

CENTRE FOR DISTANCE & ONLINE EDUCATION
UNIVERSITY OF JAMMU
JAMMU



SELF LEARNING MATERIAL
FOR
M.A. SOCIOLOGY
SEMESTER - III
TITLE : SOCIAL STATISTICS AND COMPUTER
ANALYSIS

Course No. : SOC-C-302

Credits : 6

Unit : I - IV

Subject : Sociology

Lesson No. 1 - 20

Course Coordinator
Prof. Vishav Raksha

Head, Department of Sociology
University of Jammu.

Teacher In charge
Dr. Neha Vij

P.G. Sociology, CDOE
University of Jammu.

<http://www.distanceeducationju.in>

Printed and Published on behalf of the Centre for Distance & Online Education, University of Jammu, Jammu by the Director, CDOE, University of Jammu, Jammu.

**COURSE TITLE : SOCIAL STATISTICS AND COMPUTER
ANALYSIS**

Course Contributors :-

Prof. J.P. Singh Joorel

Professor Department of Statistics
University of Jammu.

Mr. Raman Gupta

Director, Computer Centre
University of Jammu.

Dr. Hema Gandotra

Associate Professor
Department of Sociology
University of Jammu.

Editing and Proof Reading By:

Dr. Neha Vij

Assistant Professor, CDOE
University of Jammu

© Centre for Distance & Online Education, University of Jammu, Jammu, 2026

- 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 CDOE, University of Jammu.
- The script writer shall be responsible for the lesson/script submitted to the CDOE and any plagiarism shall be his/her entire responsibility.

**Syllabus of Sociology M.A. 3rd Semester for the Examination to be held in
the year December 2026, 2027 and 2028 (NON-CBCS for CDOE)**

Course No: SOC-C-302

Credits: 6

Duration of examination: 3 hours.

Title: Social Statistics and Computer Analysis

Maximum Marks: 100

a) Semester Examination (External): 70

b) Session Assessment (Internal): 30

Objective: To train the students of Sociology in basic statistical methods which are applicable in Sociological problems and data analysis. The course also intends to acquaint the students with the different computer applications and their use in the Research.

UNIT-I Quantitative Methods and Survey Research

Measures of Central tendency: Mean, Median and Mode; Geometric Mean and Harmonic Mean; Measurement and Scaling; Reliability and Validity in Quantitative Research.

UNIT-II Statistics in Social Research

Measures of Dispersion: Standard Deviation, Quartile Deviation, and Mean Deviation; Correlation Analysis, Regression Analysis & their Relationship; Association of Attributes.

UNIT-III Sampling and Statistical Tests

Meaning and Methods of Sampling; Procedure of testing a hypothesis; Tests of Significance - Student's t-test, f-test and Chi-square test.

UNIT-IV Computer Applications

Statistical data and use of Computers; Introduction to Windows Operating System. M.S.- Word, M.S. Excel, M.S. Power Point, Introduction of MS Office.

NOTE FOR PAPER SETTING:

A. Session Assessment Internal (30 marks)

There will be three Internal Assessment Assignments carrying the total weightage of 30 marks.

IAA1. Long Answer Type Questions ($10 \times 1 = 10$ Marks)

Long answer type question (of maximum 1200 words) of 10 marks (one to be attempted, out of two, each from unit I and II).

IAA2. Short Answer Type Questions ($5 \times 2 = 10$ Marks)

Two short answer type questions (of about 600 words), each of 5 marks. These shall be taken from units III and IV.

IAA3. Very Short Answer Type Questions ($2.5 \times 4 = 10$ Marks)

Four very short answer type questions (of about 250 words), each of 2.5 marks. These shall be taken from all the four units.

B. Semester Examination External (70 marks)

The question paper will consist of two sections A and B.

Section A will consist of eight long answer-type questions, two questions from each unit. The candidate will be required to answer four questions, selecting one from each unit. Each question will carry 13 marks ($13 \times 4 = 52$ marks).

Section B will consist of eight short answer-type questions, two questions from each unit. The candidate will be required to answer four questions, selecting one from each unit. Each question will carry 4.5 marks ($4.5 \times 4 = 18$ marks).

Prescribed Readings

1. Agarwal, B.L. 2000. *Basic Statistics*, New Delhi: New Age International (P) Limited Publisher.
2. Argyrous, George. 1997. *Statistics for Social Research*, New York: Mc Millan Press Ltd.
3. Druckman, Daniel. 2005. *Doing Research: Methods of Inquiry for Conflict Analysis*, Sage Publication, New Delhi.
4. Goods, W.J. & Hatt, P.K. 1981. *Methods in Social Research*, New York: Mc Graw Hill.
5. Gupta, S.C. 1981. *Fundamentals of Statistics*, Bombay: Himalayan Publishing House.
6. Gupta, S.P. 2004. *Statistical Methods*, New Delhi: Sultan Chand and Sons.
7. Healey, Joseph H. 1990. *Statistics: A Tool for Social Research*, California: Wadsworth Publication Co.
8. Majumdar, P.K. 2005. *Research Methods in Social Sciences*, Viva Books, New Delhi.
9. Majumdar, P.K. 2002. *Statistics: A Tool for Social Sciences*, New Delhi: Rawat Publications.
10. Nachmias, C.F. and D. Nachmias. 1996. *Research Methods in Social Sciences*, Arnold.
11. Neuman, W. Lawrence. 1997. *Social Research Methods*, Allyn & Bacon, London.
12. Ram, B. 2000. *Computer Fundamentals*, New Delhi: New Age International (P) Limited Publishers.
13. Sarantakos, S. 2005. *Social Research*, Palgrave MacMillan, New York.
13. Scale, Clive (ed). 2004. *Social Research Methods - A Reader*, Routledge, London.
14. Smith, Gray. 1998. *Introduction to Statistical Reasoning*, New York: Mc Graw-Hill.
15. Srivastava. 2004. *Methodology and Field Work*, OUP, New Delhi.
16. Vaus, D.A. DE. 2002. *Surveys in Social Research*, Rawat Jaipur.
17. Xavier, C. 2000. *Introduction to Computers and Basic Programming*, New Delhi: New Age International (P) Limited, Publishers.
18. Yadava, Surender & Yadava, K.N. S. 1995. *Statistical Analysis for Social Sciences*.

TABLE OF CONTENTS

UNIT-I		
QUANTITATIVE METHODS AND SURVEY RESEARCH		
L. NO.	LESSON NAME	PAGE NO.
1.	Measures of Central Tendency: Mean, Geometric Mean and Harmonic Mean	5
2.	Median and Mode	27
3.	Measurement and Scaling	45
4.	Reliability and Validity in Quantitative Research	62
UNIT-II		
STATISTICS IN SOCIAL RESEARCH		
5.	Measures of Dispersion: Range, Interquartile Range, Quartile Deviation, Mean Deviation	70
6.	Standard Deviation and Co-efficient of Variation	85
7.	Correlation Analysis	102
8.	Regression Analysis and their Relationship	123
9.	Association of Attributes	148
UNIT-III		
SAMPLING AND STATISTICAL TESTS		
10.	Meaning and Methods of Sampling	176
11.	Procedure of Testing a hypothesis	197
12.	Tests of significance – Student's T-test	203
13.	Chi Square Test	222
14.	F-Test	239
UNIT-IV		
COMPUTER APPLICATIONS		
15.	Introduction to Computer Operating System	260
16.	Word Processing (MSWord)	287
17.	Introduction to Microsoft Excel XP (MS Excel)	308
18.	MS PowerPoint	323
19.	Introduction to MS Office	381
20.	Statistical data and Use of Computers	405

MEASURES OF CENTRAL TENDENCY:
MEAN, GEOMETRIC MEAN AND HARMONIC MEAN

STRUCTURE

- 1.0 Objectives
- 1.1 Introduction
- 1.2 Concept of Central Tendency
- 1.3 Different Measures of Central Tendency
- 1.4 Arithmetic Mean
 - 1.4.1 Computation of Arithmetic Mean
 - 1.4.2 Properties of Arithmetic Mean
- 1.5 Geometric Mean
 - 1.5.1 Computation of Geometric Mean
- 1.6 Harmonic Mean
- 1.7 Let us sum up
- 1.8 Glossary
- 1.9 Self-Assessment Questions
- 1.10 Lesson End Exercise
- 1.11 Suggested Readings
- 1.12 Answers to check your progress

1.0 OBJECTIVES

After going through this lesson, you should be able to:

- Understand the concept and significance of measures of central tendency.
- Learn about different types of measures of central tendency.
- Compute various measures of central tendency, such as Arithmetic mean, Geometric mean and Harmonic mean.
- explain the properties and merits of central tendency, and
- State the limitations of the central tendency.

1.1 INTRODUCTION

With this lesson, we begin our formal discussion of the statistical methods for summarizing and describing numerical methods for summarizing and describing numerical data. The first step in that direction is to find one representative value which can be used to locate and summaries the entire set of varying values (data).

This one value can be used to make many decisions concerning the entire set of values. We can define measures of central tendency or location to find some central value around which the data and to cluster.

In this lesson you will study the concept of measures of central tendency and its types, Computation of Arithmetic mean Geometric mean and Harmonic mean. You will further learn in detail the properties, and limitations of these measures of central tendency.

1.2 CONCEPT OF CENTRAL TENDENCY

In the general pattern of frequency distribution in the data, we may identify a single value around which many other items or values of the data congregate. This is a value which is somewhere in the central part of the range of all values. When this typical item/value of the data aims towards the central part of the data, it is known as Central Tendency. As it indicates the location of the clustering of items, is also called measures of location. Thus, the central tendency (value) of the numerical data gives the central idea of the entire data. Such a value is called central value or an average or the expected value of the variable. The word average is very commonly used in day-to-day conversation. Measures of central tendency enable us to compare two or more sets of data to facilitate comparison. For example, the average sales figures of a particular item of June may be compared with the sales figures of previous months.

It should be clear that the concept of a measure of central tendency is concerned only with quantitative

variables (data) and is undefined for qualitative variables (data) as these are un-measurable on a scale.

As the average is a single representative value of the mass of complex data, it must have the following characteristics:

- (i) It should be rigidly defined.
- (ii) It should be easy to understand and simple to compute.
- (iii) It should be based on all the observations of data.
- (iv) It should not be affected by extreme values of the observations. As single extreme value i.e., a maximum or a minimum value can unduly affect the average. A too small item can reduce the value of an average, and a too large item can also inflate its value to a large extent.

It should be capable of further algebraic treatment. That is an average should be amenable to further algebraic treatment. That should add to its utility. For example, if we are given the averages of three data sets of same type, it should be possible to obtain the combined average of all those three data sets.

CHECK YOUR PROGRESS-I

Tick the correct one

1. The simplest of all averages is: -
 - (a) Arithmetic mean b) Geometric Mean
 - b) Harmonic mean d) None of the above
2. In case of continuous Series, arithmetic mean is calculated conveniently by.....
 - (a) Direct method b) Step Deviation Method
 - c) Short cut method d) All of the above

1.3 DIFFERENT TYPES OF MEASURES OF CENTRAL TENDENCY

In statistics, various types of measures of central tendencies or averages are studied. For simplification and clarity, the different measures of central tendencies are classified as under:

(a) Mathematical Averages:

1. Arithmetic Mean or Mean

2. Geometric Mean
3. Harmonic Mean

(b) Positional Averages:

1. Median
2. Mode
3. Partition values.

At presents however, we are confined to the first set of measures which A.M., G.M. and H.M.

1.4 ARITHMETIC MEAN

The arithmetic mean (A.M.) or mean or average is the most popular and widely used measure of central tendency. It is obtained by dividing the sum of the values of all the observations in the given data by the number of observations in that data.

1.4.1 Computation of Arithmetic Mean

In classification and tabulation of data, we observed that the values of the variable or observations can be put in the form of any of the following statistical series; namely: –

- (i) Individual series or ungrouped data.
- (ii) Discrete Frequency Distribution.
- (iii) Continuous Frequency Distribution.

Arithmetic mean for the above series is calculated as under:

(I) Individual Series or Ungrouped Data: Following two methods are used for calculating arithmetic mean of an individual series or ungrouped data or data without any frequencies:

Direct Method: Computation of A.M. is very simple when the data is ungrouped i.e., frequency distribution is not given or done. Just add all the values of the observation and divide it by the number of observations. Normally, the A.M. is denoted by which $\bar{X}$ which is read as ‘X bar’. If the values of N observations on a variable X are $X_1, X_2, X_3, \dots, X_N$; then the A.M., denoted by $\bar{X}$, is defined by

$$\bar{X} = \frac{X_1 + X_2 + X_3 \dots X_N}{N} = \frac{\sum X}{N}$$

Where $\sum X$ means sum of all the observations of variable X and N be the number of observations.

NOTE: Σ (read it as sigma is the Greek symbol denoting the summation over all values.

Example 1: The following table gives the daily income of 8 employees in an office. Find out the average income of the employees.

Employees:	1	2	3	4	5	6	7	8
Income (Rs.):	150	350	200	180	250	100	350	200

Solution: Average income can be computed as follows:

$$\begin{aligned}\bar{X} &= \frac{\sum X}{N} \\ &= \frac{150 + 350 + 200 + 180 + 250 + 100 + 350 + 200}{8} \\ &= \frac{1780}{8} = 222.50\end{aligned}$$

Thus, average income of the employees is Rs. 222.50 per day.

(b) Short-cut Method: When the values of the observation in the given data are too large or they are in fraction; then the computation of mean through direct method becomes difficult.

This difficulty can be solved by using the short-cut method. Under this method, the following steps are to be followed:

Step-I: Assume any arbitrary mean (A) to find out the deviations of items from assumed mean. The assumed mean is usually chosen to be a round number in the middle of the range of the given observations, so that deviations can be easily obtained by subtraction.

Step-II: Compute the deviation (D) of each observation from the assumed mean i.e. $D = X - A$.

Step-III: Obtain the sum of all deviations i.e., $\sum D$.

Step-IV: Compute the A.M. (average or mean) by using the following formula.

$$\bar{X} = A + \frac{\sum D}{N}$$

Example 2: Obtain average income of data given in example 1 with short-cut method.

Solution: Suppose assumed mean is 200, and then make the following table:

Employee No.	Daily income (X)	D = X-A
1	150	-50
2	350	150
3	200	0
4	180	-20
5	250	50
6	100	-100
7	350	150
8	200	0
N = 8		$\Sigma D = 180$

Thus, the average income is

$$\begin{aligned}\bar{X} &= A + \frac{\sum D}{N} = 200 + \frac{180}{8} \\ &= 200 + 22.50 \\ &= \text{Rs. } 222.50\end{aligned}$$

NOTE : It may be observed here that answer obtained by the direct method and the short-cut method is the same.

II. **Discrete Frequency Distribution or Discrete Series or Grouped Data :** The following methods are used for computing arithmetic mean in a discrete series or grouped data :

- (a) Direct Method, (b) Short Cut Method
- (c) Step-Deviation Method.

(a) **Direct Method :** In grouped data or discrete series the average or mean can be obtained by using the following formula :

$$\bar{X} = \frac{f_1 X_1 + f_2 X_2 + \dots + f_n X_n}{f_1 + f_2 + \dots + f_n} = \frac{\sum fX}{\sum f}$$

$$= \frac{\sum fX}{N}$$

Where $X_1, X_2, X_3, \dots, X_n$ are the observations of the variable and $f_1, f_2, f_3, \dots, f_n$ are the respective frequencies.

The procedure will be more clear from the following examples :

Example 3. Following table gives the wages paid to 125. Workers in a factory per day. Calculate the average of wages :

Wages (Rs.)	200	210	220	230	240	250	260
No. of Workers	5	15	32	42	15	12	4

Solution : Calculation of Arithmetic Mean

Wages (Rs.)	Number of Workers	fX
X	f	
200	5	1000
210	15	3150
220	32	7040
230	42	9660
240	15	3600
250	12	3000
260	4	1040
	$N = \sum f = 125$	$\sum fX = 28490$

The average wage is

$$\bar{X} = \frac{\sum fX}{\sum f} = \frac{28490}{125} = 227.92$$

Thus, average wage is Rs. 227.92 per day.

