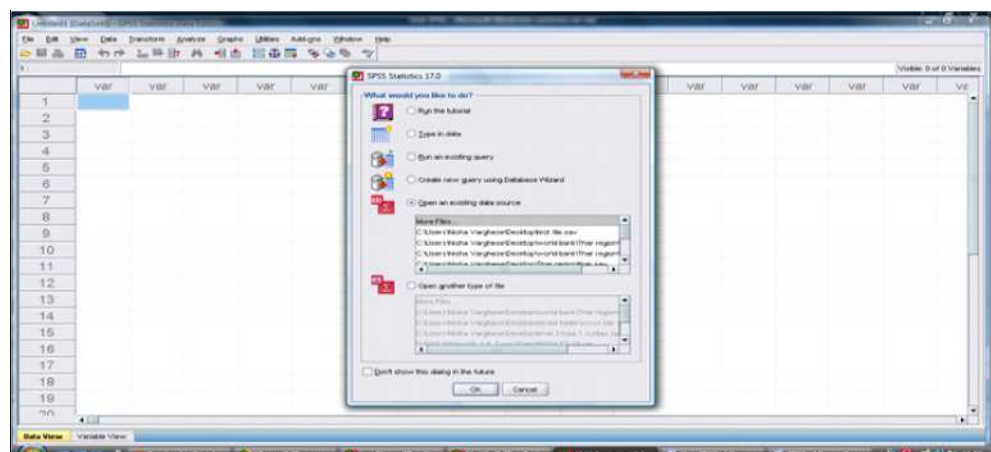


2.5 DATA MANAGEMENT

Once you have collected the data for your research, the next step would be to enter the data in appropriate format so as to facilitate statistical analysis. In this section you will read about data management in SPSS.

2.5.1 Creating a Data File

The first step is to create a data file. SPSS can also read files created by other programmes such as Microsoft Excel. When you first open the SPSS, you will be prompted to choose to either run a tutorial or type new data or open an existing file.



It is important to note here that the rows are used to enter information on respondents or cases and columns are used to enter variable information. For example you have to enter information on area, production and yield of a particular crop in 19 districts in Rajasthan, 'area, production and yield' will be represented in rows and districts in columns.

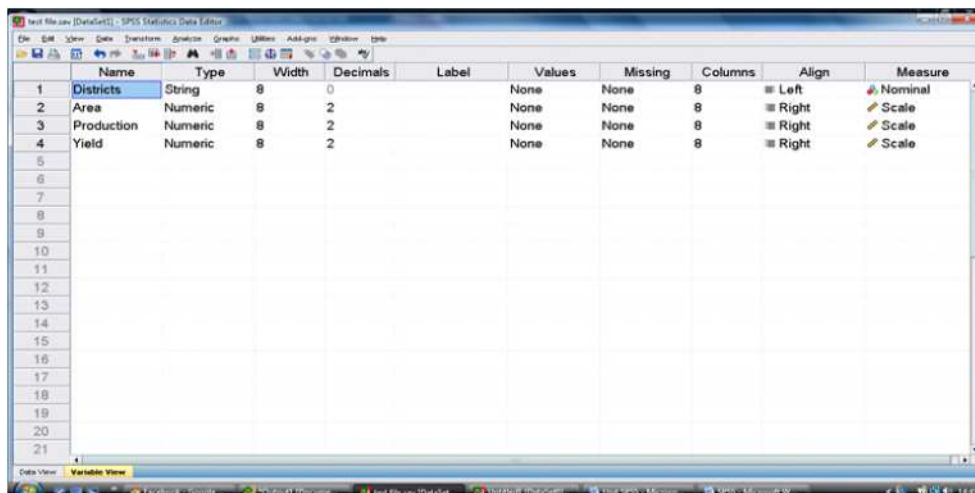
SPSS data editor will have two views: **Data view** and **Variable view** and you can select the relevant view by clicking at either of the option given at the bottom left of the window. To proceed logically, you would first click on the **Variable View**. Here you will define your variables.

Under **Name**, you would be required to type the desired variable name. While naming a variable, you need to take care that the first letter should be a alphabet and no spaces can appear in the variable name. The next is to define the type of the variable under the **Type** option. The functions and corresponding default values of each of these columns are given below:

Characteristics of Variables and their Default Values:

Characteristics of Variables	Explanation	Default value taken by SPSS
Type	Data type: Numeric, date, string etc.	Numeric
Width	Total number of Characters	8
Decimals	Number of characters after decimal	2
Label	Description of Variable	None
Values	Give labels to different levels of variables	None
Missing	Enables you to designate certain data as missing	None
Columns	Maximum number of characters in a column	8
Align	Alignment of values in a given column	Right
Measure	Specifies the measurement scale of the variable	Scale

The picture below gives the details of the variable view for our dataset.



The next step would be to enter the data. The variable information entered in the variable view can be altered at later stage. Next click the **Data View** at the bottom left of the window. Once the variable information is entered in the variable view, you would be able to see all the variable names that you entered in the **Data View**. Now type the name of the districts and the data pertaining to area production and yield. Once you are done with entering the data, you need to save the data by clicking the **File → Save** option at the top left corner of the window. You can give a relevant name to the file when the dialogue box opens. The SPSS files are saved with **.sav** as suffix.

Some important tips in data management in SPSS:

- You can open an existing data file by clicking **File → Open → Data** in sequence.
- You can insert an additional variable/column by clicking at the column next to the desired place where you want to insert the variable, go to **edit** option at top and click **insert variable**. A new variable can be inserted by following these steps in both data and variable view.
- In order to insert a new row, just above the selected row, click on the row above which you want to insert the additional row. Then click **edit** option at top and click **insert case**. A blank row will appear just above the selected row.

So far you have read about working with excel, an overview of SPSS and data management in SPSS. Now answer the questions given in *Check Your Progress-1*.

Check Your Progress - 1

- Note:** (a) Write your answer in about 50 words.
(b) Check your answer with possible answers given at the end of the unit.

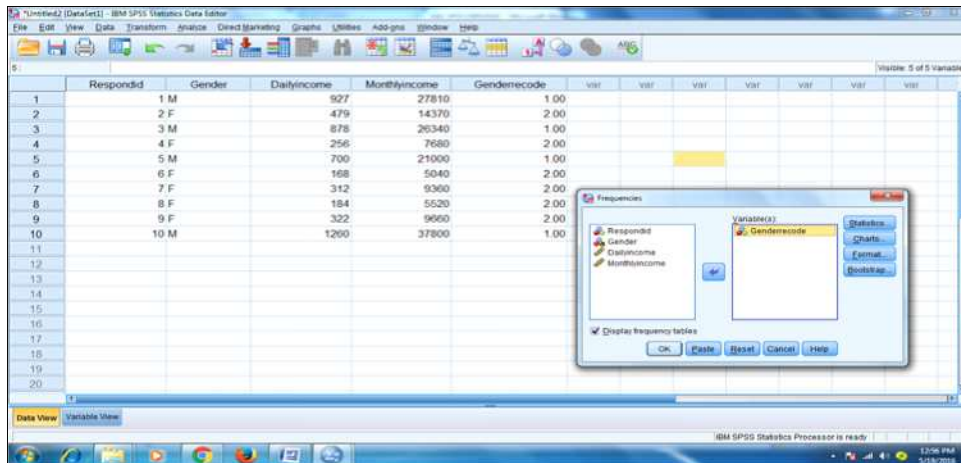
1. Explain about data editor in SPSS?
.....
.....
.....
2. How will you include data labels in a bar graph in excel?
.....
.....
.....

2.6 USING DESCRIPTIVE STATISTICS TO DESCRIBE YOUR DATA

Descriptive statistics are used to describe a variable and to study its summary measures like mean, median, mode and standard deviation. SPSS provides you an easy option to calculate all the descriptive statistics simultaneously in one go. The steps for calculating descriptive statistics are as follows:

1. **Frequency**
 - i. In order to calculate frequencies, the following steps need to be followed. Click on **Analyse** on the main menu and **Descriptive Statistics**

followed by **Frequencies** on the sub menu. A dialogue box will open. Shift the desired variable in the **Variable(s)** text box and click **OK**.



- ii. The gender has been recoded as 1 for Males and 2 for Females. The results will be displayed in the output file. The results of working out frequencies of the variable gender using SPSS will be as under

Genderrecode

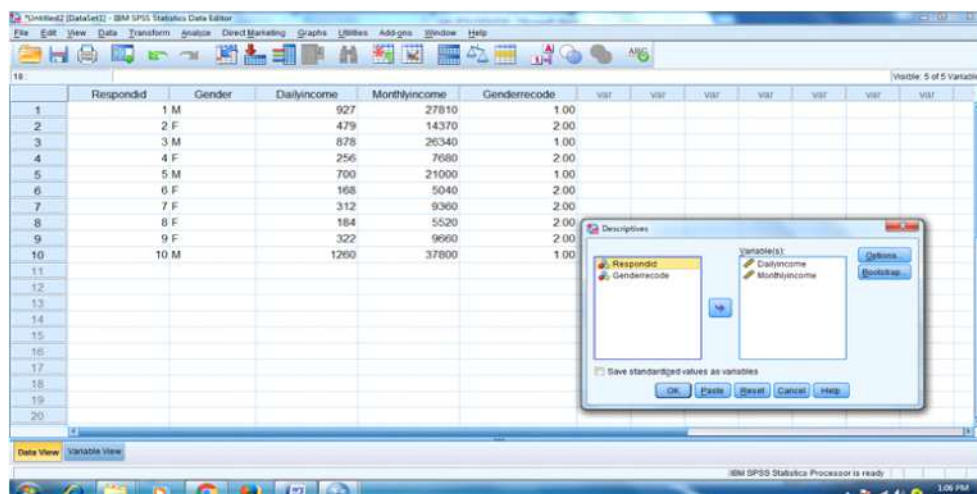
	Frequency	Percent Percent	Valid Percent	Cumulative
1.00	4	40.0	40.0	40.0
Valid 2.00	6	60.0	60.0	100.0
Total	10	100.0	100.0	

- iii. This shows that in the given data there are 4 males and 6 females.

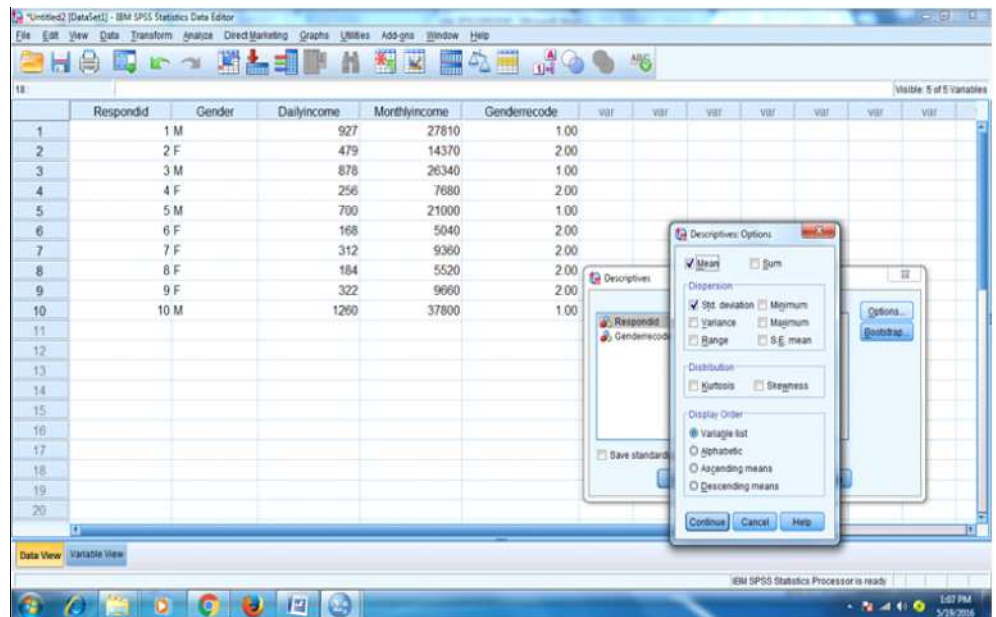
2. Descriptive Statistics

Descriptive statistics includes calculation of various measures of central tendency and measures of dispersion. These can be calculated by following the steps given below:

- i. Click **Analyze, Descriptive Statistics** and **Descriptives**. This will open a dialogue box in which you move the variables whose descriptives need to be calculated by shifting them to the **Variable(s)** text box. In this example we have shifted the daily income and monthly income to the variable box.