(b) **Short-cut method :** According to this method, the formula for calculating arithmetic mean is

$$\bar{X} = A + \frac{\sum fD}{N}$$

Where

A = Assumed mean

D = X-A, i.e. deviations from assumed mean.

ΣfD = Sum of the products of frequencies and their respective deviations.

Example 4. Calculate the arithmetic mean for the data in example 3 by using short cut method.

Solution : In this case, assumed mean (A) is taken to be 230 as it is at the centre of the range of the variable X. The other needed calculations are given in the table.

Wages ($\bar{X}$)	No. of Workers f	Deviations $D=X-230$	fD
200	5	-30	-150
210	15	-20	-300
220	32	-10	-320
230	42	0	0
240	15	10	150
250	12	20	240
260	4	30	120
	$N = \Sigma f = 125$		$\Sigma fD = -260$

The arithmetic mean will be

$$\begin{aligned}\bar{X} &= A + \frac{\Sigma fD}{\Sigma f} = 230 + \frac{-260}{125} \\ &= 230 - 2.08 \\ &= 227.92\end{aligned}$$

e) **Step-Deviation Method :** In the short cut method, the deviations taken from an assumed mean generally have a common factor. In continuous series this common factor is nothing but the uniform class interval. The computational work in short-cut method can be further simplified if the deviations are divided by this common factors. The deviations divided by the common factor are called step-deviations. According to this method the arithmetic mean is calculated by the formula :

$$\bar{X} = A + h \frac{\sum fD}{\sum f}$$

Here C = the common factor in deviations $X-A$

$$D = \frac{X-A}{h}, \text{ the step deviations}$$

Example 5. Use step deviation method for calculating arithmetic mean for the data in example 3.

Solution : Computation of Arithmetic mean by step Deviation method

Wages X	No. of Workers f	Deviations X-230	$D = \frac{X-230}{10}$	fD
200	5	-30	-3	-15
210	15	-20	-2	-30
220	32	-10	-1	-32
230	42	0	0	0
240	45	10	1	15
250	12	20	2	24
260	4	30	3	12
	N = 125			$\sum fD = -26$

Thus, the arithmetic mean becomes

$$\bar{X} = A + h \frac{\sum fD}{N} \times C$$

$$= 230 + 10 \left(\frac{-26}{125} \right)$$

$$= 230 - 2.08$$

$$= 227.92$$

(III) **Continuous Series :** In continuous series, the procedure of computing arithmetic mean is the same as in the case of discrete series. The only difference is that in continuous series the frequencies within each class are assumed to be distributed uniformly over its range. With this assumption each class is then represented by its mid-point, denoted by m. using these mid-points (m) of the classes and the corresponding frequencies, arithmetic mean for the continuous series can be calculated by using

any of the methods used in a discrete series. The following example will clearly the procedure:

Example 6. Calculate the arithmetic mean from the following data by using Direct Method and Step Deviation Method :

Class : 20–25 25–30 30–35 35–40 40–45 45–50 50–55

Frequency : 8 10 12 20 11 4 5

Solution :

Computation of A. M.

Class	Frequency f	Mid-Points m	fm	$D = \frac{m - 37.5}{5}$	fD
20–25	8	$\frac{20 + 25}{2} = 22.5$	180.00	–3	–24
25–30	10	$\frac{25 + 30}{2} = 27.50$	275.00	–2	–20
30–35	12	$\frac{30 + 35}{2} = 32.50$	390.00	–1	–12
35–40	20	$\frac{35 + 40}{2} = 37.50$	750.00	0	0
40–45	11	$\frac{40 + 45}{2} = 42.5$	467.50	1	11
45–50	4	$\frac{45 + 50}{2} = 47.50$	190.00	2	8
50–55	5	$\frac{50 + 55}{2} = 52.50$	262.50	3	15
Total	N = 70		$\Sigma fm = 2515$		–21

Therefore, arithmetic mean is

By Direct Method :

$$\bar{X} = \frac{\sum fm}{\sum f} = \frac{2515}{70} = 35.93$$

By Step Deviation Method :

$$\begin{aligned}\bar{X} &= A + h \frac{\sum fD}{\sum f} = 37.50 + 5 \left(\frac{-22}{70} \right) \\ &= 37.50 - 1.57 = 35.93\end{aligned}$$

1.4.2 Properties of Arithmetic Mean : The following are the main properties of arithmetic mean :

- (i) The sum of the deviations of the observations from the arithmetic mean is always zero i.e., $\Sigma(X - \bar{X}) = 0$. This is explained in the following illustration.

X	$X - \bar{X}$
10	-25
20	-15
30	-5
40	5
50	15
60	25
$\Sigma X = 210$	$\Sigma(X - \bar{X}) = 0$

Here $\bar{X} = \frac{\Sigma X}{N} = \frac{210}{6} = 35$

- (ii) If the number of items and mean are known, the total of the items can be obtained as $\bar{X}$

$$= \frac{\Sigma X}{N}, \text{ thus } \Sigma X = N \bar{X}.$$

- (iii) The sum of squares of deviations from the arithmetic mean is minimum i.e., it is always less than the sum of squares of deviations of the items taken from any other value. In other words $\Sigma(X - \bar{X})^2$ is always minimum. We can verify this by taking the example

X	$X - \bar{X}$	$(X - \bar{X})^2$	$X - 10$	$X - 20$	$(X - 10)^2$	$(X - 20)^2$
5	-10	100	-5	-15	25	225
10	-5	25	0	-10	0	100
15	0	0	5	-5	25	25
20	5	25	10	0	100	0
25	10	100	15	5	225	25
$\Sigma X = 75$	$\Sigma(X - \bar{X}) = 0$	250			375	375

Here $\bar{X} = \frac{\sum X}{N} = \frac{75}{5} = 15$

$$\Sigma(X-\bar{X})^2 = 250, \Sigma(X-10)^2 = \Sigma(X-20)^2 = 375$$

$$\therefore \Sigma(X-\bar{X})^2 \text{ is minimum.}$$

- (iv) If we add or delete an observation which is equal to mean, the A.M. remains unaffected.
- (v) If each observation is increased or decreased by some constant C, the A. M. also increase or decrease by C. Similarly, when each observation is multiplied by a constant, say K the A.M. is also multiplied by the same quantity K.
- (vi) If the means and the number of observations of two or more related groups are known, we can obtain the combined mean of these groups as follows :

Combined Mean

$$\bar{X} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

Here $\bar{X}$ = Combined mean

N_1 = No. of observations in first group

N_2 = No. of observations in second group.

$\bar{X}_1$ = Mean of the first group

$\bar{X}_2$ = Mean of the second group

The formula may also be extended for K-groups as

$$\bar{X} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2 + \dots + N_k\bar{X}_k}{N_1 + N_2 + \dots + N_k}$$

Example : The A. M. of the production of a item. during the period January to August 4000 items per months and A.M. for the period September to December is 4300 items per month. Find the

average production for the whole year.

Solution : Here $N_1 = 8$ (Jan. to August)

$$N_2 = 4 \text{ (Sept. to Dec.)}$$

$$\bar{X}_1 = 4000$$

$$\bar{X}_2 = 4300$$

The average production for the whole year ($\bar{X}$) can be obtained by combined mean formula i.e.

$$\begin{aligned}\bar{X} &= \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2} \\ &= \frac{8 \times 4000 + 4 \times 4300}{8 + 4} \\ &= \frac{32000 + 17200}{12} \\ &= \frac{49200}{12} = 4100 \text{ items per month.}\end{aligned}$$

1.5 GEOMETRIC MEAN.

In the situations where we deal with quantities that change over a period of time, we may be interested to know the average rate of change. In such cases the simple arithmetic mean is not suitable and we have to resort to the geometric mean. The geometric mean (G. M.) of a set of N positive items is defined as the N^{th} root of their product.

1.5.1 Computation of Geometric Mean : Like arithmetic mean, computation procedure of geometric mean is different for grouped and ungrouped data. As defined, the geometric mean is given by

$$\text{G. M.} = (X_1 \cdot X_2 \cdot X_3 \cdot \dots \cdot X_N)^{1/N} \text{ for individual series}$$

$$\left(X_1^{f_1} \cdot X_2^{f_2} \cdot \dots \cdot X_n^{f_n} \right)^{\frac{1}{N}} \text{ for discrete series}$$

$$= \left(m_1^{f_1} m_2^{f_2} \dots m_n^{f_n} \right)^{\frac{1}{N}} \text{ for continuous series}$$

here $N = f_1 + f_2 + \dots + f_n$ (in case of discrete and continuous series) and m be the mid-points.

For example, the G.M. of three numbers 4, 8 and 16 would be

$$\begin{aligned} \text{G. M.} &= (4 \times 8 \times 16)^{1/3} = (512)^{1/3} \\ &= 8 \end{aligned}$$

If the number of items is four or more the task of multiplying the numbers and of finding the N^{th} root becomes difficult. Therefore, computations can be simplified by the use of logarithm. The G. M. is then calculated as follows :

$$\begin{aligned} \log \text{ G.M.} &= \frac{\log X_1 + \log X_2 + \dots + \log X_N}{N} \\ &= \frac{\sum \log X}{N} \end{aligned}$$

$$\therefore \text{ G.M.} = \text{Antilog} \left(\frac{\sum \log X}{N} \right)$$

Similarly in discrete and continuous series it becomes.

$$\text{G.M.} = \text{Antilog} \left(\frac{\sum f \log X}{N} \right) \text{ for discrete case}$$

$$= \text{Antilog} \left(\frac{\sum f \log m}{N} \right) \text{ for continuous case.}$$

Example 8 : The daily income of 10 persons in a locality are given below (in Rs.): 85, 70, 15, 75, 500, 8, 45, 250, 40, 36. obtain the geometric mean.

Solution : Computation of G.M.

X	log X
85	1.9294
70	1.8451
15	1.1761
75	1.8751
500	2.6990
8	0.9031
45	1.6532
250	2.3979
40	1.6021
36	1.5563
Total	$\Sigma \log X = 17.6373$

$$\therefore \text{G.M.} = \text{Antilog} \left[\frac{\sum \log X}{N} \right]$$

$$= \text{Antilog} \left[\frac{17.6373}{10} \right]$$

$$= \text{Antilog} [1.76373]$$

$$= 58.03$$

Thus the geometric mean is Rs. 58.03.

Example 9. Find the geometric mean in the following distribution.

Marks	:	0-10	10-20	20-30	30-40	40-50
No. of Students	:	5	7	15	25	8

Solution :

Computation of G.M.

Marks (Class)	No. of Students (<i>f</i>)	Mid-point <i>m</i>	$\log m$	$f \log m$
0–10	5	5	0.6990	3.4950
10–20	7	15	1.1761	8.2327
20–30	15	25	1.3979	20.9685
30–40	25	35	1.5441	38.6025
40–50	8	45	1.6532	13.2256
Total	N = 60			85.5243

Here $N = 60$, $\Sigma f \log m = 85.5243$.

Thus, G.M. is

$$\text{G.M.} = \text{Antilog} \left[\frac{\Sigma f \log m}{N} \right]$$

$$= \text{Antilog} \left[\frac{85.5243}{60} \right]$$

$$= \text{Antilog} [1.4087]$$

$$= 25.63 \text{ marks.}$$

1.6 HARMONIC MEAN

As you know, generally the data is in varied forms. The manner in which the data is given counts heavily for judging the appropriateness of the use of the measures of central tendency. For example, when the total distance is constant and the speed per unit time is given then harmonic mean would be more appropriate measure to find average speed. Further, suppose that production rate per unit of time is given and we are interested in knowing the average, then harmonic mean is preferable.

The harmonic mean (H.M.) of a set of observations is the reciprocal of the arithmetic mean of the reciprocals of the observations. Thus,

$$\begin{aligned}
\text{H.M.} &= \frac{N}{\frac{1}{X_1} + \frac{1}{X_2} + \dots + \frac{1}{X_N}} \\
&= \frac{N}{\sum \left(\frac{1}{X} \right)}; \text{ for individual series} \\
&= \frac{N}{\sum \left(\frac{f}{X} \right)}; \text{ for discrete series} \\
&= \frac{N}{\sum \left(\frac{f}{m} \right)}; \text{ for continuous series}
\end{aligned}$$

Here symbols have their usual meanings.

Example 10. Obtain harmonic mean of the following observations : 40, 45, 30, 35, 55, 65, 37, 42

Solution.

Computation of H. M.

X	1/X
40	0.0250
45	0.0222
30	0.0333
35	0.0286
55	0.0182
65	0.0154
37	0.0270
42	0.0238
	0.1935

$$\text{Here } N = 8, \sum \left(\frac{1}{X} \right) = 0.1935$$

The H.M. is

$$\text{H.M.} = \frac{N}{\sum \left(\frac{1}{X} \right)} = \frac{8}{0.1935} = 41.34$$

Example 11. The distribution of marks obtained by students in a class is as under

Marks :	20	21	22	23	24	25
No. of students :	4	12	15	20	11	8

Find the Harmonic mean.

Solution :- Computation of H.M.

X Marks	f No. of students	$\frac{1}{X}$	$f \left(\frac{1}{X} \right)$
20	4	0.0500	0.2000
21	12	0.0476	0.5712
22	15	0.0455	0.6825
23	20	0.0435	0.8700
24	11	0.0417	0.4587
25	8	0.0400	0.3200
	N = 70		$\sum \left(\frac{f}{X} \right) = 3.1024$

Here $N = 70, \sum \left(\frac{f}{X} \right) = 3.1024$, thus

$$\text{H. M.} = \frac{N}{\sum \left(\frac{f}{X} \right)} = \frac{70}{3.1024}$$

= 22.56 marks.

Example 12. A train goes at a speed of 20 miles per hour for the 16 miles, at a speed of 40 miles per hour for 20 miles. It covers the last 10 miles at a speed of 15 miles per hour. Find out average speed.

Solution : Computation of H.M.

Speed	Distance	W/X
M.P.H. X	W	
20	16	0.800
40	20	0.500
15	10	0.667
Total	46	1.967

Here $N = \sum W = 46$, $\sum \frac{W}{X} = 1.967$,

$$\text{H.M.} = \frac{\sum W}{\sum \left(\frac{W}{X} \right)} = \frac{46}{1.967} = 23.38$$

Average speed is 23.38 M.p.h.

1.7 LET US SUM UP

The main characteristics of the data are represented by a single figure known as ‘an average’ or a mean’. It is the point of location around which individual values cluster. An ideal average must satisfy certain properties such as ease of calculation, rigidity in its definition, should be based on all items, should remain unaffected by extreme values, and also should have sampling stability. An average gives a bird’s eye view of the entire data, facilitates comparison and become and useful in statistical inference. In the present lesson we have explained A.M., H.M., and G.M. in detail along with their computational procedures.

1.8 SUMMARY

The main characteristics of the data are represented by a single figure known as ‘an average’ or a mean’. It is the point of location around which individual values cluster. An ideal average must satisfy certain properties such as ease of calculation, rigidity in its definition, should be based on all

items, should remain unaffected by extreme values, and also should have sampling stability. An average gives a bird's eye view of the entire data, facilitates comparison and becomes useful in statistical inference. In the present lesson we have explained A.M., H.M., and G.M. in detail along with their computational procedures.

1.9 GLOSSARY

- **Central Tendency:** The central value in data around which all other values move.
- **Arithmetic Mean:** The Simple average value calculated in data set is known as arithmetic mean.
- **Geometric Mean:** Average change in quantities over a period of time is called geometric mean.
- **Harmonic Mean:** It is the reciprocal of arithmetic mean.
- **Common difference:** It is that value which is common L.C.M of all the frequencies in continuous series. It is usually indicated by small letter 'c'.

1.9 SELF-ASSESSMENT QUESTIONS

1. Explain the qualities of a good measure of central tendency.

2. What are the various measures of central tendency.

3. Define Arithmetic mean. Also mention its properties and limitations.

1.10 LESSON END EXERCISE

1. Define geometric mean with its merits and demerits.

- 2 Give a brief description of harmonic mean. State the purpose of studying this average.