- ii. Click on the **Options** button to get another dialogue box. Then select the desired statistics by clicking the appropriate check boxes. In this example we have checked the Mean and Standard Deviation. Click **Continue** to go back to the previous dialogue box and then click **OK**. In this example, we have selected Mean and Standard Deviation.



- iii. The results will be displayed in the output file. The Mean and Standard Deviation for daily income and monthly income as calculated using SPSS are as given:

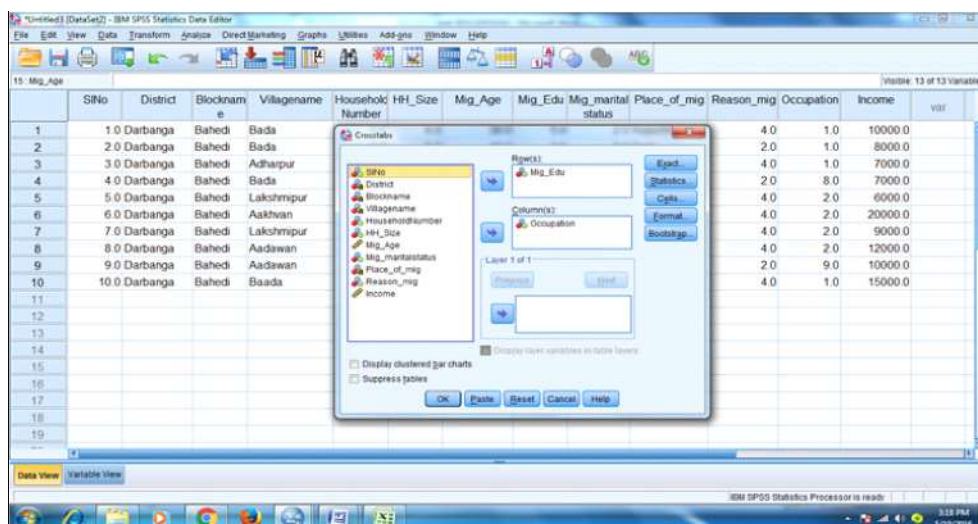
Descriptive Statistics

	N	Mean	Std. Deviation
Dailyincome	10	548.60	373.613
Monthlyincome	10	16458.00	11208.383
Valid N (listwise)	10		

3. Crosstabs

The **Crosstabs** command is used to get information on the intersection of two or more categorical variables. While information on each level of one categorical variable can be obtained by working out frequencies but it does not provide information on intersection of two variables. This can be done using the crosstabs command as follows:

- For example in a given data set we want to cross tabulate migrant education and their occupation.
- Click **Analyze, Descriptive Statistics** and **Crosstabs**. A dialogue box opens. Select **Mig_Edu** and move it to the **Rows** list box and select **Occupation** and move it to the **Columns** list box.
- Click **Cells** button which opens another dialogue box. This dialogue box allows you to select additional information you want to compute for the selected variables. Select the relevant information and click **Continue** button to go to the previous dialogue box and then click **OK**.



iv. The result that we get in the output file is as under.

Migrant Education was classified into 9 categories viz. 1. Illiterate, 2. Literate with formal education, 3. Primary, 4. Matriculate, 5. Intermediate, 6. Technical Education, 7. Graduate, 8. Post Graduate, 9. Others.

Migrant occupation was classified as 1. Unskilled worker, 2. Skilled worker, 3. Shopkeeper, 4. Self Employed, 5. Salesman, 6. Officer/executive, 7. Software Professional, 8. Others, 9. Unemployed.

Thus in this example, out of a total of 4 literates with formal education, one was employed as unskilled worker, 2 were working as skilled workers and one was engaged in other occupation not listed here.

Mig_Edu * Occupation Crosstabulation

Count

		Occupation				Total
		1.0	2.0	8.0	9.0	
Mig_Edu	2.0	1	2	1	0	4
	3.0	2	2	0	0	4
	4.0	0	0	0	1	1
	5.0	1	0	0	0	1
Total		4	4	1	1	10

2.7 TESTING OF HYPOTHESIS USING SPSS

SPSS can also be used for testing of hypothesis. A hypothesis is a statement about the population. Testing of hypothesis helps in establishing the truth about the statement made about the population.

1. One Sample t Test

One sample t test is used to test if the observed mean is different from the population mean. The hypothesis that is tested for its claim about the population parameter is called the null hypothesis. Another hypothesis which is the opposite of null hypothesis is the alternative hypothesis. This is accepted in case the null hypothesis is rejected.

T test is used when the sample size is less than 30 and population standard deviation is not known. For example, a CFL tube manufacturing company claims that the average life of tube manufactured by them is 3000 hours. To test this claim, sample data is collected and analysed. Here

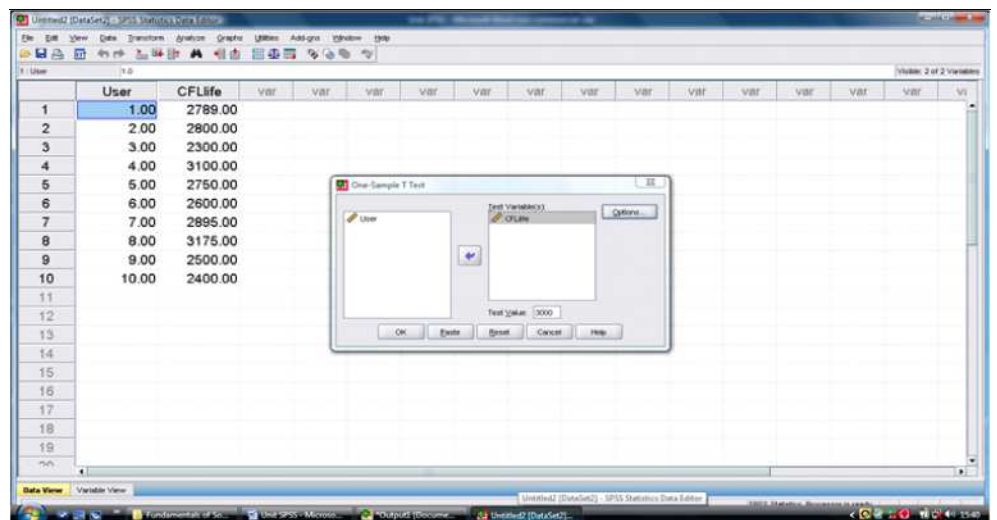
$$H_0: \mu = 3000$$

$$H_1: \mu \neq 3000$$

User	1	2	3	4	5	6	7	8	9	10
CFL life	2789	2800	2300	3100	2750	2600	2895	3175	2500	2400

For the given data, one sample t test can be done as follows:

- Enter the data in SPSS data editor.
- Click **Analyze, Compare Means, One Sample T Test**. A dialogue box will open. Move the variable **CFLlife** to the **Test Variable(s)** box.
- Click **Options** to set the confidence interval and then press **Continue** and then **OK**.



The test results on SPSS would give the following two sets of tables:

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
CFLlife	10	2730.9000	285.68104	90.34028

One-Sample Test

Test Value = 3000						
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
CFLlife	-2.979	9	.015	-269.10000	-473.4639	-64.7361

The first table showing One Sample Statistics gives the summary measures and the second table on One Sample Test gives the test values. The

Significance value or the p value shows the risk of Type-1 error. Here the p value is less than .05 indicating that there is significant difference between the hypothesised mean and sample mean. Hence we reject the null hypothesis and thus the claim of the company that the average life of the CFL tubes manufactured by them is 3000 hours.

2. Independent Sample t-Test

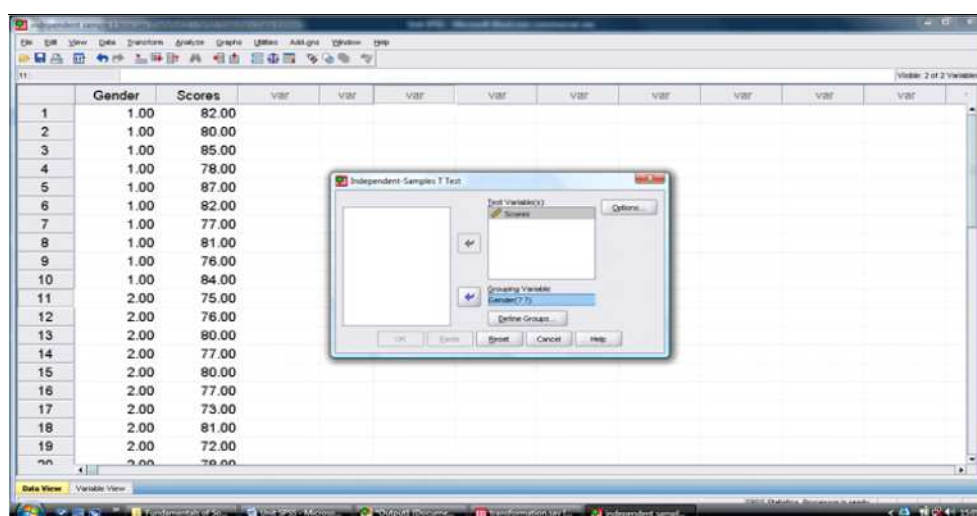
This test is employed to study the difference in two groups. Some of the examples where independent sample t-test can be employed are to study the difference in productivity of high yielding varieties (HYV) between irrigated and unirrigated farms, gender differences in IQ levels etc.

In order to take up an independent sample t test, the following steps need to be followed:

Eg. In a physical education class, the scores of males and females are compared. We need to find out if the mean score of men in class differ significantly from the mean score of women in the class.

	Scores									
Male	82	80	85	78	87	82	77	81	76	84
Female	75	76	80	77	80	77	73	81	72	78

- Open the data file and click on **Analyze** from the command menu.
- Select **Compare Means** and **Independent Sample T-test** from the **Analyze** menu. The **Test Variable** is **Scores** and the grouping variable is **Gender**.



- To define groups, click on define groups to get a new dialogue box. Assign code 1 to group 1 (males) and 2 to group 2 (females).
- Click **Options** button to set the confidence interval of 95%. Click **Continue** and then **OK** to execute the command.

The output will have the following tables:

Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
Scores	1.00	10	81.2000	3.55278	1.12349
	2.00	10	76.9000	2.99815	.94810

The **group statistics** table displays the summary measures of the selected variables. The second table displays the results of the t test. The first step is to look at the Levene's test to decide if you need to assume equal variance or not. The null and alternative hypothesis for Levene's test for equality of variances is as follows:

H_0 : Variances of two groups are equal

H_1 : Variances of two groups are unequal.

In this example, since the significance of F value is .571 which is higher than .05, we accept the null hypothesis. Thus in this example we need to interpret the results related to assumption of equal variance. The next step would be to interpret the t test. In this example, the t value is 2.925 and the significance level is .009 which is less than .05. Hence we reject the null hypothesis for equality of means. In other words, there is significant difference between the marks scored by males and females.

3. Paired Sample t-Test

This test is employed to test the difference in means of the dependent samples. Here, the observations are recorded on the same subject but at different points of time. For example, comparing the performance of the students before and after vacations, sales of a product before and after advertisement etc. These samples are very closely matched and hence are also called dependent sample t test.

Suppose 20 students were given a test before and after studying a module. We want to find out if in general, reading our module, leads to improvement in student's knowledge levels. This can be done using paired t test. The scores of students before and after reading the module are given:

Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Pre-module score	18	21	16	22	19	24	17	21	23	18	14	16	16	19	18	20	12	22	15	17
Post-module score	22	25	17	24	16	29	20	23	19	20	15	15	18	26	18	24	18	25	19	16

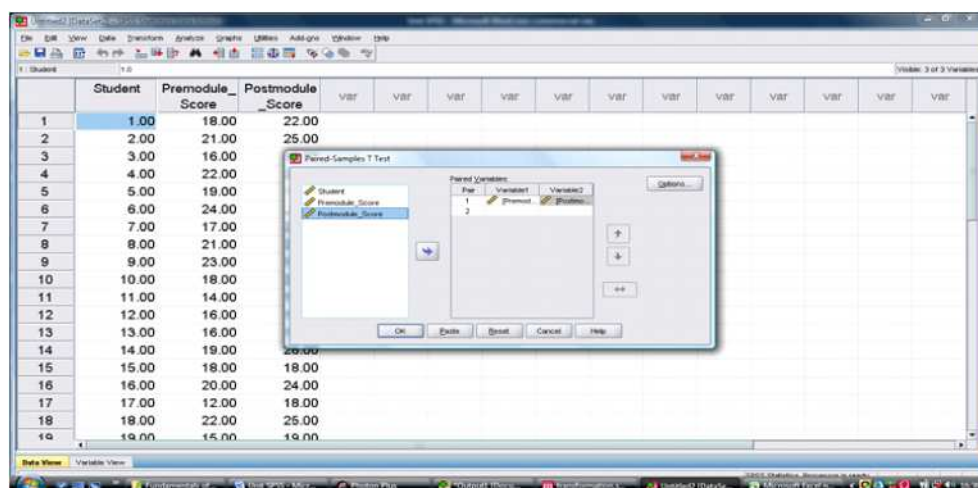
The null and alternative hypothesis for examining the difference in scores of students before and after reading the module is given as:

H_0 : There is no difference in the marks scored before and after reading the module.