- 3 The height (in cm) of 10 students is given as; 156, 154, 168, 172, 160, 168, 175, 170, 158, 162. Find the A.M., G. M. and H.M.

- 4 Find, A.M., G.M. and H.M. for the following frequency distribution:

Class	f
10–20	4
20–30	8
30–40	6
40–50	20
50–60	12
60–70	8
70–80	2

- 5 The average monthly salary of 20 male worker in a factory in Rs. 3200/- per month and that of 16 female is Rs. 2500/- per month. Find the average salary of both male and female.

1.11 SUGGESTED READINGS

1. Argyros, George. 1997. *Statistics for Social Research*. New York: Mc Millan Press Ltd.
2. Goods, W.J. & Hatt, P.K. 1981. *Methods in Social Research*. New York: Mc Graw Hill.
3. Gupta, S.C. 1981. *Fundamentals of Statistics*. Bombay: Himalayan Publishing House.
4. Gupta, S.P. 2004. *Statistical Methods*. New Delhi: Sultan Chand and Sons.

1.12 ANSWERS TO CHECK YOUR PROGRESS

Answers: 1. (a) 2. (a)

MEASURES OF CENTRAL TENDENCY: MEDIAN AND MODE**STRUCTURE**

- 2.0 Objectives
- 2.1 Introduction
- 2.2 Median
 - 2.2.1 Computation of Median
 - 2.2.1 Merits and Limitations of Median
- 2.3 Partition Values
- 2.4 Mode
 - 2.4.1 Computation of Mode
 - 2.4.2 Merits and Limitations of Mode
- 2.5 Let us sum up
- 2.6 Glossary
- 2.7 Self-Assessment Questions
- 2.8 Lesson End Exercise
- 2.9 Suggested Readings
- 2.10 Answers to check your progress

2.0 OBJECTIVES

Dear learner, after studying this lesson, you should be able to:

- understand the meaning of median
- understand the computational procedure of median for different sets of data,
- define quartiles and its computational procedure.
- understand the concept of mode and its computational procedure for different types of data, and
- state the uses and limitations of median and mode.

2.1 INTRODUCTION

As you have studied in previous lesson that measures of central tendency were classified into two categories. Mathematic averages (A.M, G.M., and H.M.) have already been explained in the previous lesson. As you know, these averages are affected by extreme values and also cannot be obtained in open-end class interval. Many times, we may like to find an average which is not affected by extreme values. Median and mode are such measures. There are some other measures called portion values, which are not averages, but similar to median in concept. In the present lesson you will learn the meaning, computational procedure, limitations of these measures of central tendency.

2.2 MEDIAN

In a given 'array' i.e., when the observations are arranged in an ascending or descending order. Any point which divides the array into two equal parts, so that exactly one half of the observations are below, and one-half are above that point, is called median. Thus, the median is that value of the variable which divides the group in two equal parts. The median is usually denoted by Me or Md.

2.2.1 Computation of Median

Median can be obtained for both ungroup and grouped data. But the methods are different. Now let us study the methods of computing median for different types of data.

I. Individual series or ungrouped data: The determination of median in an individual series involves the following steps:

- (i) The data are arranged either in ascending or descending order.
- (ii) Locate median by determining the value of $\left(\frac{N+1}{2}\right)^{th}$ item in the order, N being the

number of observations in the series. However, the value of median depends on N i.e whether N is odd or even number.

- (a) **When N is odd :** If N is odd then median is given by $Me = \left(\frac{N+1}{2}\right)^{th}$ item.

For example take the series 8, 12, 6, 4, 16, 13, 9. First we arrange these values in ascending order as 4, 6, 8, 9, 12, 13, 16. As N = 7, an odd number, thus median is

$$\begin{aligned} M_e &= \left(\frac{N+1}{2}\right)^{th} \text{ item} \\ &= \left(\frac{7+1}{2}\right)^{th} \text{ item} = 4^{th} \text{ item} \\ &= 9 \end{aligned}$$

- (b) **When N is even :** When N is an even number, $\left(\frac{N+1}{2}\right)$ will involve a fraction. In such cases the median is taken as arithmetic mean of two middle values.

$$M_e = \frac{\left(\frac{N}{2}\right)^{th} \text{ item} + \left(\frac{N}{2} + 1\right)^{th} \text{ item}}{2}$$

Example 1. The following data relate to the height of 8 students in a class.

S. No.	1	2	3	4	5	6	7	8
Ht. (in cm) :	153	142	151	144	149	146	141	150

Find median.

Solution. For computing median, we first arrange the data in ascending order as

142 144 146 147 149 150 151 153

As N = 8, an even number, then median is given by

$$\begin{aligned} Me &= \frac{\left(\frac{N}{2}\right)^{th} \text{ item} + \left(\frac{N}{2} + 1\right)^{th} \text{ item}}{2} \\ &= \frac{\left(\frac{8}{2}\right)^{th} \text{ item} + \left(\frac{8}{2} + 1\right)^{th} \text{ item}}{2} \\ &= \frac{4^{th} \text{ item} + 5^{th} \text{ item}}{2} \end{aligned}$$

$$= \frac{147 + 149}{2}$$

$$= 148 \text{ Cm.}$$

II. Grouped Data– Discrete Case : As you know the data is in grouped form *i.e.*, in the form of frequency distribution, it can be either in the form of discrete series or continuous series. The methods of computing median is different for these two types of data. Let us study them separately.

Discrete Series. The procedure in this method consists of the following steps:

- (i) Arrange the data in ascending order,
- (ii) Obtain the cumulative frequencies
- (iii) Determine the size of $\left(\frac{N+1}{2}\right)^{th}$ item, N being the total frequency.
- (iv) Median is located at the value of the variable in whose cumulative frequency the value of $\left(\frac{N+1}{2}\right)^{th}$ item falls.

Example 2. Find the median size of the following data :

Size (x)	:	4	10	8	5	9	7	6
Frequency	:	6	5	20	12	14	28	15

Solution : As a first step, arrange this data (X) in ascending order and prepare the following table by finding out cumulative frequencies.

Size (X)	Frequency (f)	CQM frequency
4	6	6
5	12	6 + 2 = 18
6	15	18 + 15 = 33
7	28	33 + 28 = 61
8	20	61 + 20 = 81
9	14	81 + 14 = 95
10	5	95 + 5 = 100
Total	N = 100	

Then for locating median, we find the value of

$$\left(\frac{N+1}{2}\right)^{th} = \left(\frac{100+1}{2}\right)^{th} = 50, 50^{th} \text{ item}$$

Then, as a last step, median is located at the value of item (X) in whose cumulative frequency the value of 50, 50th item falls. Thus

$$\text{Med} = 7$$

Continuous case : In this case, the procedure for calculating median is totally different than previous method. First we locate the median class by cumulating the frequencies until $\left(\frac{N}{2}\right)^{th}$ point is reached. Finally, the median is determined within this class by using an interpolation formula. The procedure thus involves the following steps :

Step : (i) Compute cumulative frequencies

Step: (ii) Find the size of $\left(\frac{N}{2}\right)^{th}$ item

Step: (iii) Locate the median class in which cumulative frequency column where the size of $\left(\frac{N}{2}\right)^{th}$ item falls.

Step: (iv) Obtain the median by using the following formula :

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

where

L = Lower limit of the median class.

c.f. = Cumulative frequency of the class preceding the median class.

f = Simple frequency of the median class

c = Class interval of the median

NOTE : If the given frequency distribution consists of inclusive classes, then true class limits or class boundaries for these classes should be obtained before computing median.

Example 3. Following is the distribution of marks of 50 students in a class.

Marks	:	0–10	10–20	20–30	30–40	40–50	50–60
No. of Students	:	4	6	20	10	7	3

Find Median.

Solution. Computation of Median

Marks	No. of Students f	$C.f.$
0–10	4	4
10–20	6	$4 + 6 = 10$
20–30	20	$10 + 20 = 30$
30–40	10	$30 + 10 = 40$
40–50	7	$40 + 7 = 47$
50–60	3	$47 + 3 = 50$

Since $N = 50$, $\frac{N}{2} = \frac{50}{2} = 25$ which falls in the cum frequency (30) of the class 20–30, thus the median class is 20–30. Therefore to calculate median, we have

$$\frac{N}{2} = 25, \quad L = 20, \quad f = 20$$

$$C.f. = 10 \quad i = 10, \quad \text{hence}$$

$$Me = L + \frac{\frac{N}{2} - C.f.}{f} \times i$$

$$= 20 + \frac{25 - 10}{20} \times 10$$

$$= 20 + 7.5$$

$$= 27.50$$

Example 4. Determine median for the following income distribution :

Income group : Below 100 100–200 200–300 300–400 400–500 above 500

No. of Persons : 5 10 18 30 10 17

Solution : It is an example of open-end frequency distribution as the first class interval is below 100 and last class interval is above 500. In such types of situations median can easily be obtained.

Income Group	No. of Persons f	C.f.
Below 100	5	5
100–200	10	15
200–300	18	33
300–400	30	63
400–500	20	83
Above 500	17	100

Since $\frac{N}{2} = \frac{100}{2} = 50$, thus the median class is 300–400. Thus

$L = 300$, $f = 30$ $c.f. = 33$, $c = 100$

$$\begin{aligned}
 \text{Med} &= L + \frac{\frac{N}{2} - c.f.}{f} \times i \\
 &= 300 + \frac{50 - 33}{30} \times 100 \\
 &= 300 + \frac{1700}{30} \\
 &= 300 + 56.67 \\
 &= 356.67
 \end{aligned}$$

Example 5. From the following data find the value of the median :

Class int. : 11–15 16–20 21–25 26–30 31–35 36–40 41–45 46–50

Frequency : 7 10 13 26 35 22 11 6

Solution. Here inclusive classes are given. Therefore, for median determination inclusive limits need to be converted into class boundaries as shown in the table.

Given Classes	Class Boundaries	Frequency <i>f.</i>	Cum freq. <i>c.f.</i>
11–15	10.5–15.5	7	7
16–20	15.5– 20.5	10	17
21–25	20.5–25.5	13	30
26–30	25.5–30.5	26	56
31–35	30.5–35.5	35	91
36–40	35.5–40.5	22	113
41–45	40.5–45.5	11	124
46–50	45.5–50.5	6	130

Since $\frac{N}{2} = \frac{130}{2} = 65$, thus the median class 30.5 – 35.5. Thus

$$\text{Med.} = L + \frac{\frac{N}{2} - C.f.}{f} \times i$$

$$= 30.50 + \frac{65 - 56}{35} \times 5$$

$$= 30.50 + \frac{9}{7}$$

$$= 30.50 + 1.2857$$

$$= 31.7857$$

$$\cong 31.79$$

2.2.2 Merits and Limitations of Median :

You have studied the meaning and computation of median alongwith the illustrations. Now let us discuss the merits and demerits of median.

Merits

- (i) For an open-end distribution, such an income distribution, the median given more representative value.
- (ii) It is not affected by the extreme items. It is of course affected by the number of items.
- (iii) Median minimizes the total absolute deviations i.e., the sum of absolute deviations from the median is the minimum.
- (iv) For dealing with qualitative phenomena, median is the most suitable average.

Limitations

- (i) Median is not capable of algebraic treatment.
- (ii) It is not based on all items of the series.
- (iii) It is affected more by sampling fluctuations.
- (iv) The median, in some cases, cannot be computed exactly as the mean when the number of items included in the series of data is even, the median is determined approximately as the mid-point of the two middle items.

2.3 PARTITION VALUES

As we know that the median is the middle value of the variable and it splits the series into two equal parts. That is why it is called positional average. In fact, there are other positional measures that partition the series into still more number of equal parts, say four (quartiles) or 10 equal parts (Deciles) or 100 parts (percentiles). These measures are known as portion values. We shall restrict ourself only to quartiles.

The values of a variable that divide the series into four equal parts are known as Quartiles. Since three points are required to divide the data in 4 equal parts, we have three quartiles Q_1 , Q_2 , and Q_3 . The first quartile (Q_1), known as a lower quartile, is the value of a variate below which there are 25% of the observations and above which there are 75% of the observations.

The second quartile (Q_2) is the value of a variate which divides the distribution into 2 equal parts. It means Q_2 is same as Median.

The third quartile (Q_3), known as upper quartile, is the value of a variate below which there are

75% observations, and above which only 25% observations. Thus it is clear that $Q_1 < Q_2 > Q_3$. The formula for Q_1 and Q_3 is

$$Q_1 = L + \frac{\frac{N}{4} - C.f.}{f} \times ic$$

and

$$Q_3 = L + \frac{\frac{3N}{4} - C.f.}{f} \times ic$$

where symbols have usual meanings.

Example 6. Calculate Q_1 and Q_3 from the following data :

Class :	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80
Freq. :	10	15	20	25	35	15	16	14

Solution. Computation of Q_1 and Q_3

Class	Frequency (f)	$C.f.$
0–10	10	10
10–20	15	25
20–30	20	45
30–40	25	70
40–50	35	105
50–60	15	120
60–70	16	136
70–80	14	150

Computation Q_1 : Since $\frac{N}{4} = \frac{150}{4} = 37.50$ lies in class (20–30), as it falls in the $c.f.$ of 45, thus, Q_1 is given by

$$Q_1 = L + \frac{\frac{N}{4} - C.f.}{f} \times i$$

$$\begin{aligned}
&= 20 + \frac{37.50 - 25}{20} \times 100 \\
&= 20 + \frac{12.50}{2} = 20 + 6.25 \\
&= 26.25
\end{aligned}$$

Computation of Q_3 : Since $\frac{3N}{4} = \frac{3 \times 150}{4} = 112.50$ lies in (50–60) class as it falls in the *C.f.* 120, thus Q_3 is given by

$$\begin{aligned}
Q_3 &= L + \frac{\frac{3N}{4} - C.f.}{f} \times i \\
&= 50 + \frac{112.50 - 105}{15} \times 10 \\
&= 50 + \frac{7.50}{1.50} = 50 + 5.00 \\
&= 55.00
\end{aligned}$$

2.4 MODE

The word ‘Mode’ comes from the French word ‘la mode’ which means the fashion. In statistical language, the mode is that value of a variate which occurs most often in a series, i.e., a value of a variate which is repeated most often in data set. But it is not exactly true for every frequency distribution. Rather it is that value of the variate around which the other items tend to concentrate most heavily. It shows the centre of concentration of the frequency in and around a given value. It is commonly denoted by M_o . For example, take the case of a shopkeeper who sells shoes. He is interested to know the sizes of shoes which are commonly demanded. Here in such a situation, mean would indicate a size that may not fit any person. Median may also not provide a representative size because of the unevenness in the distribution. It is the mode which will help in making a choice of approximate size for which an order can be placed.

CHECK YOUR PROGRESS-I

- **MULTIPLE CHOICE QUESTIONS**

- 1. Which measure of central tendency represents the value that appears most frequently in a data set?**
 - a) Mean
 - b) Median
 - c) Mode
 - d) Range
- 2. What is the median of the data set: 3, 10, 7, 2, 9, 12, 13?**
 - a) 7
 - b) 9
 - c) 10
 - d) 12
- 3. If a dataset has an even number of observations, how is the median calculated?**
 - a) The average of the two middle numbers
 - b) The highest number
 - c) The lowest number
 - d) It cannot be calculated
- 4. If the mode of a data set is 10 and the mean is 4, what is the median?**
 - a) 5
 - b) 6
 - c) 7
 - d) 8

5. Which statement is true regarding the mode?

- a) A dataset can have only one mode.
- b) The mode is always greater than the median.
- c) A dataset can have more than one mode.
- d) The mode is always smaller than the mean.

6. What is the mode of the following data: 2, 4, 4, 7, 2, 5, 4, 8?

- a) 2
- b) 4
- c) 5
- d) 7

2.4.1 COMPUTATION OF MODE :

The method of computing mode is different for grouped and ungrouped data. Now let us study those method separately.

(i) Ungrouped Data or Individual Series : For an ungrouped data or individual series the mode can be located simply by inspection. Here, the value that occurs most frequently in the data is taken as a mode. For example, the ages (in years) of 10 boys are 9, 11, 10, 14, 17, 14, 9, 11, 12, 11. Here the number 11 appeared thrice. Therefore, mode age is 11.

In some cases there may be more than one mode. For example; 8, 7, 12, 10, 8, 12, 8, 12, 6, 5. In this case both the numbers 8 and 12 appear equal number of times (three). Therefore, there are two modes; 8 and 12.

(ii) Discrete Series : In discrete series; when the values of individual items are known, mode can be determined just by inspection. By inspection you can find out the value of the variate ordering which the items are most heavily concentrated.

Example 7. Find the mode for the following data :

Size of item	:	8	9	10	11	12	13	14
Frequency	:	10	12	18	16	19	14	12

Solution : In this frequency distribution size of item 12 has the highest frequency, implying that

there is a heavy concentration of items at this value. Therefore mode is 12.

In a series like this it is easy to obtain mode. Difficulty arises when nearly equal concentrations are found in two or more neighbouring classes. To locate modal class in situations, there is a need for Grouping Method. In grouping method the values are first arranged in ascending order alongwith their frequencies. Normally, the grouping table has the following six columns :

- Column 1. The maximum frequency is marked by putting a mark or a circle.
- Column 2. The frequencies are grouped in two's and the highest total is marked.
- Column 3. Leaving the first frequency, the remaining frequencies are grouped in twos and the highest total is marked.
- Column 4. The frequencies are grouped in threes, the highest total is marked.
- Column 5. Leaving the first frequency, the remaining frequencies are grouped in threes and the highest total is marked.
- Column 6. Leaving the first two frequencies, the remaining are grouped in threes and the highest total is marked. After completing the grouping table, an analysis table is formed for finding the value which is repeated the highest number of times. The same procedure is adopted for determining modal class in the case of continuous series.

(iii) Continuous Series : In a continuous series, frequencies are given in various classes. A class having maximum frequency is called modal class. In case nearly equal concentration of frequencies is observed in two or more classes, the grouping method may be used to determine the modal class. After determining the modal class, the mode can be obtained by using the following formula :

$$M_o = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

where

L = Lower limit of the modal class

f_1 = Frequency of the modal class

f_0 = Frequency of the class preceding the modal class

f_2 = Frequency of the class succeeding the modal class

i = Class interval of the modal class

Note : (i) While applying this formula, it is necessary to have uniform class intervals. If they are unequal, first they should make equal.

(ii) Where mode is ill-defined, its value may be obtained by using mean and median as

Mode = 3 Median – 2 Mean This is called empirical mode.

Example 8. Obtain mode of the following distribution :

Class	:	10–20	20–30	30–40	40–50	50–60	60–70
Frequency	:	8	12	25	45	11	9

Solution.

Class	Frequency
0–20	8
20–30	12
30–40	25 → f_0
40–50	45 → f_1 →
50–60	11 → f_2
60–70	9

highest frequency

$$\begin{aligned}
 \text{MQ} &= L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i \\
 &= 40 + \frac{45 - 25}{2 \times 45 - 25 - 11} \times 10 \\
 &= 40 + \frac{20}{54} \times 10 \\
 &= 40 + 3.70 = 43.70
 \end{aligned}$$

2.4.2 Merits and Limitations of Mode:

Merits

2.4.2.1 In certain cases, mode is the only suitable average. For example, modal size of garments, shoes, modal wages.

2.4.2.2 Its value can be determined in open-end frequency distributions of garments, shoes, modal wages.

- (ii) Its value can be determined in open-end frequency distributions.
- (iii) The value of mode can also be obtained graphically.
- (iv) It can be used to describe qualitative phenomena.
- (v) Like median, mode is not unduly affected by extreme values.

Limitations:

- (i) The value of mode cannot always be determined.
- (ii) The value of mode is not unique.
- (iii) The value of mode is not based on all the observations.
- (iv) The value of mode is affected unduly by the size of class intervals.

2.5 LET US SUM UP

In case of calculation of mean, we see there was a limitation of affecting its value by changing the value of extreme item.

But while calculating the value of median or mode, this limitation is automatically overcome. Here in this lesson, we learned how to calculate median, mode and other partition values such as quantities, deciles and percentiles. All these have their different utilities of calculating

2.6 GLOSSARY

- **Median:** This measure of central tendency divides whole series into two equal parts.
- **Mode:** This refers to those values in data which repeats most/ maximum times.
- **Quartile:** Quartile divides whole of the series into four equal parts.
- **Deciles:** It divides whole data into ten equal sets.
- **Percentiles:** It divides whole data set into 100 equal parts.

2.7 SELFASSESSMENT QUESTIONS

1. What is Median? Explain its limitations.

-
-
2. Define mode. How is it different than median
-
-
-

3. From the following data, obtain

(i) Median (ii) First and third quartiles (iii) Mode

Wages Daily (in Rs)	No. of employees
20–40	8
40–60	12
60–80	20
80–100	30
110–120	40
120–140	35
140–160	18
160–180	7
180–200	5

4. Determine the model size of the collar by using grouping method.

Collar Size (in cm) :	32	33	34	35	36	37	38	39	40	41
No. of Students :	7	14	30	28	35	34	16	34	36	16

2.8 LESSON END EXERCISE

1. Which tool of statistics divides series into two equal parts
- a) Mean b) Median c) Mode

2. _____ Indicates the quantity of data repeated maximum times in series.
3. The formula of calculating Q3 is _____

2.9 SUGGESTED READINGS

1. Argyrous, George. 1997. Statistics for Social Research. New York: Mc Millan Press Ltd.
2. Goods, W.J. & Hatt, P.K. 1981. Methods in Social Research. New York: Mc Graw Hill.
3. Gupta, S.C. 1981. Fundamentals of Statistics. Bombay: Himalayan Publishing House.
4. Gupta, S.P. 2004. Statistical Methods. New Delhi: Sultan Chand and Sons.

2.10 ANSWERS TO CHECK YOUR PROGRESS

Answers: 1- c) Mode

2- Answer: b) 9 (Ordered: 2, 3, 7, **9**, 10, 12, 13)

3- a) The average of the two middle numbers

4- Answer: b) 6 (Formula: $\text{Mode} = 3(\text{Median}) - 2(\text{Mean})$ $10 = 3(\text{Median}) - 2(4)$ $3(\text{Median}) = 18$)

5- c) A dataset can have more than one mode

6- b) 4 (4 appears three times)

MEASUREMENT AND SCALING

STRUCTURE

- 3.0 Objectives
- 3.1 Introduction
- 3.2 Meaning of Scale
- 3.3 Types of Scale
- 3.4 Levels of Measurement
- 3.5 Let us sum up
- 3.6 Glossary
- 3.7 Self-Assessment Questions
- 3.8 Lesson End Exercise
- 3.9 Suggested Readings
- 3.10 Answers to check your progress

3.0 OBJECTIVES

Dear learner, after going through this lesson, you should be able to:

- Understand the meaning of scaling.
- Types of Scale
- Understand the steps involved in constructing scales

3.1 INTRODUCTION

Scales are techniques employed by social scientists in the area of attitude measurement. They consist of a number of statements or questions and a set of response categories, related to a score. They place respondents in a continuum between very low (or negative), over a neutral, to a very high (or positive) position. Each item is chosen so that persons with different points of view on this item react to be in a different way. In this sense they are a part of surveys and questionnaires and are considered during the process of questionnaire construction.

3.2 MEANING OF SCALE

Scaling involves a high degree of operationalization and allows researchers to measure complex issues. Furthermore, it enables researchers to summate values of several variables into one score and this with a relatively high degree of reliability. In general, it offers respondents a choice of picking their answers out of given sets of alternatives, which as we shall see, are established in a very careful but also a cumbersome way.

There is nominal, ordinal and interval/ratio scales. Of these, nominal scales are not very common. Most popular are the Likert scales, the Thurstone scales and the Guttman scales, which do not use nominal measurement.

Scales vary not only in their level of measurement but also in their aims and their method of construction. Some are constructed by means of a very complicated process, while others are built in a relatively simple manner. In all cases, however, there are some basic points the experts such as Edwards (1957) and Likert (1932) some time ago said should be considered during scale construction—points that are still respected and practiced in social research today many investigators. The following are some examples:

1. Language must be simple, clear and direct.
2. Items must be brief (up to 20 words) and contain one issue only. Complex sentences must be avoided.
3. Items referring to past events and factual items must be avoided.
4. Ambiguous and irrelevant items must be avoided.
5. Items that may be accepted or rejected by all respondents must be avoided.
6. Words such as all, always, no one, never, only, exactly, almost should be avoided.
7. Use of professional jargon and double negations should be avoided.
8. Response categories must be mutually exclusive, exhaustive and unidimensional (i.e. measuring one single construct).

- **Reasons for using scales**

Scales are used for a number of reasons. Apart from general methodological motives, the following reasons are most common (see Vlahos. 1984):

- **High coverage:** Scales help to cover all significant aspects of the concept
- **High precision and reliability:** Scales allow a high degree of precision and reliability.
- **High comparability:** The use of scales permits comparisons between sets of data.
- **Simplicity:** Scales help to simplify collection and analysis of the data.

Scales are a most useful tool of social research and also one that is very difficult to construct. Construction and statistical testing are very involving and time-consuming tasks and therefore not easily accessible to the ordinary researcher. However, researchers developed and tested in the past a very large number of scales which have been adequately tested and are available to other researchers to use. In this sense, scale construction is less common than scale use. Scale construction may be a step to consider after having completed your current course of study. In the meaning using already available scales may be the way to go when addressing issues for which scales are available.

3.3 TYPES OF SCALE

1. THE THURSTONE SCALE

Description

This scale was developed in the USA in the 1920s; it consists of a list of items constructed with the aid of experts who are very closely related to the construction of the scale. It is employed mainly in the area of attitude measurement, and is developed through a cumbersome and demanding process, as explained below.

Construction

The construction of the scale is as follows:

Steps 1

The researcher selects a number of relevant statements containing a set of response categories ('agree', 'disagree') allowing respondents to express their attitudes to the issue in question freely.

Step 2

These statements are given to a number of judges, who are asked to order them on a continuum

from 1 to 11, according to the way they judge the statements. If in the opinion of the judge the statement describes the most favourable attitudes to the study object, it is given the score 1; if it describes slightly less favourable attitudes, it is given the score 2 and so on. In this way, statements are allocated a scale value.

Step 3

The statements are scrutinized in terms of the value they received from the judges. Statements that were ordered by the judges uniformly are retained and given an average scale value (the closest to the average); those that received a diverse value are discarded.

Step 4

The remaining statements are processed further by the researcher, and their number, reduced. The resulting scale is constructed so that statements are distributed evenly between 1 and 11 and each statement is identified through its scale value.

Evaluation

Although Thurstone scales are still used, they are criticized. Among other things, for their demanding and time-consuming manner of construction, and the emphasis they place on the views of the judges. They are a valuable tool of methodology, and are employed as the sole technique or together with other methods of attitude measurement.

2. THE LIKERT SCALE

Description

Developed by Likert in 1932, this scale operates in a way similar to that of the Thurstone scale. It consists of a set of items of equal value and a set of response categories constructed around a continuum of agreement/disagreement to which subjects are asked to respond. It is very popular among social scientists, is relatively easy to construct and is believed to be more reliable than the Thurstone scale.

Construction

Likert scales are constructed in the following way:

Step 1

A number of items related to an issue are collected. In general, 80 to 120 items are thought to be sufficient, but four times as many items as needed are generally considered.

Step 2

Five-answer response categories are assigned to each item, ranging from 'strongly agree' through

‘agree’, ‘undecided’ and ‘disagree’ to ‘strongly disagree’ including numerical values, for example from 1 to 5 respectively.

Step 3

Statements are administered to respondents in a pilot study, and total scores are computed and further processed to determine, for instance, one-dimensionality, that is measuring one and the same concept (usually through factor analysis), and internal consistency (e.g. correlation with the total score is calculated).

Step 4

Items with a substantial correlation are retained; items with low correlation are discarded. The constructed scale is then administered to all respondents.

Example. There is a lot of sexism going on in this community.

<i>Strongly agree</i>	<i>Agree</i>	<i>Undecided</i>	<i>Disagree</i>	<i>Strongly disagree</i>
<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>

Evaluation

Likert scales are very popular among social scientists and have been so for more than half a century. The reason for this is that they: (1) have a high degree of validity even if the scale contains only a few items; (2) provide single scores from a set of items; (3) have a very high reliability (between 0.85 and 0.94); (4) allow ranking of the respondents; and (5) are relatively easy to construct. Nevertheless, researchers point to some drawbacks of this method. For example, total scores referring to many and diverse items say little about a person’s response to the various aspects of the research object; also, it is difficult to achieve equal items in the scale (Kimmon, 1990).

3. THE BOGARDUS SOCIAL DISTANCE SCALE

Description

This scale was developed in the USA and was employed to measure ‘social distance’ between the respondents and persons of other nationalities or races: it is still used to determine how close a respondent is willing to place himself or herself to persons of other races or nationalities.

The scale consists of a number of statements that indicate the degree of distance between the respondent and the groups under study. More particularly the respondents are asked to state their reactions to a set of statements varying in intensity of closeness to a population group. As a concrete example, respondents could be asked to state which of the following seven statements (which actually make up the scale) reflect accurately and honestly their true feelings towards Aborigines, and whether they would accept

an Aborigine as a :

- close relative by marriage
- personal friend
- neighbor
- colleague at work
- speaking acquaintance only
- visitor to their country
- person to be kept out of their country

Interpretation

The results obtained through this procedure are evaluated as shown below:

- Compute the mean values for each group.
- Rank each group according to the value of the mean.
- The higher the value, the greater the social distance, that is the lower the willingness to assume contacts with that group; and the stronger the negative prejudice and attitude to that group.

Application

This scale, although originally developed to measure distance among ethnic groups, can be equally successfully employed in other areas, for example in market research and studies of race relations. One could, for instance, develop a range of questions related to a certain item (car, television set, record player, etc.) that could best describe a person's intention and willingness to buy this item. For example, questions ranging from 'I would most certainly buy this product' to 'I will never buy such a thing in my life' can be used to measure the degree of a person's readiness to purchase the item.

Evaluation

This scale has been used very extensively by social scientists. The three most common advantages of the scale are the following (see Kimmon, 1990):

1. A very high split-half reliability (r is equal to or greater than 0.90).
2. A high content validity of the scale items.

3. A satisfactory overall validity and reliability.

Although there are some problems associated with the construction of the steps of the scale and their order, the scale is considered to be a very useful tool of social research.

4. THE GUTTMAN SCALE

Description

This is another scale that measures social distance, or rather proximity'. It consists of a number of statements placed in a hierarchical order ranging from low to high in such a way that if respondents reject one statement, they will also reject all other statements above it; and if they accept one statement they will accept all other statements below it.

Respondents are normally asked to state whether they agree or disagree with each of the statements. The results obtained are expected to show the degree of proximity or distance of the subjects from the research object (e.g. migrants, blacks, homosexuals etc.). More particularly, it will show how far the respondents will allow certain people to come close to them.

Construction

'Construction of such a scale is complicated and time consuming. In a simplified form it can be constructed in the following way:

Step 1

A number of statements thought to be cumulative, that is, they fall in a hierarchical order ranging from low to high, are formulated in such a way that if respondents reject one statement, they will also reject all other statements above it; and if they accept one statement, they will accept all other statements below it.

Step 2

These statements are presented to a number of subjects (say, 10), who are asked to state whether they agree or disagree with each statement.

Step 3

A table with the numbers of the statements on the top, and the side, is constructed; the agreements of the subjects with each statement are entered (note that disagreements are not recorded).

Step 4

The statements are then ordered so that the one accepted by one subject only is placed first, the

statement accepted by two subjects second, the statement accepted by three subjects third and so on.

Step 5

The reproducibility value, which is I minus the fraction consisting of the number of errors (numerator) and the number of responses (denominator), is computed. If the score is 0.90 or better the scale is satisfactory.