H_1 : There is difference in the marks scored before and after reading the module.

This can be done using SPSS by following these steps:

- Open the data file. Click on **Analyze, Compare Means** and **Paired Samples t-test**. A dialogue box opens.
- Select the two variables **Pre-module Score** and **Post-module Score** and shift them to **Paired Variable** list box. Click on **Options** button and set the confidence interval, click **Continue** and then **OK**.



SPSS gives the following output tables:

Paired Samples Statistics

	Mean	N	Std. Deviation	Std. Error Mean
Pair 1 Premodule_Score	18.4000	20	3.15228	.70487
Postmodule_Score	20.4500	20	4.05845	.90750

Paired Samples Correlations

	N	Correlation	Sig.
Pair 1 Premodule_Score & Postmodule_Score	20	.717	.000

Paired Samples Test

	Paired Differences					t	df	Sig. 2-tailed)
	Mean	Std. Devia- tion	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Pair 1 Premodule_ Score - Postmodule_ Score-	2.0500	2.8372	.63443	-3.37787	-.72213	-3.231	19	.004

Paired Samples Statistics table displays the summary measures.

Paired Samples Correlation table gives the correlation between the scores before and after reading of the module.

Paired Sample Test table gives the t value of -3.231 with associated significance value of .004 which is less than .05. Therefore we reject the null hypothesis and can say that there is improvement in marks scored after reading the module as compared to marks scored before reading the module.

4. Chi square Test:

Chi square test measures the association between two or more categorical variables. Chi square tests the hypothesis whether two or more sample drawn from the same population have similar characteristics. Chi square test is used to test statistical significance of results reported in bi-variate tables.

For example, there are 125 children who are three television commercials for breakfast cereals and are asked to pick any one of the three which they liked best. The results are shown in the following table:

Gender	Cereals			Total
	A	B	C	
Boys	30	29	16	75
Girls	12	33	5	50
Total	42	62	21	125

You would like to know if the choice of commercial was related to the gender of the child.

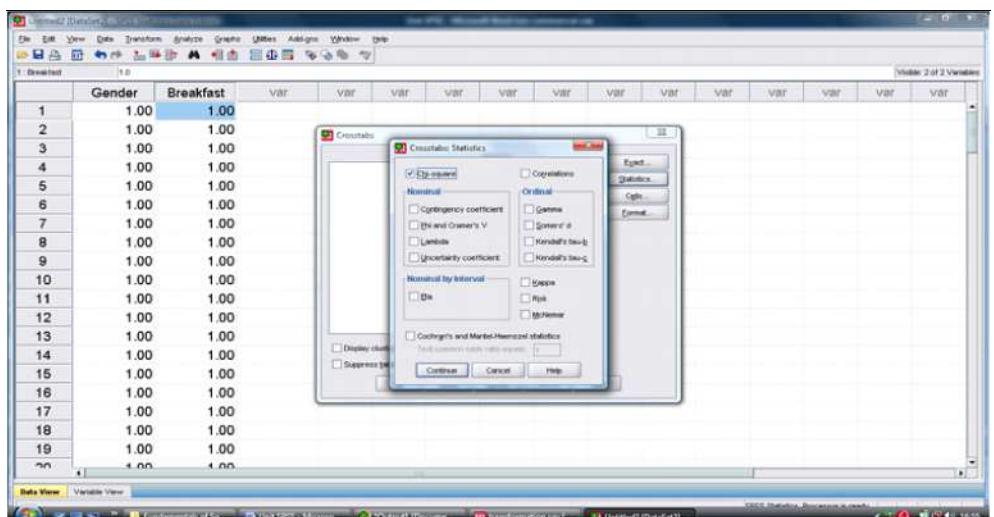
The null and the alternative hypothesis would be defined as follows:

H_0 : There is no association between gender and choice of commercial

H_1 : There is association between gender and choice of commercial

Procedure to conduct Chi square test is as follows:

- Click **Analyze, Descriptives and Crosstabs**. A dialogue box will open.
- Select **Gender** and move it into the **Row** list box and **Location** in the **Column** list box.
- Click **Statistics** button to get another dialogue box. Select **Chi square** check box and then click **Continue** and then **OK**.



You will get the following output tables:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Gender * Breakfast	125	98.4%	2	1.6%	127	100.0%

Gender * Breakfast Crosstabulation

Count

	Breakfast				Total
		1.00	2.00	3.00	
Gender	1.00	30	29	16	75
	2.00	12	33	5	50
	Total	42	62	21	125

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	9.098 ^a	2	.011
Likelihood Ratio	9.254	2	.010
Linear-by-Linear Association	.136	1	.712
N of Valid Cases	125		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 8.40.

Case Proceedings Summary table gives the summary information of the variables.

Crosstabulation of gender and breakfast is given in the second table showing 30 boys and 12 girls prefer the first option, 29 boys and 33 girls prefer the second option and 16 boys and 5 girls prefer the third option.

The Chi square test table gives the test results. The value of Pearson Chi Square is 9.098 and the associated significance is .011 which is less than .05. Therefore the null hypothesis is rejected and we can say that there is association between gender and breakfast commercial preference.

2.8 MULTIPLE REGRESSION USING SPSS

Regression analysis is one of the most widely used mathematical techniques. Regression is commonly used for prediction of values of the dependent variable based on the values of independent variables. Regression analysis with only one independent variable is called simple regression and one with more than one independent variable is called multiple regression.

A single variable regression line may be specified as

$$Y = \alpha + \beta X_1$$

A multiple regression model can be specified as

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \dots + \beta_n X_n$$

The null and alternative hypothesis for a regression analysis would be:

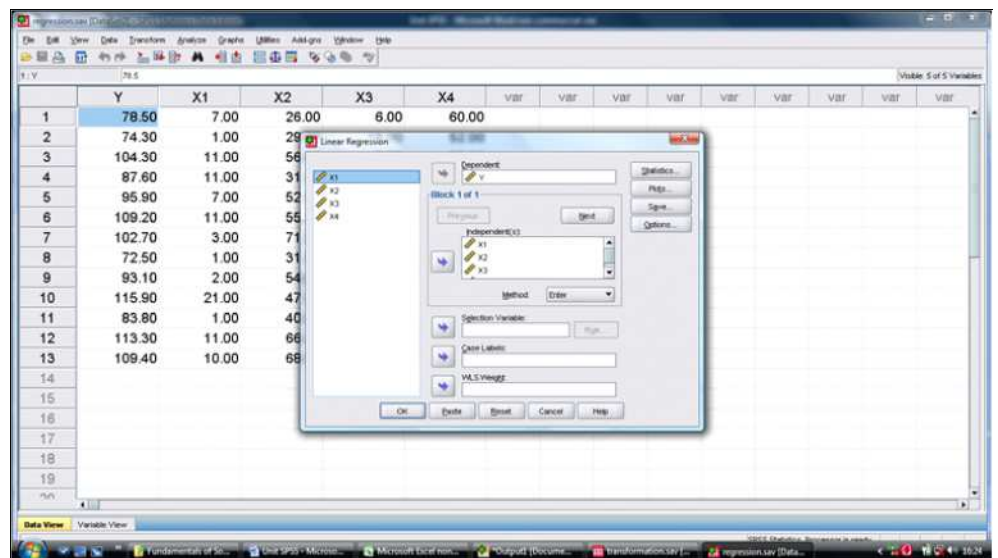
$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

Let us see how multiple regression can be worked out on SPSS taking a hypothetical example in which Y is the dependent variable and X1, X2 and X3 are independent variables.

In order to do multiple regression, you need to follow the steps given below.

- i. Click **Analyze, Regression, Linear** to get a dialogue box.
- ii. Shift Y to the **Dependent** text box and X1, X2, and X3 to the **Independent(s)** text box. Prior to running the regression, you need to choose the method of regression which is to remove or retain the independent variables in the regression equation. If you are not very familiar with the methods, you can choose the default method **Enter**.
- iii. Click **OK** button to execute the command.



The results are given below:

Variables Entered/Removed

Model	Variables Entered	Variables Removed	Method
1	X3, X1, X2 ^a	.	Enter

a. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.600 ^a	.360	.343	6.162

a. Predictors: (Constant), X3, X1, X2

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2473.601	3	824.534	21.716	.000 ^a
	Residual	4404.365	116	37.969		
	Total	6877.967	119			

a. Predictors: (Constant), X3, X1, X2

b. Dependent Variable: Y

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	60.029	2.644		22.702	.000
	X1	.005	.003	.116	1.558	.122
	X2	.457	.264	.134	1.731	.086
	X3	-.489	.071	-.533	-6.888	.000

a. Dependent Variable: Y

Variables Entered/Removed gives information about the independent variables that were entered and removed in the regression analysis. Since we have chosen the Enter method, all the four independent variables have been entered.

In **Model Summary**, R is the square root of R Square and R Square or coefficient of determination is the percentage of total variation of dependent variable that can be explained by the independent variables included in the model. It shows how well the regression line fits the data. In our example, the independent variables explain only 36 percent of variations in the dependent variable. Adjusted R square gives a more realistic picture of the fit of the regression line. Standard Error of the Estimate is the standard deviation of the error term in the model.

The **ANOVA** table is used to test the overall significance of the model. In this case, when we find that the F value is significant showing that the group of independent variables chosen in this model show a statistically significant relationship with the dependent variable.

In the **Coefficients** table, the first column gives a list of variables in the regression equation. The term **constant** refers to the intercept or the height of regression line when it crosses the Y axis. In other words, it gives the value that the dependent variable will assume when the value of all the independent variables is 0. **B** represents the values of the coefficients. Since the natural units of measurement of variables are taken hence the coefficients are known as unstandardized coefficients. Based on the coefficients, the estimated regression equation will be $Y = 60.029 + 0.457X_1 - 0.489X_2 + 0.005X_3$. The standard error is to estimate the t value. The Standardized Beta coefficients are those that would be obtained if you standardized the X and the Y variables which means you arrange them on an identical scale and then execute the regression. The t and Sig. Column have the t value

and 2 tailed p-value which are used to accept or reject the null hypothesis, as the case may be. If you have Sig. Values or p-values less than or equal to .05, then those coefficients are statistically significant and you can reject the null hypothesis and say that the coefficients are significantly different from 0. In our example, the constant and coefficient for X3 are significant at 5 percent and the coefficient for X2 is significant at 10 percent. The coefficient for X1 is not significant.

So far you have read about the use of descriptive statistics in SPSS. You have also learned the usage of testing of hypothesis, correlation and regression using SPSS. Now answer the questions given in *Check Your Progress – 2*.

Check Your Progress 2

Note: a) Write your answer in about 50 words.

b) Check your answer with possible answers given at the end of the unit

1. How will you calculate frequencies using SPSS?

.....

.....

.....

2. Explain the use of crosstabs in SPSS.

.....

.....

.....

2.9 LET US SUM UP

SPSS is a widely used and popular statistical package. The use of statistical package has become indispensable in research. In this unit you have read about the basics of SPSS. This unit deals with data feeding, analysis and interpretation of quantitative data. It makes use of elementary statistical techniques that are explained and put into use in SPSS with the help of lucid examples. Besides the data handling and analysis explained in this unit, SPSS has several other interesting and useful features. However, in this unit we have tried to cover the basic topics that are required by research students in performing statistical analysis. Students are however encouraged to explore other options available on SPSS and put them to use.

2.10 REFERENCES AND SELECTED READINGS

Pandya, K., Bulsari S. And Sinha S. (2014). SPSS in Simple Steps, Dreamtech Press, Delhi.