Evaluation

This scale has been employed very extensively in the past and is still considered to be valid and useful way of measuring social proximity. But it is considered to be more cumbersome than the Bogardus social distance scale, which is used more frequently.

5. THE SEMANTIC DIFFERENTIAL SCALE

Description

This technique was developed by Osgood, Suci and Tannenbaum in 1957 and has been used by social scientists to measure the impression concepts make on people and the meaning they invoke. Concepts are measured independently as well as in comparison with other concepts, and can be related to a variety of contexts, issues or objects, in this way allowing the researcher to draw relevant conclusions about the respondents.

The semantic differential scale consists of a number of opposite concepts, which may range from 7 to over 70. Examples of such opposites are given below. The data sheet containing the sets of opposites is administered to the respondents with instructions to place an individual (e.g. a teacher) or a group of individuals (e.g. Asian migrants) in a specific position between the extremes of a continuum.

Example. Some opposites

Good	6	5	4	3	2	1	0	Bad
Democratic	6	5	4	3	2	1	0	Authoritarian
Sociable	6	5	4	3	2	1	0	Unsociable
Strong	6	5	4	3	2	1	0	Weak
Flexible	6	5	4	3	2	1	0	Rigid
Cooperative	6	5	4	3	2	1	0	Uncooperative
High	6	5	4	3	2	1	0	Low
Hard	6	5	4	3	2	1	0	Soft

Conformist	6	5	4	3	2	1	0	Non-conformist
Fair	6	5	4	3	2	1	0	Unfair
Difficult	6	5	4	3	2	1	0	Easy
Active	6	5	4	3	2	1	0	Passive
Sharp	6	5	4	3	2	1	0	Dull
Independent	6	5	4	3	2	1	0	Dependent
Irritable	6	5	4	3	2	1	0	Calm
Hot	6	5	4	3	2	1	0	Cold
Harmonious	6	5	4	3	2	1	0	Unharmonious

The numerals indicate the degree of agreement or disagreement of the subjects regarding the concepts under evaluation. In the example, 6 stands for very good, strong, high, etc., 5 for moderately good, strong, etc., 4 for fairly good, strong, etc., 3 for undecided, 2 for fairly bad, unsociable, weak low, etc., 1 for moderately bad, unsociable, weak, etc., and 0 for very bad, unsociable weak, low, etc.

The subject's judgment is based on three distinct characteristics, namely evaluation of the individual, judgment of the potency or power of the individual, and judgment of the activity of the individual. General evaluation is judged by opposites such as good-bad, sociable-unsociable, high-low and harmonious-unharmonious. Potency is judged by means of opposites such as hard-soft, large-small, difficult-easy and unyielding-lenient. Activity is judged by opposites such as hot-cold, active-passive, sharp-dull and irritable-calm. Of these three dimensions the first (evaluation) seems to be the most important.

Respondents are advised to evaluate the study person or group, by indicating the number that corresponds to their feelings on the specific item. If the respondents think that the person in question is moderately good, they are advised to circle '5' at the 'good' and 'bad' item; if they feel that this person is fairly unsociable, they should circle '2' in the second line, and so forth. Each circle represents a score which can be high or low depending on the subject's judgment of the concept or the individual, for example the teacher. When the evaluation is completed, a total score for the impression of the concept or the person in question is computed by adding up all individual scores. A high score represents a high impression and a low score indicates a low impression of the concept or the person.

This scale can be employed successively in a number of different groups, such as Asian migrants, Italian migrants and British migrants, allowing comparisons to be made between these groups.

Interpretation

The results of this procedure can be interpreted and presented in many ways. The method of adding up the individual scores mentioned above is one. Drawing profiles, computation of correlation

coefficients and of the semantic distances are other ways.

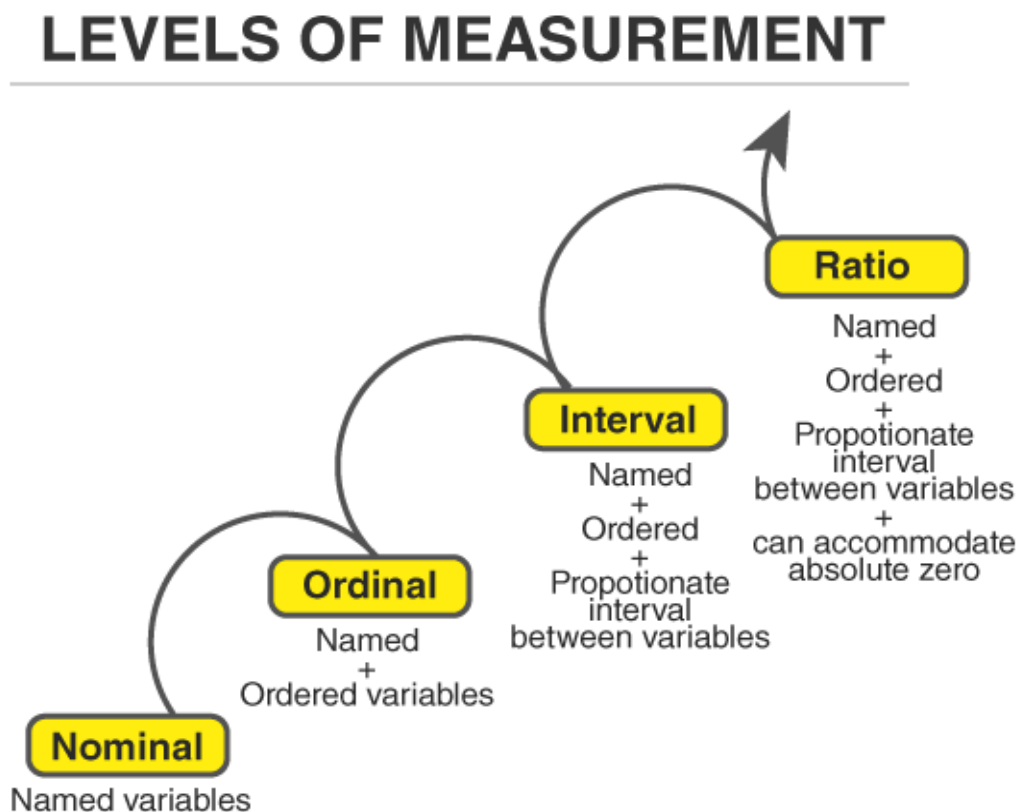
Evaluation

The semantic differential method offers precise information about the attitudes of people toward others. It allows evaluation of concepts, comparisons and measurement of different types on the same measure, and is relatively easy to construct. It has, however, to be treated with caution. For instance, a long list of points to choose from might cause confusion and also inaccurate results. The use of equal intervals or ordinal data is another issue. Definitions of the concepts and their meanings might vary from one respondent to another, causing problems and distortions.

3.4 LEVELS OF MEASUREMENTS

There are four different scales of measurement. The data can be defined as being one of the four scales. The four types of scales are:

- Nominal Scale
- Ordinal Scale
- Interval Scale
- Ratio Scale



1. Nominal Scale

A nominal scale is the 1st level of measurement scale in which the numbers serve as “tags” or “labels” to classify or identify the objects. A nominal scale usually deals with the non-numeric variables or the numbers that do not have any value.

Characteristics of Nominal Scale

- A nominal scale variable is classified into two or more categories. In this measurement mechanism, the answer should fall into either of the classes.
- It is qualitative. The numbers are used here to identify the objects.
- The numbers don't define the object characteristics. The only permissible aspect of numbers in the nominal scale is “counting.”

Example:

An example of a nominal scale measurement is given below:

What is your gender?

M- Male

F- Female

Here, the variables are used as tags, and the answer to this question should be either M or F.

2. Ordinal Scale

The ordinal scale is the 2nd level of measurement that reports the ordering and ranking of data without establishing the degree of variation between them. Ordinal represents the “order.” Ordinal data is known as qualitative data or categorical data. It can be grouped, named and also ranked.

Characteristics of the Ordinal Scale

- The ordinal scale shows the relative ranking of the variables
- It identifies and describes the magnitude of a variable
- Along with the information provided by the nominal scale, ordinal scales give the rankings of those variables
- The interval properties are not known
- The surveyors can quickly analyse the degree of agreement concerning the identified order of variables

Example:

- Ranking of school students – 1st, 2nd, 3rd, etc.
- Ratings in restaurants
- Evaluating the frequency of occurrences
- Very often
- Often
- Not often
- Not at all
- Assessing the degree of agreement
- Totally agree
- Agree
- Neutral
- Disagree
- Totally disagree

3. Interval Scale

The interval scale is the 3rd level of measurement scale. It is defined as a quantitative measurement scale in which the difference between the two variables is meaningful. In other words, the variables are measured in an exact manner, not as in a relative way in which the presence of zero is arbitrary.

Characteristics of Interval Scale:

- The interval scale is quantitative as it can quantify the difference between the values
- It allows calculating the mean and median of the variables
- To understand the difference between the variables, you can subtract the values between the variables
- The interval scale is the preferred scale in Statistics as it helps to assign any numerical values to arbitrary assessment such as feelings, calendar types, etc.

Example:

- Likert Scale
- Net Promoter Score (NPS)
- Bipolar Matrix Table

4. Ratio Scale

The ratio scale is the 4th level of measurement scale, which is quantitative. It is a type of variable measurement scale. It allows researchers to compare the differences or intervals. The ratio scale has a unique feature. It possesses the character of the origin or zero points.

Characteristics of Ratio Scale:

- Ratio scale has a feature of absolute zero
- It doesn't have negative numbers, because of its zero-point feature
- It affords unique opportunities for statistical analysis. The variables can be orderly added, subtracted, multiplied, divided. Mean, median, and mode can be calculated using the ratio scale.
- Ratio scale has unique and useful properties. One such feature is that it allows unit conversions like kilogram – calories, gram – calories, etc.

Example:

An example of a ratio scale is:

What is your weight in Kgs?

- Less than 55 kgs
- 55 – 75 kgs
- 76 – 85 kgs
- 86 – 95 kgs
- More than 95 kgs

CHECK YOUR PROGRESS-I

Answer the below Questions:

1.Q: Which scale simply classifies or labels items into categories (e.g., gender)?

2.Q: Which scale ranks data in order but lacks equal intervals (e.g., satisfaction level)?

3.Q: Which scale features equal intervals but no true zero point (e.g., Celsius)?

4.Q: Which scale possesses an absolute zero, allowing for ratio comparisons (e.g., weight)?

5.Q: What is the most powerful level of measurement?

6.Q: What is the least powerful level of measurement?

7.Q: Which scale is used when asked to rank your preference from 1st to 5th?

8.Q: Which scale is used to measure temperature?

9.Q: What is the, “0” in a ratio scale called?

10.Q: Which scale is used when assigning numbers to players on a team?

3.5 LET US SUM UP

The concept and practice of measurement are two important and also controversial issues. However, the controversy in this case is not about whether to employ measurement in social research or not but rather about how and in what way measurement should be employed. The practice of measurement is well accepted in social research, regardless of type and nature. Some studies may use nominal measurement; others may use ordinal and others interval/ratio measurement. All types of measurement are employed. The notion that one type of research is better than the other is incorrect. Qualitative researchers may opt for nominal measurement, but this does not make other types of measurement less effective. In one and the

same research instrument one may find some variables being measured at the nominal level and others at the ordinal or interval/ratio level. The latter provides different types of information than the former but it nevertheless produces equally useful information.

The level of measurement is useful for itself, but more so for further research and analysis. The level of measurement determines the type of measures that are too employed in the analysis. As we shall see in Chapters, there is a close

relationship between level of measurement, type of variable and statistical tests. For this reason, having a clear understanding of the level of measurement is important for doing research, and for assuring high level of accuracy.

Measurement together with objectivity and ethics on one hand and with validity and reliability on the other constitute major principles of social research. The latter are central to any type of research, regardless of its nature and ideological affiliation. Adherence to reliability and validity is a fundamental requirement which researchers have to consider seriously when doing research. Reliability and validity are indicators of consistency, truthfulness and accuracy, and such concepts are structural ingredients of any type of research.

Measurement, validity and reliability, together with scaling, which were discussed in the last part of this chapter, are very useful research tools. They help establish the parameters for producing well-founded and respectable findings.

3.6 GLOSSARY

- **Scaling** makes research work simple by removing the complexities involved in measurement of data.
- **Likert Scale** is the most popular Scale.

3.7 SELF-ASSESSMENT QUESTIONS

1. What do you mean by scaling.

2. Give the reasons, why Scales are used.

3. Explain in detail various types of Scales and steps involved in each scale for constructing.

3.8 LESSON END EXERCISE

1. Scale helps primarily in_____
- a) Measuring the observation b) collecting data
- c) Sampling the data collection d) All of the above
2. _____ is the basic feature of a good scale.
3. Likert Scale was developed by _____ in _____

3.9 SUGGESTED READINGS

1. Argyros, George. 1997. Statistics for Social Research. New York: Mc Millan Press Ltd.
2. Gupta, S.C. 1981. Fundamentals of Statistics. Bombay: Himalayan Publishing House.
3. Gupta, S.P. 2004. Statistical Methods. New Delhi: Sultan Chand and Sons.

3.10 ANSWERS TO CHECK YOUR PROGRESS

Answer

- | | | | |
|---|-----------------------|----|-----------------------|
| 1 | <u>Nominal</u> | 2 | <u>Ordinal</u> |
| 3 | Interval | 4 | Ratio |
| 5 | Ratio | 6 | Nominal |
| 7 | Ordinal | 8 | Interval |
| 9 | Absolute | 10 | Nominal |

RELIABILITY AND VALIDITY IN QUANTITATIVE RESEARCH**STRUCTURE**

- 4.0 Objectives
- 4.1 Introduction
- 4.2 Meaning of Reliability
- 4.3 Types of Reliability
- 4.4 Meaning of Validity
- 4.5 Types of Validity
- 4.6 Difference between Validity and Reliability
- 4.7 Let us sum up
- 4.8 Glossary
- 4.9 Self- Assessment Questions
- 4.10 Lesson End Exercise
- 4.11 Suggested Readings
- 4.12 Answers to check your progress

4.0 OBJECTIVES

Dear learner, the main objectives of this lesson is to understand:

- Meaning and Definitions of Validity and Reliability.
- Types of Validity and Reliability

- Importance of Validity and Reliability in Measurement.
- Difference between Validity and Reliability.

4.1 INTRODUCTION

Science depends on accurate and systematic measurement. Because researchers must demonstrate that they are recording events accurately, scientific instruments are tested regularly for accuracy. Obviously, instruments that do not give true reading are not useful. Though dependence on instruments is necessary for all science demonstrating reliability and validity in the social science is often more difficult than it is in the natural sciences. In the natural sciences *e.g.*, official standards for items such as weight, temperature or chemical purity are available for testing instruments.

Social scientists do not have this luxury. Measuring such things as attitude or intelligence is very difficult because there is no universally accepted official standards. The credibility of field studies, naturalistic observations, and archival research depends on clear and convincing evidence that recording techniques are acceptable. Thus, investigation must demonstrate that behavioral measures are reliable and valid. Reliability and validity refer to data collection. That is, they refer to whether data recording devices are reliable and valid and to whether surveys, tests or observational systems really addresses what the investigator is studying.

4.2 MEANING OF RELIABILITY

Reliability refers to the ability of an instrument to produce consistent or same results. Since a grocer obtains the true measure of a commodity by a kilogram, a cloth merchant obtains true length of cloth by a meter, and a tailor by an inch-tape, these measuring instruments have to be reliable. Reliability is the degree to which measures are free from error so that they give same results when repeat measurements are made under constant conditions. If there are imperfections in the measuring process and the respondent misunderstands the question, or understands the question but does not give a truthful response, it will be the cause of low reliability of measurement.

4.3 TYPES OF RELIABILITY

There are four types for testing the reliability of an instrument, these are:

1. **Test-Retest Reliability:** This means administering the same scale or measure to the same respondents at two separate times for stability. It will be reliable if the reported test administered under conditions similar to the first test obtains similar results.
2. **Internal consistency reliability:** It refers to the degree of agreement between various items on the measurement device. While assessing aggression among children on a playground, one could

record many types of behavior. These could be acts of physical violence, vocal outbursts, angry gestures of facial expressions etc. One would record many types of each and then check to see if certain behavior correlates with others.

3. **Split half reliability:** Here responses, to the items of an instrument are divided and the scores correlated. The degree of co-relation indicates the degree of reliability of measurement. The test could alternatively be divided into more parts– thirds, quarters etc.; provided all the items are comparable. The correlation is then corrected to give the stepped-up reliability of the whole test.
4. **Equivalent form reliability:** It is utilized when two alternative instruments are designed to be as equivalent or possible. Each of the two measurement scales is administered to the same group of subjects. If there is high correlation between the two forms, the researcher assumes that the scale is reliable.

4.4 MEANING OF VALIDITY

Validity means the ability to produce findings that are in agreement with conceptual or theoretical values. *e.g.*, an attitude measurement technique may indicate that 80 percent people are in favor of using family planning measures. But 80 percent people may not actually use these methods. A reliable but invalid instrument will yield consistently inaccurate results. So, validity refers to the success of the scale in measuring what is meant to be measured. Many a times, the scale used may be reliable but it measures something other than what it was designed to measure.

4.5 TYPES OF VALIDITY

1. **Empirical Validity:** It tests pragmatic or criterion validity. If an instrument has, for instance, produced results indicating that students involved in student union activities do better in their exams, and if this is supported by available data, the instrument in question has pragmatic validity. Again, validity here is assumed if the findings are supported by already existing empirical evidences. In this case validity is concurrent validity.
2. **Internal/External Validity:** Internal validity ensures that the study design eliminates confounding variables, while external validity relates to the generalizability of the findings.
3. **Theoretical Validity:** It is employed when empirical confirmation of validity is difficult or not possible. A measure is taken to have theoretical validity if its findings comply with the theoretical principles of the disciplines, that is, if they don't contradict already established rules of the discipline.
4. **Face Validity:** An instrument has face validity if it seems to measure what it is expected to measure "on the face of it". In such a case, it appears to have validity, *e.g.*, a questionnaire aimed at studying sex discrimination has face validity if its questions refer to discrimination due to sex. The standard of evidence here is not based on empirical evidence, as it was in the case of the other types of

validation, but on general theoretical standards and principles, and on the subjective judgments of the researcher.

5. **Content Validity:** A measure is supposed to have content validity if it covers all possible aspects of the research topic. If a measure of operation, for instance, does not include normlessness or powerlessness the researcher cannot claim content validity for this instrument.
6. **Construct Validity:** A measure can claim construct validity if its theoretical construct is valid. For this reason, validation concentrates on the validity of the theoretical construct. For example, if discrimination of female students is the research topic, we proceed as follows: an instrument is constructed to study this topic. Then two student groups known to differ in their views on basic issues related to the research question are identified. Next the instrument whose validity is to be checked is administered to both groups of the results recorded separately for each group. If the findings obtained from each group differ, the instrument is thought to have construct validity.

CHECK YOUR PROGRESS-I

Multiple Choice Questions

1. Validity in quantitative research refers primarily to:

- A. The consistency of measurements over time.
- B. The degree to which a tool measures what it is intended to measure.
- C. The ability of a study to be replicated.
- D. The stability of the findings across different observers.

2. A questionnaire can be reliable but not valid, but a valid questionnaire is always:

- A. Very short.
- B. Extremely expensive to produce.
- C. Validated by thousands of participants.
- D. Reliable.

3. Which type of reliability measures the consistency of items within the same instrument, often using Cronbach's alpha?

- A. Test-retest reliability
- B. Inter-rater reliability

C. Internal consistency

D. Parallel forms reliability

4. When two or more observers rate the same behavior and provide similar scores, the study has high:

A. Internal validity

B. Inter-rater reliability

C. Construct validity

D. Test-retest reliability

5. Which form of validity ensures the instrument includes a representative sample of all possible questions on the topic?

A. Content validity

B. Criterion validity

C. Construct validity

D. Face validity

6. Which method is commonly used to establish the reliability of a test by administering it twice to the same group at different times?

A. Split-half

B. Item analysis

C. Test-retest reliability

D. Convergent validity

7. If a researcher checks if a new depression scale corresponds with a well-established existing depression scale, they are testing:

A. Face validity

B. Criterion validity

C. Internal consistency

D. Construct validity

8. Which factor is most likely to affect the reliability of a research instrument?

A. The sampling method

B. The phrasing/wording of the questions

C. The statistical test used

D. The research design

9. “Known groups method” is used to establish:

A. Content validity

B. Construct validity

C. Test-retest reliability

D. Internal consistency

10. What is a key indicator of validity for a set of Multiple-Choice Questions (MCQ) used in a test?

A. Participants finish the test quickly.

B. All questions have similar difficulty levels.

C. The MCQs measure the specific learning outcomes intended.

D. The test has a high Cronbach’s alpha value.

4.6 DIFFERENCE BETWEEN VALIDITY AND RELIABILITY

Zikmund has illustrated the difference between reliability and validity by an example of an old and a modern rifle. The shots by a marksman from the old rifle (target A) are considerably scattered but from the new rifle (target B) are closely clustered, showing thereby that the old rifle is less reliable. In target C, shots with the modern rifle may be reliable but if his vision is not proper, the marksman may not be able to hit the bull’s eye.

4.7 LET US SUM UP

So, we can say that research in social sciences involves studying behaviors. Accurately recording what subjects are doing is difficult and research is always in danger of being influenced by the expectations

of the researchers. The concept of validity and reliability are employed to ensure the soundness of consistency of measurement techniques. Validity refers to whether research actually measures what it was intended to measure; reliability refers to whether the research produces consistent results. So, sound measurement must meet the tests of validity and reliability.

4.8 GLOSSARY

- **Reliability** is the ability of a scale to measure something without any bias.
- **An instrument (questionnaire)** is said to be valid if it produces some generally acceptable results from a given study.
- **A good research instrument** is that which possess sufficient reliability as well as Validity

4.9 SELF ASSESSMENT QUESTIONS

1. What do you mean by Reliability and Validity.

2. What is the important role played by Reliability and Validity.

3. Differentiate between Reliability and Validity.

4.10 LESSON END EXERCISE

1. Tailor uses inch-tape to measure the length of suit is an example of
2. The degree of agreement between various items of an instrument is known as
 - a) Content Validity
 - b) Internal consistency

c) Face Validity d) None of these

3. Reliability refers to consistency in the results whereas Validity refers to measuring what is meant to be measured by a scale (True/ False)

4.11 SUGGESTED READINGS

1. Argyrous, George. 1997. *Statistics for Social Research*. New York: Mc Millan Press Ltd.
2. Gupta, S.C. 1981. *Fundamentals of Statistics*. Bombay: Himalayan Publishing House.
4. Gupta, S.P. 2004. *Statistical Methods*. New Delhi: Sultan Chand and Sons.

4.12 ANSWERS TO CHECK YOUR PROGRESS

Answers: 1- b 2- d 3- c 4-b 5-a 6- c 7- b 8- b 9- b 10-c

MEASURES OF DISPERSION**STRUCTURE**

- 5.0 Objectives
- 5.1 Introduction
- 5.2 Types of Measures of Dispersion
 - 5.2.1 Range
 - 5.2.2 Interquartile Range and Quartile Deviation
 - 5.2.3 Mean Deviation
- 5.3 Let us sum up
- 5.4 Glossary
- 5.5 Self- Assessment Questions
- 5.6 Lesson End Exercise
- 5.7 Suggested Readings
- 5.8 Answers to check your progress

5.0 OBJECTIVES

Dear learner, after successful completion of this lesson, you should be able to:

- understand the concept of dispersion,
- know the significance of measuring dispersion,
- know the different types of measures of dispersions,

- distinguish between absolute and relative measures of dispersion,
- know the computational procedure of these measures, and
- understand the various importance of these measures.

5.1 INTRODUCTION

In the previous lessons, we have studied the various measures of central tendency that are used to provide a single representative value of a given set of data. This value tells us where the center of the set of data lies but does not tell us how the data is scattered around this central value. Two sets of data may have the same average but the items in one set may scatter widely around its average while in the other case, items may be close to the average. In this way the central value alone cannot describe the distribution adequately. A further description about the scatteredness is necessary to get a better description of data. The extent or degree to which data tend to spread around an average is called the dispersion or variation. Measures of dispersion help us in studying the extent to which observations are scattered around the central value. Such measures are helpful in comparing two or more series with regard to their variability.

A good measure of dispersion should possess, as far as possible, the same properties as those of a good measure of central tendency.

5.2 TYPES OF MEASURES OF DISPERSION

Following is some of the well-known measures of dispersion which provide a numerical index of the variability of the given data:

- (i) Range (ii) Semi-Interquartile Range
- (iii) Mean Deviation (iv) Standard Deviation

Further, measures of dispersion are of two types, namely-Absolute measure of dispersion and Relative Measure of dispersion.

Absolute measures of dispersion are expressed in the same unit in which the observations are given. These measures are useful for comparing two or more sets of data where units of measurements are the same. Relative measures of dispersion are expressed as ratio or percentage or the coefficient of the absolute measures of dispersion. These measures are pure number, free from unit of measurement and are generally called coefficient of dispersion. Relative measures are useful for comparing variability in two or more sets of data where units of measurement may be different. Now we shall discuss various measures.

5.2.1 Range

Range is the simplest measure of dispersion. It is defined as the difference between the highest value and lowest value of the data. In symbols, Range is given by

$$\text{Range} = L - S$$

where L = largest value, and S = smallest value.

In case of grouped data, the range is defined as the difference between the upper limit of the highest class and the lower limit of the smallest class.

Range is the absolute measure of dispersion which is not suitable for comparison. To overcome this difficulty a relative measure of dispersion, called the coefficient of Range is calculated by the following formula

$$\text{Coefficient of Range} = \frac{L - S}{L + S}$$

Range is not suitable measure of dispersion because it is not based on all the observations. It is also affected by extreme observations. It does not tell us about the variation in the observations relative to the average. Despite various limitations, ranges us a quick and simple measure and it is useful in quality control for analyzing the variations in the quality of the product. It is very useful in weather forecasting.

5.2.2 The Inter-quartile Range or Quartile Deviation :

The quartile deviation, also known as semi-interquartile range, is computed by taking the average of the difference between third quartile (Q_3) and the first quartile

(Q_1). In symbols, this can be written as

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2}$$

As quartile deviation is an absolute measure of dispersion, thus its relative measure of dispersion called coefficient of quartile deviation, is defined as

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

It is also not suitable because it is not based on all the items and it is very much affected by sampling fluctuations. However, it is a good measure of dispersion. In case of open-end frequency distribution.

Example 1. For the following data, compute quartile deviation and its coefficient.

Weekly Wages (in Rs.)	No. of Workers
Below 850	12
850-900	16
900-950	39
950-1000	56
1000-1050	62
1050-1100	75
1100-1150	30
1150 and Above	10

Solution. To compute quartile deviation, we first need the values of Q_1 and Q_3 which can be obtained from the following table :

Weekly Wages (in Rs.)	No. of Workers f	$c.f.$
Below 850	12	12
850-900	16	28
900-950	39	67
950-1000	56	123
1000-1050	62	185
1050-1100	75	260
1100-1150	30	290
1150 and Above	10	300
N = 300		

Since $\frac{N}{4} = \frac{300}{4} = 75$ which falls in the class 950-1000, thus Q_1 is given by

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

$$\begin{aligned}
&= 950 + \frac{75-67}{56} \times 50 \\
&= 950 + \frac{50}{7} = 950+7.14 \\
&= 957.14
\end{aligned}$$

Similarly Q_3 is

$$\begin{aligned}
Q_3 &= 1050 + \frac{225-185}{75} \times 50 \\
&= 1050 + \frac{2000}{75} = 1050+26.67 \\
&= 1076.67
\end{aligned}$$

$$\begin{aligned}
\text{Thus Q.D.} &= \frac{Q_3 - Q_1}{2} = \frac{1076.76 - 957.14}{2} \\
&= \frac{119.53}{2} = 59.765
\end{aligned}$$

and Coefficient of Q.D. is

$$\begin{aligned}
\text{Coefficient of Q.D.} &= \frac{Q_3 - Q_1}{Q_3 + Q_1} \\
&= \frac{1076.67 - 957.14}{1076.67 + 957.14} \\
&= \frac{119.53}{2033.81} \\
&= 0.0587 \\
&= 5.87\%
\end{aligned}$$

The quartile deviation is superior to range as it is based on two extreme values i.e. it is based on middle 50% of the observations.

5.2.3 Mean Deviation :

As discussed, the range and quartile deviation are not based on all the observations. More

so, these are not measures of dispersion in strict sense of the term as they do not measure scatteredness in observations around an average. The mean deviation is an improvement as it considers all the observations. Mean deviation is defined as the arithmetic mean of the absolute deviations of observations taken from some central value, such as mean, median, mode. Here all the deviations are treated as positive regardless of sign. In symbols, the mean deviation about mean, median or mode can be expressed as follows :

Mean Deviation from mean ($\bar{X}$)

$$\text{M.D. } (\bar{X}) = \frac{1}{N} \sum |X - \bar{X}|$$

Mean Deviation from median (M_e) :

$$\text{M.D. } (M_e) = \frac{1}{N} \sum |X - M_e|$$

Mean Deviation from Mode (M_0) :

$$\text{M.D. } (M_0) : = \frac{1}{N} \sum |X - M_0|$$

where N is the number of observations.

In case of grouped data, the above formulae can be expressed as :

$$\text{M.D. } (\bar{X}) = \frac{1}{N} \sum f |X - \bar{X}|$$

$$\text{M.D. } (M_e) = \frac{1}{N} \sum f |X - M_e|$$

$$\text{M.D. } (M_0) = \frac{1}{N} \sum f |X - M_0|$$

where N be the sum of frequencies.

The relative measure corresponding to the mean deviation, called the coefficient of Mean Deviation, is obtained by dividing the mean deviation by the particular average used in computing the mean deviation. Thus, if mean deviation has been computed from mean, the coefficient of mean deviation would be :

$$\text{Coefficient of Mean Deviation} = \frac{\text{M.D.}(\bar{X})}{\bar{X}}$$

Although the mean deviation is a good measure of dispersion, its use is limited.

Example 2. Consider the following data which are relate to the sales of acompany for 100 days :

Sales (Rs. thousand)	No. of days
40-50	10
50-60	15
60-70	25
70-80	30
80-90	12
90-100	8

Compute mean deviation from mean, median and mode. Also obtain their coefficient of mean deviation.