2.11 CHECK YOUR PROGRESS – POSSIBLE ANSWERS

Check Your Progress – 1

1. Data editor is used to enter data in a spreadsheet format. Data editor lets you see and manipulate your data. Data files are saved as a file

type .sav. SPSS lets you to open many datasets simultaneously. Creating a data file will be dealt with in subsequent sessions.

You can include the data labels by selecting the bars by a left click on any of bar.

Check Your Progress – 2

1. In order to calculate frequencies, the following steps need to be followed. Click on **Analyse** on the main menu and **Descriptive Statistics** followed by **Frequencies** on the sub menu. A dialogue box will open. Shift the desired variable in the **Variable(s)** text box and click **OK**.
2. The **Crosstabs** command is used to get information on the intersection of two or more categorical variables. While information on each level of one categorical variable can be obtained by working out frequencies but it does not provide information on intersection of two variables. This can be done using the crosstabs command as follows:
 - i) Click **Analyze, Descriptive Statistics** and **Crosstabs**. A dialogue box opens. Select **Gender** and move it to the **Rows** list box and select **Place** and move it to the **Columns** list box.
 - ii) Click **Cells** button which opens another dialogue box. This dialogue box allows you to select additional information you want to compute for the selected variables. Select the relevant information and click **Continue** button to go to the previous dialogue box and then click **OK**.

UNIT 3 DATA PROCESSING AND ANALYSIS

Structure

- 3.1 Introduction
- 3.2 Data Measurement and Its Types
- 3.3 Tabulation and Interpretation of Data
- 3.4 Let Us Sum Up
- 3.5 Keywords
- 3.6 References and Selected Readings
- 3.7 Check Your Progress – Possible Answers

3.1 INTRODUCTION

The purpose of data analysis is to identify whether research assumptions were correct or not, and to highlight possible new views on the problem under study. The ultimate purpose of analysis is to answer the research questions outlined in the objectives with the collected data. However, before we look at how variables may be affecting one another, we need to summarize the information obtained on each variable in simple, tabular form, or, in a figure.

Some of the variables may produce numerical (continuous) data, while other variables produce categorical data. In analyzing our data, it is important, first, to determine the type of data that we are dealing with. This is crucial because the type of data used largely determines the type of statistical techniques that should be used to analyze the data. Once the data is processed, tables and graphs are prepared, then the report writing work may be initiated.

After studying this unit, you should be able to:

- define data and describe various types and nature of data.
- describe techniques of data processing, tabulation, and presentation.
- describe and interpret data from tables that have been generated.

3.2 DATA MEASUREMENT AND ITS TYPES

3.2.1 Data Measurement

Measurement is the process of observing and recording the observations that are collected as part of a research effort and the process of assigning numbers to objects or observations. But do we measure abstract concepts like happiness, quality of life, personality, opinion, etc? You have to understand that the qualities of good measurement are: (i) Precise; (ii) Unambiguous; (iii) Free from errors; (iv) Valid; (v) Reliable, and; (vi) Practical.

The tools, which are developed for measurement/ collection of data, should be

- valid to measure the characteristic which it is intended to measure
- reliable to the extent to which an experiment/ test/ or any measuring procedure yields the same result on repeated trials

- sensitive enough to detect differences in a characteristics
- specific enough to represent only the characteristic of interest
- appropriate to the objectives of the study
- provides adequate distribution of response in the study population
- meets the objective of the study

These are key points to keep in mind while developing effective tools for data collection. However, the following factors may affect the validity of the data, so remedial steps should be taken accordingly:

Respondents: reluctance, modesty, having little knowledge about details of the research problem to answer, may not admit ignorance, guessing about response, feel boredom due to long questioning, fatigue, and anxiety, etc.

Situation: lack of support to field investigators, lack of assurance on anonymity and confidentiality, etc.

Investigator/ Interviewer: may reword or reorder questions, style, and look, carelessness in filling the reply, incorrect coding, or calculation of scores.

Instrument: complex words, ambiguous meaning, poor printing, inadequate space for answer, response choice omission.

3.2.2 Types of Data

Data can be classified as either

- i. Primary or secondary data
- ii. Time series, cross section and panel data
- iii. Categorical and numerical data

i. Primary or Secondary Data

Primary Data

Primary data means original data that have been collected specially for the purpose in mind. Research where one gathers this kind of data is referred to as field research. Tools used for gathering primary data, include a questionnaire, an interview schedule etc.

Secondary Data

Secondary data are data that have been collected for another purpose. Also, after performing statistical operations on primary data the results become known as secondary data. Research where one gathers this kind of data is referred to as desk research. The source for gathering secondary data include for example data from a book.

ii. Time series, Cross-Section and Panel Data

Cross-sectional data refers to data collected by observing many subjects (such as individuals, firms or countries/regions) at the same point of time, or, without regard to differences in time. Analysis of cross-sectional data usually consists of comparing the differences among the subjects. For example, we want to measure current obesity levels in a population. The cross-sectional data provides us with a snapshot of that population, at that one point in time. Note that we do not know, based on one cross-sectional sample, if obesity is increasing or decreasing; we can only describe the current proportion.

Time series data is also known as longitudinal data, which follows one subject's changes over the course of time. For example, the average production of wheat from 1990 to 2009.

Panel data combines both cross sectional and time series data and looks at multiple subjects and how they change over the course of time. Panel analysis uses panel data to examine changes in variables over time, and differences in variables between subjects.

iii. Categorical and Numerical Data

Categorical or Nominal: Data that can be divided into categories or groups such as male and female, and can take only discrete and not decimal values. They are called categorical or nominal data. There are two types of categorical data: they are nominal or ordinal. In nominal data, the variables can be numbered and the numbers are simply labels. You can count but not order or measure the nominal data eg. sex. The ordinal data on the other hand can be ranked or ordered but not measured eg. house numbers or test levels in music.

Continuous or Numerical: Data that can take any value, including decimal values, are called continuous data. Data that can be measured on a scale is said to be scalar. We speak of numerical data if they are expressed in numbers. There are two types of numerical data: they are discrete or continuous. Discrete data are a distinct series of numbers. Discrete data cannot be divided and can take only certain values eg. number of students. Continuous data on the other hand can be divided infinitely eg. height of students.

In this section, you studied about the data measurement and its types. Now answer the questions given in *Check Your Progress - 1*.

Check Your Progress-1

Note: a) Write your answer in about 50 words.

b) Check your answer with possible answers given at the end of the unit

1. What is panel data?

.....

.....

.....

2. What do you understand by primary data?

.....

.....

.....

3.3 TABULATION AND INTERPRETATION OF DATA

The following operations need to be done to bring data into a presentable form:

— Data coding, editing and feeding

- Data tabulation
- Figures and graphs

We will now discuss these operations, one after the other, for both categorical and numerical data.

3.3.1 Data Coding, Editing and Feeding

Once data is collected the researcher thinks about the processing and analysis of the data. This is a crucial stage of research work. Here, the researcher has to consult the guide and other academicians who have done research in the related field. Therefore, it is advised that data processing must be planned in advance, and, necessarily, during the time of questionnaire formulation; and, during the time of data collection. Nowadays, with the advent of the computer, researchers might think it is the job of computer assistant, but it is not so. Before either giving it to the computer assistant, or doing it manually, a lot of activities have to be completed: (i) editing of the data which you received from the field; (ii) coding of the data; (iii) preparation of the master chart if the data is to be computed manually; and (iv) presentation of data.

(i) Editing of the data

The editing of data is the first step of data processing and analysis. After collecting data, either through the questionnaire or schedule, you have to edit it. You have to carefully check for missing and wrongly entered data. Therefore, in large scale surveys and research, the company undertaking the research project appoints supervisors, or, editors, for proper checking of data. An example of editing of data is given below.

Example – Suppose that you conduct a survey of the occupation status of the people in a village, for which you have to collect data from heads of households. The questionnaire for the purpose is as follows

Name of the Respondent:	- Rama Singh
Age:	- 56
Sex:	- Male
Cast:	- General
Occupation:	- Private Service
Income of the Respondent:	- 10,000
Wife's name:	- Rita Singh
Age:	- 51
Occupation:	- House wife
Monthly Income:	- 800/-

In this questionnaire, the wife of Rama Singh is a housewife, but she earns Rs. 800. She may be a housewife, but by knitting she may be earning Rs. 800 per month. Therefore, her occupation may be categorized as self-employment. Like this, you have to cross check the data and try to get accurate information which will minimize error and strengthen your research findings. Therefore editing of the data is very essential before sending it for processing.

(ii) Coding of the data

The questionnaire must be properly coded. Before either, sending it for feeding into the computer, or entering it in the master chart, the coding of data is necessary. The coding of data will make data entry easy. Coding of data means assigning numerical symbol 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. An example of coding of the marital status is given below.

Marital Status

Married:	-	01
Separated:	-	02
Widowed:	-	03
Divorced:	-	05
Never Married:	-	06
Any Other:	-	07

The coding of data and checking that codes are properly inserted in the questionnaire must be done during the editing of the questionnaire. It is always better to prepare a code book for your questionnaire. The coding can also be done during the time of data collection, if the code book is available with you.

(iii) Entering data in the Master Chart

If you are doing tabulation manually, it is always wise to enter the data into a master chart. The master chart is a large sheet which will enable you to enter all the codes of different variables into it. It will help you to generate tables easily. An example of master chart is given below.

Master Chart

Sl. No	Name of the respondent	Variables				
		Sex	Age	Marital Status	Education	Occupation
1						
2						
3						
4						
5						
Total						

In the master chart you can enter the data of 14 sample respondents. Likewise you can expand the number of respondents in the columns and variables in the rows. It is always better to enter code (numerical number) in the master chart.

(iv) Entering Data into the Computer

Computers are widely used for the analysis of data. It makes the calculation much faster. The excel sheet and the SPSS package can be used in social

science research. The following steps are used in the analysis of data by using the SPSS package.

- i. Entering of data in the SPSS statistical package
- ii. Selection of analytical procedure from the menu
- iii. Selection of variables for analysis
- iv. Examination of the output.

(v) Presentation of Statistical Data

Statistical data are collected to serve a purpose. Therefore, data may be presented in such a way that it may be easily grasped and the conclusion may be drawn promptly. Generally, the following three methods are used for the presentation of statistical information.

a) Textual Presentation: In this method, statistical information is presented in text form. Generally, this type of presentation is made in a descriptive way. It requires careful reading of the text in order to grasp the meaning and significance of the facts and figures given, therein. But, for most people, it is not a suitable and effective method of presentation of statistical information, because it is not easy for the reader to single out the individual information and figures. The advantage of this method is that an ordinary person can prepare and present the text and a layman can read it and grasp it.

b) Tabular Presentation: In this method, statistical information is presented in the form of a table. Facts and figures are gathered, and then, incorporated in tables. Generally, this type of presentation is made in a tabular form with rows and columns. Tables summarize statistical data in a logical and orderly manner. The main advantage of this method is that tables are brief, concise, and contain only relevant figures. Tables also facilitate the comparison of figures. The only disadvantage of this method is that the presentation of tables and the interpretation of data require some skills and techniques.

c) Graphical Presentation: In this case, statistical information is presented in the form of graphs and charts. Facts and figures are gathered first, and then, they are depicted in graphs and charts for presentation. Generally, this type of presentation is made through figures, diagrams, charts, or graphs. The main advantage of this method is that the facts and figures become more attractive and appealing to eye. A disadvantage of this method is that facts cannot be shown in detail and accurately.

3.3.2 Data Tabulation

Tabulation is an orderly and systematic arrangement of numerical data presented in rows and columns for the purpose of information, comparison, and interpretation. So, a statistical table is a systematic arrangement of statistical data into rows and columns. It summarizes the data in a logical and orderly manner for the purpose of presentation, comparison, and interpretation. Tabulation is, thus, a scientific process and means of recording, statistical data in a systematic and orderly manner.

A statistical table has the following five parts.

i) Title: each table must have a title which convey the contents of the table. It should be clear, concise and self-explanatory. It should be written on the top of the table.

ii) Stub: this is a column used for mentioning the items and their heading. It is the left most column of the table. A stub is generally marked with rows and in each row, an item is mentioned. The stub should be clear and self-explanatory.

iii) Caption: this is the heading for columns, other than the stub. It is the upper part of the Table. Caption should be properly columned and worded. Sometimes, below the caption, the units of measurement and column-numbers, is called a box head.

iv) Body: the main part of the table. It contains the data which are exhibited in the table. The figures inserted therein should be distinct.

v) Source & Footnote: the last part of a table. If the researcher is procuring data from a secondary source, then the source of the data needs to be mentioned. For example, if you are citing data from the Census of India then the year and the department and the state need to be mentioned. After mentioning the sources, the researcher has to provide a footnote, e.g., in the same cell, if you are giving the figure, and in parenthesis, the percentage to the total, then, it must be mentioned in the footnote. An example of a table is given below:

Title		
Caption		
STUB	BODY	BOX HEAD

Source:

Footnote:

3.3.3 Types of Tabulation

Tabulation is done based on the data. The following types of tables are generally constructed.

- i. Construction of frequency distribution table
- ii. Construction of cross- tabulation
- iii. Construction of figures and graph

i. Construction of Frequency Distribution Table

A frequency distribution table can be of two types:

- Simple frequency distribution
- Grouped frequency distribution

In constructing a simple frequency distribution, the observations are not divided into groups or classes. Only individual values are shown, whereas,

in the grouped frequency distribution the observations are divided into groups or classes.

Here is a simple frequency distribution table.

Table 3.1: Marks obtained by class fifth students in mathematics

Marks	Tally of marks	Frequency (No. of students)
20	HHH	5
21	HHH HHH II	12
25	HHH HHH III	13
30	HHH HHH	10
Total		40

Table 3.2: Grouped frequency distribution

Marks obtained	Tally of marks	Frequency
1-10	HHH HHH I	11
11-20	HHH HHH HHH HHH	20
21-30	HHH HHH HHH HHH I	21
31-40	HHH HHH III	13
41-50	HHH HHH	9

While preparing the group frequency distributions, the following points have to be taken into consideration.

- The groups must not overlap, otherwise there will be confusion about which group a measurement belongs to.
- There must be continuity from one group to the next, which means that there must be no gaps. Otherwise, some measurements may not fit in a group.
- The groups must range from the lowest measurement to the highest measurement so that all of the measurements have a group to which they can be assigned.
- The groups should normally be of an equal width, so that the counts in different groups can easily be compared.

ii. Construction of Cross Tabulation

So far, we have made tables containing frequency distributions for one variable at a time, in order to partially describe our data. Depending on the objectives of our study, and the study type, we may have to examine the relationship between several of our variables at the same time. For this purpose it is appropriate to construct cross tabulation of data. Depending on the objectives and the type of study, different kinds of cross tabulations may be required. The examples of cross tabulation are given below. Here, three different types of cross tabulation of data have been given.

Example 1: A study was carried out on the degree of job satisfaction among doctors and nurses in rural and urban areas. To describe the sample, a cross tabulation was constructed which included the sex and the residence (rural,

or urban) of the doctors and nurses interviewed. This was useful, because in the analysis, the opinions of male and female staff had to be compared separately for rural and urban areas.

Table 3.3: Type of teachers by residence

Residence	Type of teachers		Total
	Principles	Teachers	
Rural	10 (16%)	69 (38%)	79 (33%)
Urban	51 (84%)	113(62%)	164 (67%)
Total	61(100%)	182 (100%)	243(100%)

Interpretation: Table 3.3 shows that a higher percentage of teachers than principals work in rural areas, but, that, overall, a greater proportion of teaching staff works in urban areas (67%).

Table 3.4: Sex of teachers by residence

Residence	Sex of teachers		Total
	Male	Female	
Rural	54 (43%)	25 (21%)	79 (33%)
Urban	71 (57%)	93 (79%)	164 (67%)
Total	125(100%)	118 (100%)	243(100%)

Interpretation: it can be concluded from Table 3.4 that there are more male teachers serving in rural areas than females.

To obtain an overview of the distribution of principals and teachers by gender in rural and urban areas, we can construct the following two-by-four cross-table.

Table 3.5: Residence and sex of principals and teachers

Teaching staff		Residence		Total
		Rural	Urban	
Principals	Males	8 (10%)	35 (21%)	43 (18%)
	Females	2 (3%)	16 (10%)	18 (7%)
Teachers	Males	46 (58%)	36 (22%)	82 (34%)
	Females	23 (29%)	77 (47%)	100 (41%)
Total	79(100%)	164 (100%)	243(100%)	

Interpretation: this table shows, at a glance, that more of male teachers serve in rural areas. It also indicates that males outnumber females as Principals, (18% M > 7% F Principals) but, that, overall, there are more female teachers, and that the females are mainly clustered in towns.

The data in the tables is usually listed in absolute figures, as well as in relative frequencies (percentages or proportions). As already seen in Unit 1, for numerical data (such as age) the mean, median, and/or mode, with standard deviation may be calculated as well, to describe the sample.

General hints while constructing tables

- Make sure that all the categories of the variables presented in the tables have been specified, and that they are mutually exclusive (i.e., no overlaps and no gaps) and are exhaustive.
- When making cross-tabulations, check that the column and row counts correspond to the frequency counts for each variable.
- Also, check that the grand total in the table corresponds to the number of subjects in the sample. If not, an explanation is required. This could be presented as a footnote. (missing data, for example.)
- Think of a clear title for each table. Also, be sure that the headings of rows and columns leave no room for misinterpretation.
- Number your tables and graphs and keep them together with the objectives to which they are related. [Numbering of tables and graphs in a chapter (e.g., 4) may be like table 4.1, table 4.2 and Figure 4.1 Figure 4.2, etc.]. This will assist in organizing your report and ensure that work is not duplicated.

iii. Construction of figures and graph

If your report contains many descriptive tables, it may be more readable if you present the most important ones in figures. The most frequently used figures for presenting data include

• Bar charts	for categorical data
• Pie charts	
• Histograms	for continuous data
• Line graphs	
• Scatter diagrams	
• Maps	

We will now look at example of the abovementioned figures that can be used for presenting data.

1. Bar Chart

The data can be presented in a bar chart, using either absolute frequencies or relative frequencies/percentages and an example is given in Figure 3.1 below.

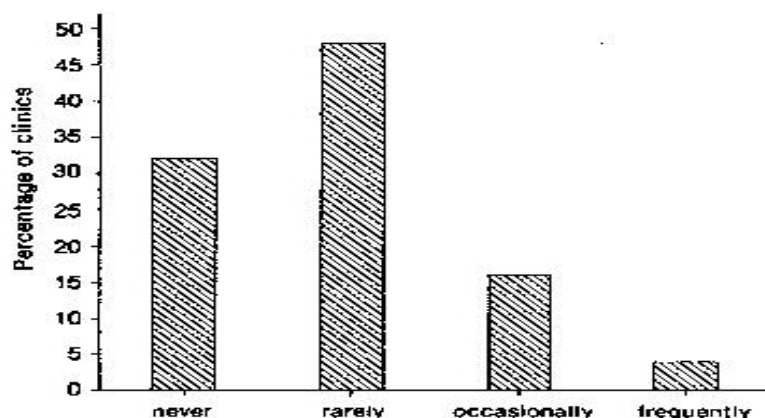


Figure 3.1: Relative frequency of shortage of anti-malaria drugs in rural health institutions (n=148)

Note that the sample size must be indicated if you present the data in percentages.

2. Pie Charts

A pie chart can be used for the same set of data, providing the reader with a quick overview of the data presented in a different form. A pie chart illustrates the relative frequency of a number of items. All the segments of the pie chart should add up to 100%.

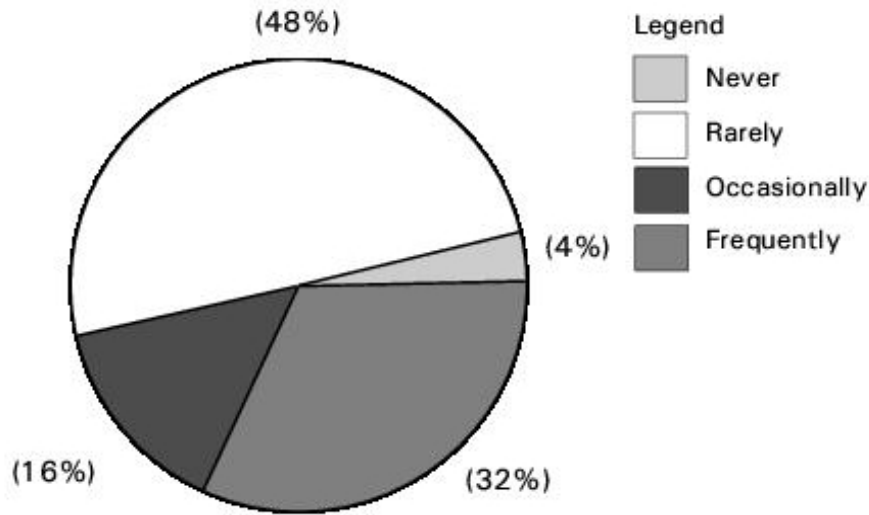


Figure 3.2: Relative frequency of shortage of anti-malaria drugs in rural health institutions (n=148)

3. Histograms

Numerical data are often presented in histograms, which are very similar to the bar charts which are used for categorical data. An important difference, however, is that in a histogram the bars are connected (as long as there is no gap between the data), whereas in a bar chart the bars are not connected, as the different categories are distinct entities. An example of histogram is given in Figure 3.3.

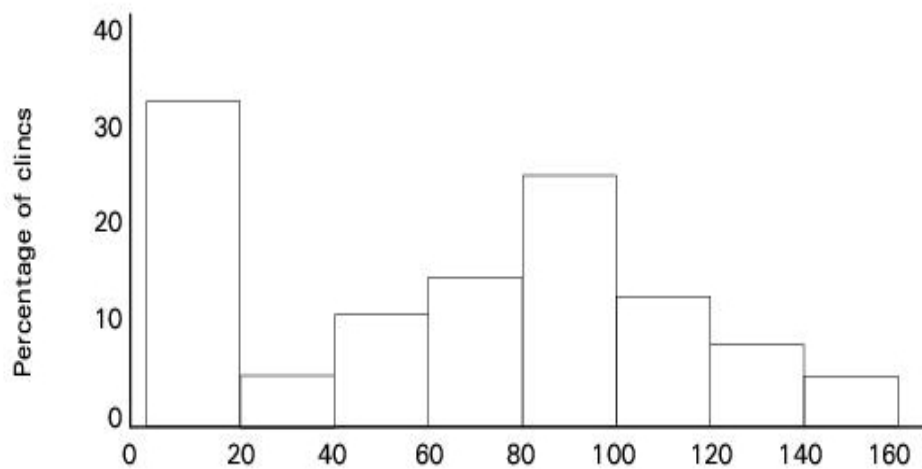


Figure 3.3: Percentage of clinics treating different numbers of malaria patients in one month (n=80).

No. of patients per month

4. Line Graphs

A line graph is particularly useful for numerical data if you wish to show a trend over time. An example of a line graph is given in Figure 3.4.

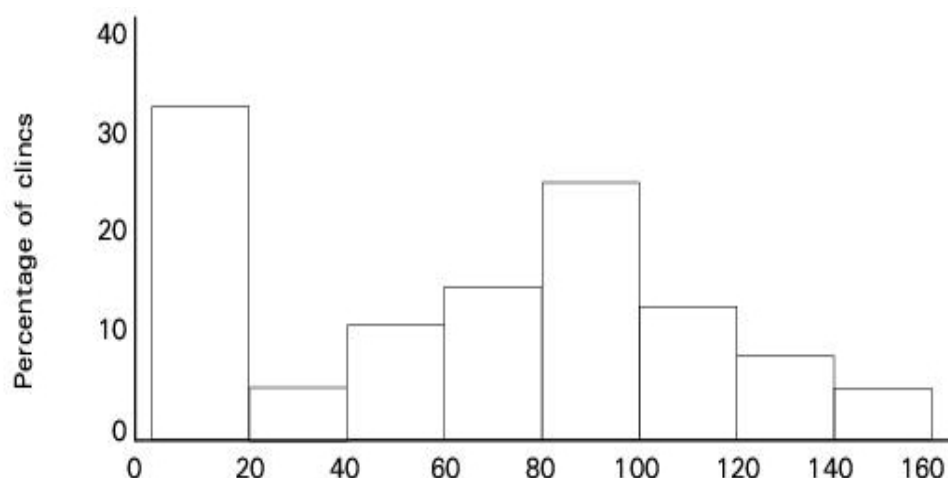


Figure 3.4: Daily number of malaria patients at the health centers in District X

It is easy to show two or more distributions in one graph, as long as the difference between the lines is easy to distinguish. Thus, it is possible to compare frequency distributions of different groups, i.e., the age distribution between males and females, or cases and controls.