Solution : To compute mean deviation from mean,median and mode, first we compute mean, median and mode. We construct the following table :

Sales (Rs. Thousand)	No. of days f	Mid-Point x	fx		$ X - \bar{X} $	$f X - \bar{X} $	$ X - M_e $
40-50	5	45	225	5	26	130	26.67
50-60	15	55	825	20	16	240	16.67
60-70	25	65	1625	45	6	150	6.67
70-80	30	75	2250	75	4	120	3.33
80-90	20	85	1700	95	14	280	13.33
90-100	5	95	475	100	24	120	23.33
Total	100		$\Sigma fx =$ 7100			$\Sigma f X - \bar{X} =$ 1040	

$f X - M_e $	$ X - M_0 $	$f X - M_0 $
133.35	28.33	141.65
250.05	18.33	274.95
166.75	8.33	208.25
66.60	1.67	50.10
66.65	11.67	233.40
116.65	21.67	108.35
$\Sigma f X - M_e $ =800.05		$\Sigma f X - M_0 $ =1016.70

$$\text{Here } \bar{X} = \frac{\Sigma fX}{N} = \frac{7100}{100} = 71.00$$

$$M_e = L + \frac{\frac{N}{2} - c.f.}{f} \times i$$

$$= 70 + \frac{\frac{100}{2} - 45}{30} \times 10$$

$$= 70 + \frac{50 - 45}{3}$$

$$= 70 + \frac{5}{3} = 71.67$$

and

$$M_0 = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

$$= 70 + \frac{30 - 25}{2 \times 30 - 25 - 20} \times 10$$

$$= 70 + \frac{50}{15} = 73.33$$

(i) The mean deviation from mean is

$$\text{M.D. } (\bar{X}) = \frac{\Sigma f|X - \bar{X}|}{N} = \frac{1040}{100} = 10.40$$

and coefficient of mean deviation is

$$\text{coefficient of M.D. } (\bar{X}) = \frac{\text{M.D. } (\bar{X})}{\bar{X}}$$

$$= \frac{10.40}{71.00} = 0.146478$$

$$= 14.65\%$$

(ii) The mean deviation from median

$$\begin{aligned}\text{M.D. } (M_e) &= \frac{\sum f |X - M_e|}{N} = \frac{800.05}{100} \\ &= 8.0005 \simeq 8.00\end{aligned}$$

and its coefficient is

$$\begin{aligned}\text{coefficient of M.D. } (M_e) &= \frac{\text{M.D. } (M_e)}{M_e} \\ &= \frac{8.00}{71.67} = 0.11162 \\ &= 11.16\%\end{aligned}$$

(iii) The mean deviation of mode

$$\begin{aligned}\text{M.D. } (M_0) &= \frac{\sum f |X - M_0|}{N} = \frac{1016.70}{100} \\ &= 10.17\end{aligned}$$

and

$$\begin{aligned}\text{Coeff. of M. D. } (M_0) &= \frac{\text{M.D.}(M_0)}{M_0} \\ &= \frac{10.17}{73.33} = 0.1387 \\ &= 13.87\%\end{aligned}$$

Note: You may observe here that the mean deviation from median is 8.00 which minimum as compared to other mean deviations. It has also been discussed while discussing median.

CHECK YOUR PROGRESS-I

Fill in the blanks:

1.is also known as the semi-interquartile range.
2. The formula for quartile deviation is $\frac{Q3-Q1}{2}$
3. Mean deviation is calculated based on the absolute deviations from the.....
4. Quartile deviation is based on the of the data.
5. If $Q1=20$ and $Q3=60$, then the quartile deviation is.....
6. The sum of deviations of individual values from their mean (ignoring signs) is used to calculate.....

5.3 LET US SUM UP

The average does not enable us to draw a full picture of a set of observations. Two sets of observations may have the same averages but the observations in one may scatter wildly around this average while in the other case, all the observations may be close to this average. Thus, the measure of scatteredness of observation around their average is necessary to get a better description of data. The extent or degree to which data tend to spread around an average is called dispersion or variation. Measures of dispersion may be absolute or relative. Absolute measures of dispersion are expressed in the unit of given observations. Such measures are useful for comparing variations in two or more distributions in which the units of measurement are the same. On the other hand, relative measures of dispersion, also called co-efficient of dispersion, are pure unitless numbers useful for comparing the variability in two or more distributions in which units of measurements are different. We study four important absolute measures of dispersion, namely—Range, Interquartile range and Quartile deviation, mean deviation and standard deviation. The Range is defined as the difference between two extreme observations. Interquartile range is the difference between the third and the first quartile. Quartile deviation gives the average amount by which the two quartiles differ from the median. Range, interquartile range and quartile deviation are not measures of dispersion in the strict sense of the term as they do not measure scatteredness in observations around an average and more so, their computation is not based on all the observations. Mean-deviation is defined as the arithmetic mean of the absolute deviations of various items from an average value, such as mean, median or mode. In the computation of mean-deviation, we ignore signs of deviations and consider absolute values only. The standard-deviation is defined as the positive square-root of the arithmetic mean of the squares of deviations of observations from the arithmetic mean. This is also known as ‘root mean square deviation.’ The square of standard-deviation is known as variance. The standard deviation is expressed in the unit of observations in the series while variance is measured in square units. The standard deviation is the best measure of dispersion as it satisfies most of the desirable properties.

After defining absolute measures of dispersion, we face with the difficulty of comparing variability in two or more distributions in which observations are expressed in different units or the means of the distributions are widely different. Thus, relative measure of dispersion is defined to overcome such situation. Co-efficient of Variation (C.V.) is the best measure of relative dispersion to deal with such situations. C.V. is useful for comparing distributions in respect of their variability, homogeneity, uniformity or consistency. The distribution having greater C.V. is considered more variable than the other, and the distribution with lesser C.V. shows greater consistency, homogeneity and uniformity.

5.4 GLOSSARY

- **Dispersion:** It is the variations in all data when we compare it with its mean value so it can also be called Spread in data
- **Range:** A simplest measure of dispersion
- **Mean Deviation:** It measures the average deviation of data from its central value.
- **Standard deviation** overcomes the limitations of mean deviation i.e. effect of negative items.

5.5 SELF ASSESSMENT QUESTIONS

1. What do you understand by dispersion? What is the need of studying dispersion.

2. What is meant by absolute and relative measure of dispersion.

3. Explain briefly the essentials of a good measure of variation.

4. What do you mean by dispersion? Enlist the important measures of dispersion.

5. Define range and quartile deviation. Also write their merits and demerits as a measure of dispersion.

6. Define mean deviation with its merits and demerits.

7. Define standard deviation. Explain its uses.

8. State the properties of standard deviation. Why is it called the best measure of dispersion.

9. What is meant by relative dispersion? When are they used? How are they measured.

10. Define variance and co-efficient of variance.

11. Discuss some important algebraic properties of S.D.

12. In what way measures of variation supplement measures of central tendency? Explain.

13. What is co-efficient of variation? What purpose does it serve.

14. Distinguish between variance and co-efficient of variation.

15. From the following data, find range and quartile deviation. Also determine their co-efficient.

Months	:	1	2	3	4	5	6	7	8	9	10	11	12
Sales (Rs.)	:	78	80	80	82	82	84	84	86	86	88	88	90

16. Calculate Q.D. and its co-efficient from the following data:

Weight (Kg.)	:	60	61	62	63	65	70	75	80
No. or workers	:	1	3	5	7	10	3	1	1

17. Calculate range and Q.D. for the following data. Also find their co-efficient.

Classes	:	0–5	5–10	10–15	15–20	20–25	25–30	30–35	35–40
Frequency	:	4	5	6	10	11	9	4	1

18. You are given two variables A and B. Using Q.D., state which is more variable.

Mid points	:	15	20	25	30	35	40	45	
Frequency	:	15	33	56	103	40	32	10	
Mid points	:	100	150	200	250	300	350	400	450
Frequency	:	340	492	890	1420	620	360	187	140

19. Calculate mean deviation from median of the prices given below:

Prices (Rs.)	210	220	225	225	225	235	240	250	270	280
---------------------	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

Also find co-efficient of mean deviation.

20. From the table given below, find mean deviation from the median and also its coefficient.

Size	:	0–10	10–20	20–30	30–40	40–50	50–60	60–70
Frequency	:	4	8	11	15	11	7	4

21. Find S.D. and C.V. from the following data:

Marks:	0–10	10–20	20–30	30–40	40–50	50–60	60–70
No. of Student:	10	15	25	25	10	10	5

22. Find mean and S.D. from the following data. Also find C.V.

Age (Less than)	:	10	20	30	40	50	60	70	80
No. of persons	:	15	30	53	75	100	110	115	125

[Hint. Change the data into a simple frequency distribution]

5.6 LESSON END EXERCISE

- The formula for computing Range is
- deviation is not affected by the negative values in data.
- is the statistical tool that can be used to compare the degree of deviations in two different data sets.

5.7 SUGGESTED READINGS

1. Argyrous, George. 1997. *Statistics for Social Research*. New York: Mc Millan Press Ltd.
2. Goods, W.J. & Hatt, P.K. 1981. *Methods in Social Research*. New York: Mc Graw Hill.
3. Gupta, S.C. 1981. *Fundamentals of Statistics*. Bombay: Himalayan Publishing House.
4. Gupta, S.P. 2004. *Statistical Methods*. New Delhi: Sultan Chand and Sons.

5.8 ANSWERS TO CHECK YOUR PROGRESS

Answers: 1. Quartile Deviation

Answers: 2. $Q_3 - Q_1/2$

Answers: 3. mean (or median)

Answers: 4. middle 50%

Answers: 5. 20

Answers: 6. mean deviation

STANDARD DEVIATION AND COEFFICIENT OF VARIATION**STRUCTURE**

- 6.0 Objectives
- 6.1 Introduction
- 6.2 Standard Deviation
- 6.3 Coefficient of Variation
- 6.4 Let us sum up
- 6.5 Glossary
- 6.6 Self-Assessment Questions
- 6.7 Lesson End Exercise
- 6.8 Suggested Readings
- 6.9 Answers to check your progress

6.0 OBJECTIVES

Dear learner, the main objectives of this lesson are:

- To acquaint you with the knowledge of standard deviation;
- To give you information about the coefficient of correlation;
- To enhance your understanding about the Karl Pearson's and Spearman's coefficient of rank Correlation.

6.1 INTRODUCTION

Standard deviation is a statistical measure indicating how spread-out data points are around the

mean (average). A low standard deviation means data is tightly clustered near the mean, while a high standard deviation indicates data is spread over a wider range. It is essential for determining data variability, risk in investments, and identifying outliers.

6.2 STANDARD DEVIATION

Standard deviation is the most popular and important measure of dispersion. It satisfies most of the properties of a good measure of dispersion. The standard deviation is defined as the positive square root of the arithmetic mean of the squares of deviations of the observations taken from the mean. It is also known as “Root mean square

Deviation” and is generally denoted by the small Greek letter σ (called sigma). Symbolically, for ungrouped data or individual series, the standard deviation (σ) is defined as

$$\text{S.D.} = s = \sqrt{\frac{1}{N} \sum (X - \bar{X})^2}$$

In case of frequency distribution, the formula becomes

$$s = \sqrt{\frac{1}{N} \sum f (X - \bar{X})^2}$$

The square of the standard deviation is known as variance. It is denoted by s^2 and given by

$$s^2 = \sqrt{\frac{1}{N} \sum (X - \bar{X})^2}, \text{ for individual series}$$

$$\sqrt{\frac{1}{N} \sum f (X - \bar{X})^2}, \text{ for continuous series.}$$

Remarks:

1. The Standard Deviation is an absolute measure of dispersion or variability in a distribution. The greater the amount of dispersion or variability, the greater the standard deviation. On the other hand, a smaller standard deviation means a higher degree of uniformity of the observations.
2. Standard deviations of two or more distributions with nearly identical means may be compared in respect of variability in observation around the mean. The distribution with the smallest standard deviation has the most representative mean.
3. The Standard Deviation is expressed in the unit of the observations of the series while the variance is measured in square unit. For example, if the observations are measured in inches, the Standard Deviation will be in inches while the variance will be in sq. inches.

Computation of Standard Deviation (S.D.) (I) Individual Series or Ungrouped data

For ungrouped data, the following two methods are used for computing standard deviation—

1. Direct Method.
2. Short-cut Method.

Direct Method:

In this procedure, the following formula is used for calculating S.D. (0)—

$$\sigma = \sqrt{\left[\frac{1}{n} \sum (x - \bar{x})^2 \right]} \quad \dots(1)$$

or

$$\sigma = \sqrt{\left[\frac{\sum x^2}{n} - \left(\frac{\sum x}{n} \right)^2 \right]} = \sqrt{\left[\frac{\sum x^2}{n} - \bar{x}^2 \right]} \quad \dots(2)$$

Remarks:

1. Formula (1) is taken to be appropriate only when the mean is a whole number.
2. Formula (2) does not involve any deviations and thus, should be recommended when the observations

are not too large.

The following examples will clarify their use.

Example 3. Calculate S.D. from the following set of observations:

$x :$ 10 11 17 25 7 13 21 10 12 14

Computing S.D.

x	$(x - \bar{x})$	$(x - \bar{x})^2$
10	-4	16
11	-3	9
17	3	9
25	11	121
7	-7	49
13	-1	1
21	7	49
10	-4	16
12	-2	4
14	0	0
$\Sigma x = 140$		$\Sigma (x - \bar{x})^2 = 274$

Solution :

$$\text{Here } \bar{x} = \frac{\Sigma x}{n} = \frac{140}{10} = 14$$

Using formula (1)

$$\sigma = \sqrt{\left[\frac{1}{n} \Sigma (x - \bar{x})^2 \right]}$$

$$= \sqrt{\left[\frac{1}{10} \times 274 \right]}$$

$$= \sqrt{27.4} = 5.23$$

Example 4. Use formula (2) to calculate S.D. for the data in example 3.

Calculation of standard deviation

x	x^2
10	100
11	121
17	289
25	625
7	49
13	169
21	441
10	100
12	144
14	196
$\Sigma x=140$	$\Sigma x^2=2234$

$$\sigma = \sqrt{\left[\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n} \right)^2 \right]}$$

$$= \sqrt{\left[\frac{2234}{10} - \left(\frac{140}{10} \right)^2 \right]}$$

$$= \sqrt{[223.4-196]}$$

$$= \sqrt{[27.4]}=\mathbf{5.23}$$

Which is the same as that obtained in example 3.

Short-cut-Method

In most of the cases the arithmetic mean of the given distribution happens to be a fractional value and then the process of taking deviation and equating them becomes quite tedious and time consuming in the computation of S.D. To overcome this difficulty, short-cut method of computation is used which involves deviations from assumed mean. The short-cut formula for calculating S.D. is—

$$\text{S.D.} = \sigma = \dots\dots\dots (3)$$

Here, d = deviation from assumed mean, say A ; i.e., $d = (x-A)$

$\sum d$ = the sum of deviations

$\sum d^2$ = the sum of squares of deviations

n = the number of observations.

Example 5. Find S.D. from the data in example 3.

Computation of S.D.

X	$d=x-A$	d^2
10	—2	4
11	—1	1
17	+5	25
25	+13	169
7	—5	25
13	+1	1
21	+9	81
10	—2	4
12	0	0
14	2	4
$n = 10$	$\sum d=20$	$\sum d^2=314$

Solution :

Solution :

Let $A = 12$

Using formula (3) i.e.,

$$\sigma = \sqrt{\left[\frac{\sum d^2}{n} - \left(\frac{\sum d}{n} \right)^2 \right]}$$

$$= \sqrt{\left[\frac{314}{10} - \left(\frac{20}{10} \right)^2 \right]}$$

$$= \sqrt{(31.4-4)} = \sqrt{(27.4)}$$

$$= 5.23$$

(II) Computation of S.D. in Grouped data

Any of the following three procedures may be applied to find S.D. for grouped data.

1. Direct Method.
2. Short-Cut Method.
3. Step-Deviation Method.

Direct Method:

In direct method of computing S.D., the following formula is used—

$$\text{S.D.} = \sigma = \sqrt{\left[\frac{1}{N} \sum f (x - \bar{x})^2 \right]} \quad \dots(4)$$

where symbols have their meaning.

Example 6. Use direct-method to calculate the S.D. of the following discrete frequency distribution.

Size (x)	:	4	5	6	7	8	9	10
Frequency	:	6	12	15	28	20	14	5

Solution.

Computation of S.D. (Direct-method)

Size X	Frequency F	fx	$(x - \bar{x})$	$(x - \bar{x})^2$	$f(x - \bar{x})^2$
4	6	24	-3.06	9.3636	56.1816
5	12	60	-2.06	4.2436	50.9232
6	15	90	-1.06	1.1236	16.8540
7	28	196	-0.06	0.0036	0.1008
8	20	160	+0.94	0.8836	17.6720
9	14	126	+1.94	3.7636	52.6904
10	5	50	+2.94	8.6436	43.2180
$N = 100$		$\sum fx = 706$	$\sum f(x - \bar{x})^2 = 237.6400$		

Here, we first calculate —

$$\bar{x} = \frac{\sum fx}{N} = \frac{706}{100} = 7.06$$

Using formula 4

$$\text{S.D.} = \sigma = \sqrt{\left[\frac{1}{N} \sum f(x - \bar{x})^2 \right]} = \sqrt{\left(\frac{237.64}{100} \right)} = \sqrt{(2.3764)} = 1.54.$$

Remark: This method involves deviations from actual mean x . Thus, the computation procedure becomes difficult when mean is not a whole number. However, in practical the procedure is rarely used as arithmetic mean is generally a fractional value.