5. Scatter diagrams

Scatter diagrams are useful for showing information on two variables which are possibly related. The example of a scatter diagram, given below, is used where we are dealing with the concepts of association and correlation.

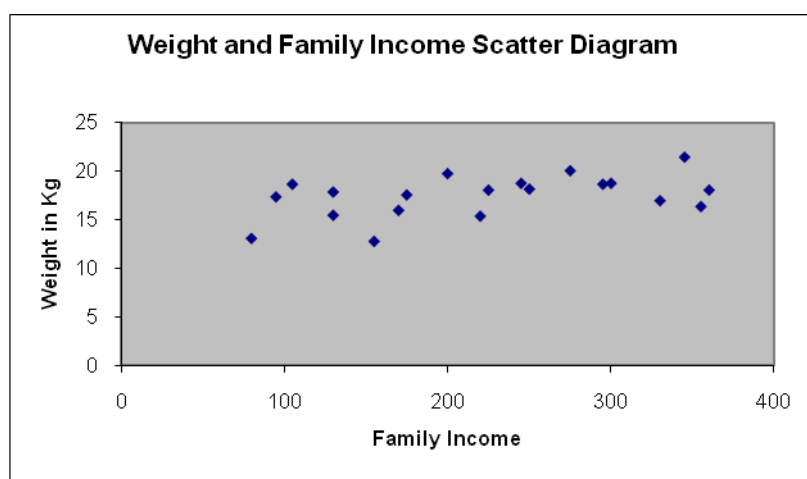


Figure 3.5: Weight of five-year-olds according to annual family income

Note: It is important that all figures presented in your research report have numbers, clear titles, and are clearly labeled (or keyed).

6. Maps

In addition to the figures above, the use of maps may be considered to present information. For instance, the area, where a study was carried out, can be shown in a map. If the study explored the epidemiology of cholera, a map could be produced showing the geographical distribution of cholera cases,

together with the distribution of protected water sources, thus illustrating that there is an association. If the study related to vaccination coverage, a map could be developed to indicate the clinic sites and the vaccination coverage among under-fives in each village, perhaps showing that home-clinic distance is an important factor associated with vaccination status.

In this section, we discussed about the tabulation and interpretation of data. Now answer the questions given in *Check Your Progress - 2*.

Check Your Progress-2

Note: a) Write your answer in about 50 words.

b) Check your answer with possible answers given at the end of the unit

1. What is meant by coding of data?

.....

.....

.....

2. What is a pie chart and where is it used?

.....

.....

.....

3.4 LET US SUM UP

Understanding data quality and its measurement are of utmost importance for any researcher. Poor quality data cannot be analyzed properly, and, may also give results which may not be valid, or, sometimes give results which could have adverse consequences for society if those results are used as a base for policy making. The presentation of data needs to be made in simple, as well as cross table format. It is always advisable to prepare a cross tabulation plan as per study requirements, and to present data in graphical way for clarity.

3.5 KEYWORDS

Data Measurement: measurement is the process of observing and recording the observations that are collected as part of a research effort.

Type of Data: broadly there are two types of data: (i) quantitative and; (ii) qualitative which can be further classified as categorical, nominal and continuous data.

Data Quality: the quality data can be characterized as: (i) precise, (ii) unambiguous, (iii) free from errors, (iv) valid, (v) reliable, and (vi) practical.

Data Processing: means the generation of frequency distribution and cross tabulation and calculation of other statistical measures.

Frequency Distribution: preparation of tables which distribute respondents according to a particular characteristic of sample, or research outcome.

Cross Tabulation: this is a process of generating tables giving the outcome of interest in columns, and various characteristics of respondents, or factors affecting outcomes in rows.

Data Interpretation: is drawing valid and meaningful conclusions from the tables generated with the help of collected data.

Report Preparation: is the process of documenting the whole process of research conducted to identify the problem, or to prove some relationships, or for proving the success of some programme related activities.

3.6 REFERENCES AND SELECTED READINGS

Gibaldi, J. (1995), *MLA Handbook for Writers of Research Papers*, Modern Language Association of America, New York.

Yang, J. T. et al. (1996), *An outline of Scientific Writing: For Researchers with English as a Foreign Language*, World Scientific Publishing, Singapore.

Trochim, W. M. (1999), *The Research Methods Knowledge Base*, 2nd Edition, Online textbook, URL: <<http://www.socialresearchmethods.net/kb/>>

Training modules of International Development Research Council, Canada.

Health System Research Modules published by WHO.

Modules on Primary Health Care, Agha Khan Foundation, Geneva.

3.7 CHECK YOUR PROGRESS - POSSIBLE ANSWERS

Check Your Progress 1

- 1) Panel data combines both time series and cross sectional data and looks at multiple subjects and how they change over the course of time. Panel analysis uses panel data to examine changes in variables over time, and differences in variables between subjects.
- 2) Primary data is original data that have been collected specially with a purpose in mind. Research where one gathers this kind of data is referred to as field research, for example: a questionnaire.

Check Your Progress 2

- 1) A questionnaire must be properly coded. Before feeding it into a computer, or entering the data into the master chart, coding of data is necessary. The coding of data will make data entry easy. Coding of data means assigning a numerical symbol 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.
- 2) A pie chart can be used for providing the reader with a quick overview of the data presented in a different form. A pie chart illustrates the relative frequency of a number of items. All the segments of the pie chart should add up to 100%.

UNIT 4 REPORT WRITING

Structure

- 4.1 Introduction
- 4.2 Types of Report
- 4.3 Writing the Research Report
- 4.4 The Preliminary Pages of Research Report
- 4.5 Main Components or Chaptering the Research Report
- 4.6 Style and Layout of the Report
- 4.7 Common Weaknesses in Report Writing and Finalizing the Text
- 4.8 Let Us Sum Up
- 4.9 References and Selected Readings
- 4.10 Check Your Progress – Possible Answers

4.1 INTRODUCTION

A research report is considered a major component of any research study as the research remains incomplete till the report has been presented or written. No matter how good a research study, and how meticulously the research study has been conducted, the findings of the research are of little value unless they are effectively documented and communicated to others. The research results must invariably enter the general store of knowledge. Writing a report is the last step in a research study and requires a set of skills somewhat different from those called for in actually conducting a research.

After reading this unit you will be able to:

- describe the various steps involved in writing a research report.
- explain the various components of a research report
- identify common mistakes committed while writing a research report.

4.2 TYPES OF REPORT

Research reports vary greatly in length and type depending on the subject. For example banks and other financial institutions prefer short balance sheet type of tabulations for their annual report. In mathematics, the report may consist of many algebraic notations, whereas a chemists report may be in the form of symbols and formulae. Students of literature usually write a long report critically analysing a writer or book.

The news items found in newspapers are also a form of report writing. Other examples of reports include book reviews, reports prepared by government bureaus, PhD theses, etc. Any research investigation may be presented in the form of a technical report, a popular report, an article, a monograph, or, at times, even in the form of an oral presentation. The technical report is prepared for specialists who have interest in understanding the technical procedure and

terminology used in the research project. The report will be in technical language. In the technical report, the main emphasis is on: (i) the methods employed; (ii) assumptions made in the course of study and; (iii) the detailed presentation of the findings, including their limitations and supporting data.

Popular data is intended for persons who have limited interest in the technical aspects of the research methodology and research findings. The audience will include laymen and even top executives who want summary reports. The popular report is one which gives emphasis on simplicity and attractiveness. The simplification should be sought through clear writing, minimizing of technical, particularly mathematical details, and liberal use of charts and diagrams. Attractive layouts along with large print and many subheadings is another feature of a popular report. In such a report, emphasis is given on practical aspects and policy implications.

4.3 WRITING THE RESEARCH REPORT

Once the data collection and analysis work is over, the researcher will start writing the research report. Social and development research reports need to

- have a logical, clear structure
- be to the point
- use simple language, and have a pleasant layout

Just as an architect has to draw a layout plan for a house that is being designed, you first have to make an outline for your report. This outline will contain a head, a body, and a tail. **The head** consists of a description of your problem within its context (the country and research area), the objectives of the study and the methodology followed. This part should not comprise more than one quarter of the report, otherwise it becomes top-heavy. **The body** will form the bigger part of your report: it will contain the research findings. **The tail**, finally, consists of the discussion of your data, conclusions, and recommendations.

Before you start writing, it is essential to group and review the data you have analysed by objective. Check whether all data has indeed been processed and analysed as you planned in the research protocol/proposal which is duly approved. Draw major conclusions and relate these to the research literature. Again, you may be inspired to go back to your raw data and refine your analysis, or to search for additional literature to answer questions that the analysis of your data may evoke. Compile the major conclusions and tables or quotes from qualitative data related to each specific objective. You are now ready to draft the report.

The research report will have, broadly, three parts.

Part I: The Preliminary Pages

Part II: The Main Text of the Research Report

Part III: The End Matter

4.4 THE PRELIMINARY PAGES OF RESEARCH REPORT

The preliminary pages of the research report should have the following main constituents.

- Title and cover page
- A foreword
- Preface
- Acknowledgements
- Table of contents
- List of tables
- List of figures
- List of appendices
- List of abbreviations
- Executive Summary

(i) Title and Cover page

The cover page should contain the title, the names of the authors with their designations, the institution that is publishing the report with its logo, (e.g., Health Systems Research Unit, Ministry of Health), the month, and the year of publication. The title could consist of a challenging statement or question, followed by an informative subtitle covering the content of the study and indicating the area where the study was implemented. However, this is suggestive in nature and should not be considered standard. It would be appropriate if the cover page is designed by an expert in computer graphics who may be suggested to include some important photograph related to identity of organization or problem under study or from the field within the background. Design software may be used. An example of a title of a research report is given in the box below.

Title of the research report

Labour Migration and its Implication on Rural Economy of Indo-Gangetic Plains of India

(ii) Foreword

A foreword is usually a short piece of writing found at the beginning of a book or other piece of literature, before the **introduction**. This may or may not be written by the primary author of the work. Often, a foreword will tell of some interaction between the writer of the foreword and the story, or, the writer of the story. A foreword to later editions of a work often explains how the new edition differs from previous ones. Unlike a preface, a foreword is always signed. An example of a foreword is given in the box below.

Foreword

Migration of all kinds, particularly income seeking migration across state boundaries, has attracted much attention in recent scholarly and policy literature. This study provides sufficient evidences of the effect of labour migration, more specifically, male outmigration on the rural economy of the Indo-Gangetic region. The number of districts of high and moderately high male outmigration has increased. The findings reveal the holistic scenario of migration led changes in agricultural

and household domains. I am sure that this volume would be of great interest to researchers, policy makers, and development agencies while framing strategies for agricultural and rural development.

(iii) Preface

A **preface**, by contrast, is written by the author of the book. A preface generally covers the story of how the book came into being, or how the idea for the book was developed; this is often followed by thanks and acknowledgments to people who were helpful to the author during the time of writing. A preface is an introduction to a **book** or other **literary** work written by the work's author. An example of preface is given in the box below.

Preface

The present study was conducted in three states of Bihar, Uttar Pradesh, and Punjab to study various aspects of labour migration, and its impact on rural economy in the Indo-Gangetic plains in India. The study focused on labour outmigration across two states of the Indo-Gangetic Region and in-migration in Punjab. The results of this study would help researchers, policy makers and planners as well as development agencies in addressing various issues of labour migration and its implication in India.

(iv) Acknowledgements

It is good practice to thank those who supported you technically or financially in the design and implementation of your study. You should not forget to thank your research guide and your employer too, who has allowed you to invest time in the study; and, the respondents may be acknowledged. You should not forget to acknowledge the contribution of computer professionals, library staff, local officials, and the community at large that provided the information. Acknowledgements are usually placed right after the title page or at the end of the report, before the references. An example of acknowledgement is given in the box below.

Acknowledgements

I take this opportunity to thank the (Name of the funding agency) for providing funds and facilities for the project. I offer my sincere thanks to the (Name of your employer) for his encouragement and support for pursuing this study. I am also grateful to the head, (Name of your department) for providing all needed support, encouragement, and technical guidance. All the Research Associates, Senior Research Fellows and technical assistants working under the project deserve special appreciation for their hard work and sincere efforts in completing this project.

(v) Table of Contents

A table of contents is essential. It provides the reader a quick overview of the chapters with major sections and sub sections of your report, and page references, so that the reader can go through the report in a different

order, or skip certain sections. The sections and sub sections within each chapter may be given numbers that are specific to the chapter. For example, a section in chapter III may be given no as 3.1; and, a sub section as 3.1.1. An example of a table of contents is given below.

Contents

S. No.	Contents	Pages
1	Introduction	
2	Review of Literature	
3	Methodology	
3.1	Data	
3.2	Analytical Tools	
3.3	Profile of Area Under Study	
4	Research Findings	
4.1	Macro Level Evidences	
4.2	Evidences from filed Survey	
5	Discussion	
6	Conclusions and Policy Implications	
7	References	
	Appendix	

(vi) List of Tables

If you have many tables or figures, it is essential to list these also in a table of contents with formatted page numbers. The initial letters of the key words in the title are capitalized and no terminal punctuation is used. An example is given below.

List of Tables

S.No.	Name of the Table	Pages
2.1	Sampling Pattern of Households in the Study Area	
3.1	Migrants by Last Residence in India	
3.2	Total Inter-State Migrants by Place of Birth in Major States	
3.3	Social Characteristics of Households in the Study Area	

(vii) List of Figures

The list of figures appears in the same format as the list of tables, titled List of Figures.

(viii) List of Appendices

The appendices will contain any additional information that the researcher has collected while carrying out the study. It may be a questionnaire, a letter

of appreciation, a government notification, etc. The list of appendices appears in the same format as the list of tables.

(ix) List of Abbreviations (optional)

If abbreviations or acronyms are used in the report, these should be stated in full in the text the first time that they are mentioned. If there are many, they should be listed in alphabetical order as well. The list can be placed before the first chapter of the report.

The table of contents and lists of tables, figures, abbreviations should be prepared last, as only then can you include the page numbers of all chapters and sections, sub-sections in the table of contents. Then, you can also finalise the numbering of figures and tables and include all abbreviations. An example of a List of Abbreviations follows.

(x) List of Abbreviations

List of Abbreviations	
AI:	Agreement Index
CMIE:	Centre for Monitoring of Indian Economy
CV:	Coefficient of Variation
DEA:	Data Envelopment Analysis

(xi) Executive Summary

The summary should be written only after the first or even the second draft of the report has been completed. It should contain

- a very brief description of the problem (**Why** this study was needed)
- objectives (**What** has been studied)
- the place of study (**Where**)
- the type of study and methods used (**How**)
- the major findings and conclusions
- the major (or all) recommendations.

The summary will be the first (and for busy programme manager/decision makers most likely the only) part of your study that will be read. Therefore, it demands thorough reflection and is time consuming. Several drafts may have to be made, each discussed by the research team as a whole.

As you may have collaborated with various groups during the drafting and implementation of your research proposal, you may consider writing **different summaries** for each of these groups. For example, you may prepare different summaries for policymakers and programme managers, for implementing staff of lower levels, for community members, or for the public at large (newspaper, TV). In a later stage, you may write articles in scientific journals.

In this section, we discussed about the types of report and the contents to be included in the preliminary pages of research report. Now answer the following questions given in *Check Your Progress 1*.

Check Your Progress-1

Note: a) Write your answer in about 50 words.

b) Check your answer with possible answers given at the end of the unit

1. What should be included on the cover page of a research report?

.....

.....

.....

2. What is the importance of writing an acknowledgement in a research report and where should it be placed?

.....

.....

.....

4.5 MAIN COMPONENTS OR CHAPTERING THE RESEARCH REPORT

The Main Text includes the following chapters

- Introduction
- Review of Literature
- Methodology
- Research Findings
- Discussion
- Conclusion and Recommendations
- Summary

4.5.1 Chapter 1: Introduction

The introduction is a relatively easy part of the report that can best be written after a first draft of the findings has been made. It should certainly contain some relevant (environmental/ administrative/ economic/ social) background data and information about the topic on which you are carrying out research for example if you are doing research on primary education, then a brief about the status of primary education, such as their number, state-wise break up, expenditure on primary education, etc., need to be described. You may make additions to the corresponding section in your research proposal, including additional literature, and use it for your report.

Then, the statement of the problem should follow, again, revised from your research proposal with additional comments and relevant literature collected during the implementation of the study. It should contain a paragraph on what you hope/ hoped to achieve from the results of the study. Enough background should be given to make clear to the reader why the problem was considered worth investigating.

The general and specific objectives should also be included in this chapter. If necessary, you can adjust them slightly for style and sequence. However, you should not change their basic nature. If you have not been able to meet

some of the objectives of the project, this should be stated in the methodology section, and in the discussion of the findings. The objectives form the heart of your study. They determined the methodology you chose and will determine how you structure the reporting of your findings.

4.5.2 Chapter 2: Review of Literature

Global literature can be reviewed in the introduction to the statement of the problem if you have selected a problem of global interest. Otherwise, relevant literature from individual countries may follow as a separate literature review after the statement of the problem. A literature review is a body of text that aims to review the critical points of current knowledge and/or methodological approaches on a particular topic. Literature reviews are **secondary sources**, and, as such, do not report any new or original experimental work. Its ultimate goal is to bring the reader up to date with current literature on a topic, and forms the basis for another goal, such as future research that may be needed in the area.

A well-structured literature review is characterized by a **logical** flow of ideas; current and relevant references with consistent, appropriate **referencing style**; proper use of **terminology**; and an unbiased and comprehensive view of the previous research on the topic. One research study should be presented in one paragraph and it should mention the name of the researcher, year of study, topic and area of study, sample size, main objectives, and findings of the study. An example of a review is given in the box below.

Review of Literature

Singh (2008) conducted a study on labour out-migration from the Indo-Gangetic plains of India. The study provides sufficient evidence of the effect of male out-migration on the rural economy of the Indo-Gangetic plains of India. Male out-migration has resulted in gender role reversal in terms of decision making on important household and farm issues. Besides, the women of the migrant households had to take up many male specific activities, like land preparation, seed selection, broadcasting, irrigation, and herbicide application. The study also proved that the crop returns of non-migrant households were significantly higher than that of migrant households in case of both rice and wheat cultivation. The technical, allocative and economic efficiencies of non-migrant households was much higher than the migrant households in both rice and wheat cultivation.

4.5.3 Chapter 3: Methodology

The methodology adopted in conducting the study must be fully explained. The scientific reader would like to know about the basic design of the study, the methods of data collection, information regarding the sample used in the study, the statistical analysis adopted and the factors limiting the study. The methodology section should include a description of

- the study type
- major study themes or variables (a detailed list of variables on which data was collected may be annexed)
- the study/ target population(s), sampling method(s) and the size of the sample(s)

- data collection techniques used for the different study populations
- duration of data collection
- how the data was collected and by whom
- procedures used for data analysis, including statistical tests (if applicable)
- any constraints and its management
- limitations of the study.

If you have deviated from the original study design presented in your research proposal, you should explain to what extent you did so, and why. The consequences of this deviation for meeting certain objectives of your study should be indicated. If the quality of some of the data is weak, resulting in possible biases, this should be described as well under the heading 'limitations of the study'. An example of methodology is given in the box below.

Methodology

3.1 Data Collection / Sample

A micro level study based on primary cross section data was designed to attain the objectives of this project. The survey was conducted in three states; Bihar, Uttar Pradesh and Punjab. A systematic interview schedule was used to collect information on various aspects of labour migration and its impact on rural economy of Indo-Gangetic Plains of India. The data was collected for 200 families with migration and 200 families without migrating members.

3.2 Analytical tools

Various statistical tools were used in the analysis of data. Those are mean, standard deviation, correlation, t-test, and regression.

4.5.4 Chapter 4: Research Findings

A detailed presentation of the findings of the study with supporting data in the form of tables and charts, together with a validation of the results is the next step in writing the main text of the report. The result section of the study should contain the statistical summaries and reductions of data, rather than raw data. All the results should be presented in a logical sequence and split into readily identifiable sections.

The systematic presentation of your findings in relation to the research objectives is the crucial part of your report.

The list of data by objectives will help you to decide how to organise the presentation of data. The decision concerning where to put what can best be made after all data have been fully processed and analysed, and before the writing starts.

When all data have been analysed, a detailed **outline** has to be made for the presentation of the findings. This will help the decision-making on how to organise the data, and is an absolute precondition for optimal division of tasks among group members in the writing process.

At this stage you might as well prepare an outline for the whole report, taking the main components of a research report as a point of departure.

An **outline** should contain

- the headings of the main sections of the report
- the headings of subsections
- the points to be made in each section
- the list of tables, figures and/or quotes to illustrate each section.

The outline for the chapter on findings will predictably be the most elaborate.

The first section under findings is usually a description of the study/ target population. When different study populations have been studied, you should provide a short description of each group before you present the data pertaining to these informants.

Then, depending on the study design, you may provide more information on the problem you studied (size, distribution, characteristics). Thereafter, in an analytic study, the degree to which different independent variables influence the problem will be discussed.

For better understanding, an example of how the research findings are tabulated and presented in the form of findings is given in the following table. An analysis of table 4.1 is given in the box below.

Table 4.1: Social Characteristics of Migrants

(Percentage)

Particulars	Bihar	UP	Overall
Number	245	308	553
(i) Age Profile			
Up to 30 Years	69.80	56.49	62.39
31 to 45 Years	26.53	35.39	31.46
Above 46 Years	3.68	8.11	6.15
(ii) Literacy Status			
Illiterate	33.88	19.16	25.50
Primary	50.20	29.87	38.00
Matriculation and above	15.92	50.97	36.48
(iii) Social Status			
Upper Caste	22.86	9.42	15.37
SC/ST/BC	77.15	90.58	84.63 \

Analysis of Table 4.1

The socio-economic characteristics of the migrants are depicted in Table 4.1. The table clearly shows that in UP there were 308 migrants from 200 households while Bihar had only 245 migrants from 200 households. On an average, 62 percent of the migrants were below 30 years of age with a higher percentage of younger migrants from Bihar than from UP. Most of the migrants from both UP and Bihar were literate, and only 25 per cent of the total migrants from both UP and Bihar were illiterate. Most of the migrants belonged to a schedule caste or backward class, the percentage being higher in UP

(91%) compared with Bihar (77%). **Tables** and **Figures** in the text should be numbered and have clear titles. It is advisable to first use the number of the section to which the table belongs. In the final draft you may decide to number tables and figures in sequence. It is appreciated in case some pictures from the field are also appropriately presented to give visual presentation of the field information.

Include only those tables and figures that present main findings and need more elaborate discussion in the text. Others may be put in annexes, or, if they don't reveal interesting points, be omitted.

It is advisable to involve a statistician/data analyst from the very beginning in each process of the research so that he/she may provide meaningful tables and help remove irrelevant findings.

Note: It is unnecessary to describe in detail a table that you include in the report. Only present the main conclusions.

The first draft of your findings is never final. Therefore, you might concentrate primarily on content rather than on style. Nevertheless, it is advisable to structure the text from the beginning in paragraphs and to attempt to phrase each sentence clearly and precisely.

4.5.5 Chapter 5: Discussion

The findings can now be discussed by objective or by cluster of related variables or themes, which should lead to conclusions and possible recommendations. The discussion may also include findings from other related studies that support or contradict your own. For easy understanding, the discussion of the table given in findings is given in the box below.

The socio-economic characteristics of the migrants are depicted in Table 4.1. The number of migrants gives an idea about frequency of migrants in a household. In UP, the percentage of households having more than one migrant was relatively higher when compared with that of Bihar. Most of the migrants in both the states were up to 30 years of age. This clearly indicates that young men in their productive age were more involved in migration. Similar results were reported by Sidhu et.al. (1997) and Kumar et.al. (1998) in their studies, that most of the migrants of both the states were literate and belong to the backward sections of the society. The underlying fact is that the backward classes belonging to the lower social hierarchy were more capable of doing menial jobs and tasks, which required lot of energy.

4.5.6 Chapter 6: Conclusions and Recommendations

The conclusions and recommendations should follow logically from the discussion of the findings. As the discussion will follow the sequence in which the findings have been presented (which in turn depends on your objectives) the conclusions should logically follow the same order. Sometimes, it is advisable to present conclusion and recommendations in specific sections related to issues of importance/under investigation/objectives of the study for better clarity to different stake holders. The conclusions should be given in bullets so that it can easily catch the attention of the reader.

Remember that action-oriented groups are most interested in this section. The conclusions should be followed by suggestions or recommendations. While making recommendations, use not only the findings of your study, but also supportive information from other sources. The recommendations should be generated from the findings and conclusions. It should not be generalized; rather it should be specific to particular stake holders in pure, actionable term which is feasible in relation to social context, policy and constitution of country, political acceptability, budget, time, etc. One should **not give general recommendations** such as, “Government should provide free treatment to everyone for all health problems”.

If your recommendations are short (roughly one page), you might include them all in your summary and omit them as a separate section in Chapter 6 in order to avoid repetition.

4.5.7 Chapter 7: References

This is the list of books/articles in some way pertinent to the research which was followed while conducting research. It should contain all those works which the researcher has consulted. The references in your text can be numbered in the sequence in which they appear in the report and then listed in this order in the list of references (Vancouver system). Another possibility is the Harvard system of listing in brackets the author’s name(s) in the text, followed by the date of the publication and page number, for example: (Sharma et al., 2000: 84). The most commonly used style for citation in the text is the American Psychological Association (APA) style.

- In this, the work by a single author is cited in text as follows: Smith (1978).
- To cite work by more than two authors cite all authors the first time the reference occurs; in subsequent citations include only the surname of the first author followed by “et al.” Eg. Williams et al. (1985)

In the list of references, the publications are then arranged in alphabetical order by the principal author’s last name. For the APA style of reference,

- The entries are arranged in alphabetical order by last name of the first author.
- General format: Author last name, first initial (year). Title. Journal, issue, pages.
- Use a hanging indent for 2nd line of reference
- When your reference list contains multiple works by the same author:
- Single author entries precede multiple-author entries beginning with the same last name
- References with the same first author and different second or third authors are arranged alphabetically by the last name of the second author, etc.
- References with the same authors in the same order are arranged by year of publication, earliest first

Examples of references

Journal Article, One Author

Baumrind, D. (1967). Child care practices anteceding three patterns of preschool behavior. *Genetic Psychology Monographs*, 75(1), 43-88.

Journal Article, Multiple Authors

Ahadi, S. A., Rothbart, M. K., & Ye, R. (1993). Children's temperament in the US and China: Similarities and differences. *European Journal of Personality*, 7(5), 359-377.

Entire Book

Galvan, J. L. (2004). *Writing Literature Reviews: A guide for students of the social and behavioral sciences* (2 nd Ed). Glendale, CA: Pryczak Publishing

Article or Chapter in Edited Book

LeVine, R. A. (1988). Human parental care: Universal goals, cultural strategies, individual behavior. In R. A. LeVine & P. M. Miller & M. M. West (Eds.), *Parental behavior in diverse societies* (pp. 3-12). San Francisco: Jossey-Bass.

Website

Try to use the format above with the information you have available on the website. If you do not have an author, use website name at beginning. For example:

Hamilton College Writing Center (2004). <http://onthehill.hamilton.edu/academics/resource/wc/index.html>.

4.5.8 Annexure

The annexes should contain any additional information needed to enable professionals to follow your research procedures and data analysis.

Information that would be useful to special categories of readers but is not of interest to the average reader can be included in annexes as well.

Examples of information that can be presented in annexes are

- tables, figures (graphs) and pictures referred to in the text but not included in order to keep the report short
- lists of hospitals, districts, villages, etc., that participated in the study
- questionnaires or check lists used for data collection
- A list of research team members.

Note: Never start writing without an outline. Make sure that all sections carry the headings and numbers consistent with the outline before they are word-processed. Have the outline visible on the wall so that everyone will be aware immediately of any additions or changes, and of progress made.

4.6 STYLE AND LAYOUT OF THE REPORT

The style of writing and layout of writing are two important components of report writing. However, revising and finalizing the text may also be considered as another important aspect in same line.

4.6.1 Style of Writing

Remember that your reader

- is short of time

- has many other urgent matters demanding his or her interest and attention
- is probably not knowledgeable concerning 'research jargon'.

Therefore, the rules are

- simplify- Keep to the essentials
- justify- Make no statement that is not based on facts and data
- do not quote the name of anyone who has provided the information
- in case of sensitive findings, one should think not to clearly mention name of village/ location, etc.
- quantify when you have the data to do so; avoid 'large', 'small'- instead, say '50%', 'one in three'
- round off to the nearest whole number. The percentage 45.8 in table may be presented in the text as about 46% and 45.3% may be presented as approximately 45%
- be precise and specific in your phrasing of findings
- inform, not impress - avoid exaggeration
- use short sentences
- use adverbs and adjectives infrequently
- be consistent in the use of past and present tenses
- avoid the passive voice, if possible, as it creates vagueness (e.g., 'patients were interviewed' leaves uncertainty as to who interviewed them) and repeated use makes dull reading
- aim to be logical and systematic in your presentation.

4.6.2 Layout of the Report

A good *physical layout* is important, as it will help your report

- make a good initial impression
- encourage the readers
- give them an idea of how the material has been organised so the reader can make a quick determination of what he will read first.

Particular *attention* should be paid to make sure there is

- an attractive layout for the title page and a clear table of contents
- consistency in margins and spacing
- consistency in headings and subheadings, e.g.: **Font size 16 or 18 bold**, for headings of chapters; **size 14 bold** for headings of major sections; **size 12 bold**, for headings of sub-sections, etc.
- good quality printing and photocopying
- correct drafts carefully with spell check as well as critical reading for clarity by other team-members, your facilitator and, if possible, outsiders
- numbering of figures and tables, provision of clear titles for tables, and clear headings for columns and rows, etc.
- accuracy and consistency in quotations and references.

4.6.3 Revising and Finalising the Text

Prepare a double-spaced first draft of your report with wide margins so that you can easily make comments and corrections in the text. Have several copies made of the first draft, so you will have one or more copies to work on, and one copy on which to insert the final changes for revision. When a first draft of the findings, discussion, and conclusions has been completed, all working group members and facilitators should read it critically and make comments.

The following questions should be kept in mind when reading the draft.

- Have all important findings been included?
- Do the conclusions follow logically from the findings? If some of the findings contradict each other, has this been discussed and explained, if possible? Have weaknesses in the methodology, if any, been revealed?
- Are there any overlaps in the draft that have to be removed?
- Is it possible to condense the content? In general a text improves by shortening. Some parts less relevant for action may be included in annexes. Check if descriptive paragraphs may be shortened and *introduced or finished by a concluding sentence*.
- Do data in the text agree with data in the tables? Are all tables consistent (with the same number of informants per variable), are they numbered in sequence, and do they have clear titles and headings?
- Is the sequence of paragraphs and subsections logical and coherent? Is there a smooth connection between successive paragraphs and sections? Is the phrasing of findings and conclusions precise and clear?

The original authors of each section may prepare a second draft, taking into consideration all comments that have been made. However, you might consider the appointment of two editors amongst yourselves, to draft the complete version. The help from proof readers may also be taken to remove minor mistakes from the draft.

It is advisable to have one of the other groups and facilitators read the second draft and judge it on the points mentioned in the previous section. Then a final version of the report should be prepared. This time you should give extra care to the presentation and layout: structure, style and consistency of spelling (use spell check).

Use verb tenses consistently. Descriptions of the field situation may be stated in the past tense (e.g., 'Five households owned less than one acre of land.'). Conclusions drawn from the data are usually in the present tense (e.g., 'Food taboos hardly have any impact on the nutritional status of young children.')

Note: For a final check on readability you might skim through the pages and read the first sentences of each paragraph. If this gives you a clear impression of the organisation and results of your study, you may conclude that you did the best you could.

4.7 COMMON WEAKNESSES IN REPORT WRITING AND FINALIZING THE TEXT

It is important to know the general mistakes committed in report writing and also the points to consider while finalising the text.

(i) **Endless Description** without interpretation is a pitfall. Tables need conclusions, not detailed presentation of all numbers or percentages in the cells which readers can see for themselves. The chapter discussion, in particular, needs comparison of data, highlighting of unexpected results, your own or others' opinions on problems discovered, weighing of pros and cons of possible solutions. Yet, too often the discussion is merely a dry summary of findings.

(ii) **Neglect of Qualitative Data** is also quite common. Quotes of informants as illustration of your findings and conclusions make your report lively. They also have scientific value in allowing the reader to draw his/her own conclusions from the data you present. Presentation of important photographs also makes report attractive and explains facts better.

(iii) Sometimes qualitative data (e.g., open opinion questions) are just coded and counted like quantitative data, **without interpretation**, whereas they may provide interesting illustrations of reasons for the behaviour of informants or of their attitudes. This is serious maltreatment of data that needs correction.

In these sections you have read about the main text and end matter of the research report. You have also read about the style and layout of the research report. The general mistakes committed while writing a research report and the method of finalizing the text have also been given. Now, answer the questions that follow in *Check Your Progress-2*.

Check Your Progress - 2

Note: (a) Write your answer in about 50 words.

(b) Check your answer with possible answers given at the end of the unit.

1. What is the order of chaptering in a research report?

.....

.....

.....

2. What information do annexes and appendices contain?

.....

.....

.....

4.8 LET US SUM UP

The last part of any research is writing the research report. The report writing is an art as well as science. You have to identify who will be reading your report and the report should be prepared accordingly. A summary of report in the beginning is important. The report layout plan should be comprehensive and all aspects of report including realistic recommendations and future directions of research should be described.

4.9 REFERENCES AND SELECTED READINGS

Ackoff, R. L. (1961), *The Design of Social Science Research*, Chicago: University of Chicago Press.

- Bailey, K. D. (1978), *Methods of Social Research*, New York.
- Berelson, B. (1952), *Content Analysis in Communication Research*, Free Press, New York.
- Berenson, B. and C. Raymond (1971), *Research and Report Writing for Business and Economics*, Random House, New York.
- Kothari, C.R.(1978), *Quantitative Techniques*, New Delhi, Vikas Publishing House Pvt. Ltd.
- Gatner, E.S.M and C. Francesco (1956), *Research and Report Writing*, Barnes & Noble Inc., New York.
- Gaum, C.G., H.F. Graves and L. Hoffman (1950), *Report Writing*, 3rd ed., Prentice Hall, New York.
- Gopal, M.H. (1965), *Research Reporting in Social Sciences*, Karnatak University, Dharwad.

4.10 CHECK YOUR PROGRESS – POSSIBLE ANSWERS

Check Your Progress 1

1. The cover page should contain the title, the names of the authors with their titles and positions, the institution that is publishing the report with its logo, and the month and year of publication. The title could consist of a challenging statement or question, followed by an informative subtitle covering the content of the study and indicating the area where the study was implemented.
2. It is good practice to thank those who supported you technically or financially in the design and implementation of your study. Also your employer who has allowed you to invest time in the study and the respondents may be acknowledged. You should not forget to acknowledge the contribution of computer professionals, library staff, local officials, and community at large that provided the information. Acknowledgements are usually placed right after the title page or at the end of the report, before the references.

Check Your Progress 2

1. In any research report, the general trend of chaptering is as follows:
Chapter 1 Introduction
Chapter 2 Objectives
Chapter 3 Methodology
Chapter 4 Research Findings
Chapter 5 Discussion
Chapter 6 Conclusions and Recommendations
2. The annexes should contain any additional information needed to enable professionals to follow your research procedures and data analysis. Information that would be useful to special categories of readers but is not of interest to the average reader can be included in annexes as well.

MDV-106: RESEARCH METHODOLOGY IN DEVELOPMENT STUDIES

(6 CREDITS)

BLOCK	UNIT NOS	UNIT TITLES
1		Fundamentals of Social Science Research
	1	Social Science Research: An Overview
	2	Components of Social Science Research
	3	Research Designs
	4	Research Project Formulation
2		Development Research
	1	Basics of Development Research
	2	Methods of Development Research
	3	Development Research Applications
3		Measurement and Sampling
	1	Measurement
	2	Scales and Tests
	3	Reliability and Validity
	4	Sampling
4		Data Collection
	1	Quantitative Data Collection Methods and Devises
	2	Qualitative Data Collection Methods and Devises
	3	Data Sources
5		Data Analysis
	1	Overview of Statistical Tools
	2	Use of Computer in Data Analysis
	3	Data Processing and Analysis
	4	Report Writing