Short-Cut-Method:

In the direct method of computing S.D., we face a difficult situation when the arithmetic mean is not a whole number. To avoid such difficulties, we consider the deviations observations from a suitably chosen assumed mean, say A . In short-cut method of comping the S.D. we make use of the following formula—

$$\text{S.D.} = \sigma = \sqrt{\left\{ \left[\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N} \right)^2 \right] \right\}} \quad \text{.....(5)}$$

Where d is deviation from assumed mean. The short-cut method of calculating S.D. involves the following steps—

Example 7. Use short-cut method to find the S.D. of the data in example 6.

Solution.

Computation of S.D. (Short-cut method)

x	F	$d = (x-7)$	fd	d^2	fd^2
4	6	-3	-18	9	54
5	12	-2	-24	4	48
6	15	-1	-15	1	15
7	28	0	0	0	0
8	20	1	+20	1	20

9	14	2	+28	4	56
10	5	3	+15	9	45
$N = 100$			$\Sigma fd = 6$	$\Sigma fd^2 = 238$	

Using formula 5,

$$\sigma = \sqrt{\left\{ \left[\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N} \right)^2 \right] \right\}} = \sqrt{\left[\frac{238}{100} - \left(\frac{6}{100} \right)^2 \right]}$$

$$= \sqrt{[2.38 - 0.0036]} = \sqrt{2.3764} = 1.54.$$

Step-Deviation Method.

In short-cut method of calculation our main aim is to simplify the deviations so that computations become easier. In grouped data, especially in continuous frequency distributions, we observe that the calculations can be further simplified if deviations (d) are divided by a common factor, say h , which is usually the size of the class- interval to get the step-deviations as d' . However, such a division is then reflected in the formula for computing S.D. which now becomes —

$$\sigma = h \sqrt{\left\{ \left[\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N} \right)^2 \right] \right\}} \quad \text{.....(6)}$$

Steps involved in computing S.D. by using step-deviation method will be clear from the following example.

Example 8. Use short-cut and step-deviation method for calculating S.D. of the following distribution.

Age groups (years) : 25–30 30–35 35–40 40–45 45–50 50–55

No. of workers: 10 12 25 40 10 3

Solution.

Using Short-cut method.

Computation of S.D. (Short-cut method)

Age	Frequency F	Mid-value	$d = (x - 42.5)$	fd^2	d^2	fd^2
25–30	10	27.5	–15	–150	225	2250
30–35	12	32.5	–10	–120	100	1200

35-40	25	37.5	-5	-125	25	625
40-45	40	42.5	0	0	0	0
45-50	10	47.5	5	50	25	250
50-55	3	52.5	10	30	100	300
N=100			$\Sigma fd = -315$		$\Sigma fd^2 = 4625$	

The assumed mean is taken as $A = 42.5$. Using formula (22).

$$= \sqrt{\left[\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N} \right)^2 \right]} = \sqrt{\left[\frac{4625}{100} - \left(\frac{-315}{100} \right)^2 \right]}$$

$$= \sqrt{[46.25 - 9.9225]} = \sqrt{[36.3275]} = 6.03 \text{ years.}$$

Using Step-derivation Method :

Computation of S.D. (Step-deviation Method)

Age	Frequency F	Mid-value x	$d' = \frac{x - 42.5}{5}$	fd'	fd' ²
25-30	10	27.5	-3	-30	90
30-35	12	32.5	-2	-24	48
35-40	25	37.5	-1	-25	25
40-45	40	42.5	0	0	0
45-50	10	47.5	1	10	10
50-55	3	52.5	2	6	12
N = 100				$\Sigma fd' = 63$	$\Sigma fd'^2 = 185$

Here, $A = 42.5$ and $h = 5$. Using formula (6), i.e.,

$$\sigma = h \sqrt{\left[\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N} \right)^2 \right]}$$

Putting the values from the table,

$$\begin{aligned}\sigma &= 5\sqrt{\left[\frac{185}{100} - \left(\frac{63}{100}\right)^2\right]} = 5\sqrt{[1.85 - 0.3969]} \\ &= 5\sqrt{(1.4531)} = 5 \times 1.2054 = 6.03 \text{ years.}\end{aligned}$$

Which is the same as that obtained by short-cut method.

Remark: In grouped data case, we have demonstrated various methods of computing the S.D. It is easy to see that **step-deviation method** is, in general, best for calculating S.D. of the grouped data.

Merits and Demerits of Standard Deviation:

Merits.

1. It is rigidly defined.
2. Its computation is based on all the observations.
3. It is amenable to further algebraic treatment which makes it the most important and widely used measure of dispersion. For example, the S.D. is used in computing skewness, correlation etc. It is an important statistical measure in sampling theory.
4. Among all the measures of dispersion, it is least affected by sampling fluctuations.
5. S.D. enables us to determine the reliability of means of two or more series having equal means. In such a situation, a series having minimum S.D. will have the most representative mean. In other words, a smaller value of S.D. reflects greater compactness and smaller variability among items.

Demerits:

1. S.D. is comparatively difficult to calculate.
2. It gives greater weight to extreme observations.
3. It is an absolute measure of dispersion and cannot be used for comparing variability of two or more distributions expressed in different units.

CHECK YOUR PROGRESS-I

FILL IN THE BLANKS WITH SUITABLE ANSWERS:

1. Standard deviation is the of the
2. The standard deviation is of any change in origin but is on change of scale.
3. The standard deviation is the mean deviation taken from the arithmetic mean.
4. Standard deviation is also called the
5. If all data values are the same, the standard deviation is
6. The sum of the squares of the deviations of the values is when taken about their arithmetic mean

6.3 CO-EFFICIENT OF VARIATION (C.V.)

It is now clear that the standard deviation, as a measures of dispersion, gives us an idea about the extent to which observations are scattered around their mean. Thus, two or more distributions having the same mean can be compared directly for their variability with the help of corresponding standard deviations. Now the following two situations may arise: -

- (a) When two or more distributions having unequal means are to be compared in respect of their variability.
- (b) When two or more distributions having observations expressed in different units of measurements are to be compared in respect of their scatteredness or variability.

For making comparisons in the above two situations, we use a relative measure of dispersion, called **co-efficient of variation (C.V.)**. The **co-efficient of variation (C.V.)** is defined as —

$$\text{C.V.} = \frac{\sigma}{x} \times 100$$

Remarks:

1. Co-efficient of variation is a pure number independent of the units of measurements.
2. This is useful for making comparisons between two or more distributions in respect of their **variability, homogeneity, uniformity or consistency**.
3. The distribution having greater C. V. is considered more variable than the other and the distribution with lesser C.V. shows greater consistency, homogeneity and uniformity.

The following example will clarify the use of co-efficient of variation.

Example 9. A sample of 5 items was taken from the output of a factory.

The length and weight of 5 items are given below:

Length (Inches) : 5 6 7 9 12

Weight (Ounces) : 13 15 18 19 20

State which of the two characteristics of the two items is more variable.

Solution. The two characteristics of the items, are length and weight expressed in inches and ounces respectively. Since the units of the two characteristics are different, their variability can be compared with the help of co-efficient of variation which is computed as under —

Computation of co-efficient of variation

Length (Inches)			Weight (ounces)		
x	$d_x = (x-7)$	dx^2	dy	$d_y = (y-18)$	dy^2
5	-2	4	13	-5	25
6	-1	1	15	-3	9
7	0	0	18	0	0
9	2	4	19	1	1
12	5	25	20	2	4
$n = 5$	$\sum dx = 4$	$\sum dx^2 = 34$	$n = 5$	$\sum dy = -5$	$\sum dy^2 = 39$

For Length (x-Series)

$$\begin{aligned} \text{Mean} &= \bar{x} = A + \frac{\sum d_x}{n} \\ &= 7 + \frac{4}{5} = 7.8 \text{ inches.} \end{aligned}$$

$$\text{S.D.} = \sigma_x = \sqrt{\left[\frac{\sum d_x^2}{n} - \left(\frac{\sum d_x}{n} \right)^2 \right]}$$

$$\begin{aligned}
&= \sqrt{\left[\frac{34}{5} - \left(\frac{4}{5} \right)^2 \right]} \\
&= \sqrt{[6.8 - 0.64]} = \sqrt{[6.14]} \\
&= 2.48 \text{ inches.}
\end{aligned}$$

For Weight (y-Series)

$$\begin{aligned}
\text{Mean} = \bar{y} &= A + \frac{\sum d_y}{n} \\
&= 18 + \left(\frac{-5}{5} \right) = 17.0 \text{ ounces.}
\end{aligned}$$

$$\begin{aligned}
\text{S.D.} = \sigma_y &= \sqrt{\left[\frac{\sum d_y^2}{n} - \left(\frac{\sum d_y}{n} \right)^2 \right]} \\
&= \sqrt{\left[\frac{39}{5} - \left(\frac{-5}{5} \right)^2 \right]} \\
&= \sqrt{[7.8 - 1]} = \sqrt{(6.8)} \\
&= 2.61 \text{ ounces.}
\end{aligned}$$

Co-efficient of variation

$$\begin{aligned}
\text{C.V.} &= \frac{\sigma_x}{\bar{x}} \times 100 = \frac{2.48}{7.80} \times 100 \\
&= 31.79\%
\end{aligned}$$

Co-efficient of variation

$$\begin{aligned}
\text{C.V.} &= \frac{\sigma_y}{\bar{y}} \times 100 = \frac{2.61}{17.00} \times 100 \\
&= 15.35\%
\end{aligned}$$

on comparing the two C.V.'s we can say that the length characteristic is more variable than weight.

CHECK YOUR PROGRESS-II

- Q1: The coefficient of variation is calculated by dividing the _____ by the _____ and multiplying by 100.
- Q2. If the standard deviation of a dataset is 5 and the mean is 20, the coefficient of variation is _____.
- Q3. Team A has a CV of 60% and Team B has a CV of 45%. Therefore, Team _____ is more consistent.

6.4 LET US SUM UP

Standard deviation measures the dispersion of data points relative to their mean, indicating consistency or volatility. A low standard deviation suggests data is clustered closely around the mean (high consistency), while a high standard deviation indicates data is spread over a wider range (high variability/risk).

The coefficient of variation (CV) is a dimensionless, relative measure of data dispersion, calculated as the standard deviation divided by the mean. A lower CV indicates higher consistency and less variability, making it superior to standard deviation for comparing datasets with different units or significantly different means.

6.5 GLOSSARY

1. **Mean (Average):** The central point around which the standard deviation is calculated.
2. **Variance:** The average of the squared differences from the mean.
3. **Volatility:** In finance, SD measures the risk or fluctuation of a fund's returns, often representing how far annual returns may deviate from the average.
4. **68-95-99.7 Rule:** In a normal distribution, roughly 68% of data falls within 1 SD, 95% within 2 SDs, and 99.7% within 3 SDs of the mean, says a National Library of Medicine article.

6.6 LESSON END EXERCISE

1. **What does a standard deviation of 0 mean?**

2. Can standard deviation be negative?

3. What is the primary use of the Coefficient of Variation?

4. What are the units of a Coefficient of Variation?

6.7 SELF ASSESSMENT QUESTIONS

1. If all observations are multiplied by 3, what happens to the standard deviation?

2. What is the difference between S.D. and Mean Deviation (M.D.)?

3. What is the relationship between CV and Consistency?

6.8 SUGGESTED READINGS

1. Clark, T.C. and E.W. Jordan (1985). Introduction to Business and Economic Statistics, South-Western Publishing Co.
2. Enns, P.G. (1985) Business Statistics.
3. Gupta, S.P. Statistical Methods.

6.9 ANSWERS TO CHECK YOUR PROGRESS

CHECK YOUR PROGRESS I

1. Standard deviation is the **square root** of the **variance**.
2. The standard deviation is **independent** of any change in origin but is **dependent** on change of scale.
3. The standard deviation is **greater than or equal to** the mean deviation taken from the arithmetic mean.
4. Standard deviation is also called the **Root-Mean Square Deviation**.
5. If all data values are the same, the standard deviation is **zero**.
6. The sum of the squares of the deviations of the values is **minimum** when taken about their arithmetic mean

CO-RELATION ANALYSIS**STRUCTURE**

- 7.0 Objectives
- 7.1 Introduction
- 7.2 Meaning of Correlation
- 7.3 Methods of Studying Correlation
- 7.4 Scatter Diagram Method
- 7.5 Karl Pearson's Coefficient of Correlation and its Computation
- 7.6 Spearman's Coefficient of Rank Correlation
- 7.7 Let us sum up
- 7.8 Glossary
- 7.9 Self-Assessment Questions
- 7.10 Lesson End Exercise
- 7.11 Suggested Readings
- 7.12 Answers to check your progress

7.0 OBJECTIVES

After successful completion of this lesson, you will be able to:

- understand the concept of correlation,
- draw a scatter diagram and have an idea about types of correlation,

- compute and interpret correlation, and
- know the limitations of correlation coefficient
- calculate Karl Pearson's Coefficient of Correlation by different methods,
- learn some important properties of Coefficient of Correlation,
- understand the probable error and its application, and
- know about Coefficient of Determination and its use.
- to understand the meaning of ranks and rank correlation.
- to learn the computational procedure of rank correlation.
- to know the merits and demerits of rank correlation, and

7.1 INTRODUCTION

In the previous lessons we have confined our discussions with the distributions of the data involving only one variable. Such a distribution is called univariate distribution. However, in practice, we come across situations where more than one variable is involved. For example, demand and supply of a commodity, volume and temperature of a gas, heights and weight of student in a class etc. In such situations, our aim is to determine whether there exists a relationship between two variables. If such a relationship can be expressed by a mathematical formula, then we shall be able to use it for an analysis of data. Correlation is the method that deals with the analysis of such relationships between two variables.

7.2 MEANING OF CORRELATION

If for every value of a variable, X , we have a corresponding value of another variable Y , the resulting series of pairs of values of two variables is known as bivariate population and its distribution is known as bivariate distribution.

In a bivariate distribution if the change in one variable appears to be accompanied by a change in other variable and vice-versa, then the two variables are said to be correlated and this relationship is called correlation or co-variation. In other words, the tendency of simultaneous variation of the two variables is called correlation. Then correlation studies the degree of inter-dependence between two variables.

- **The correlation is of the following types:**

Positive and Negative Correlation: As a first step, the correlation may be classified according to the direction of change in the two variables. When the increase (or decrease) in one variable results in a

corresponding increase (or decrease) in the other, the correlation is said to be positive. Thus, positive correlation means change in both variables in the same direction. For example, increase in amount spent on advertisement and sales. However, if the two variables deviate in opposite direction, they are said to be negatively correlated. In other words, if the increase (or decrease) in one variable creates a decrease (or increase) in the other variable, then the correlation between two variables is said to be negative.

Further the correlation is perfectly positive if the change in two variables is in the same direction and same ratio. However, it is perfectly negative if the change in two variables is in opposite direction but in same ratio.

7.3 METHODS OF STUDYING CORRELATION

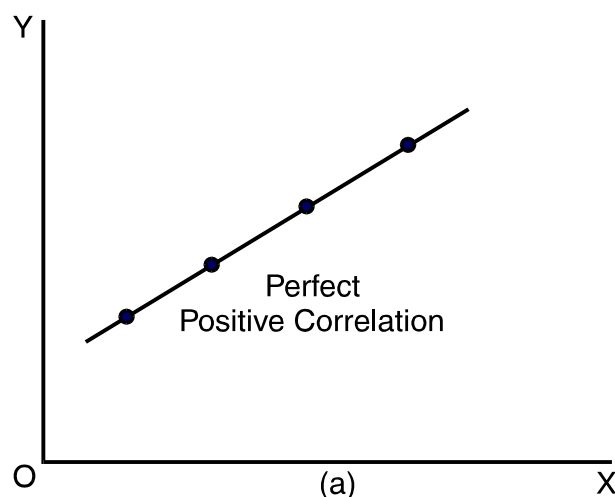
The following methods may be used for studying the correlation between two variables (For ungrouped data):

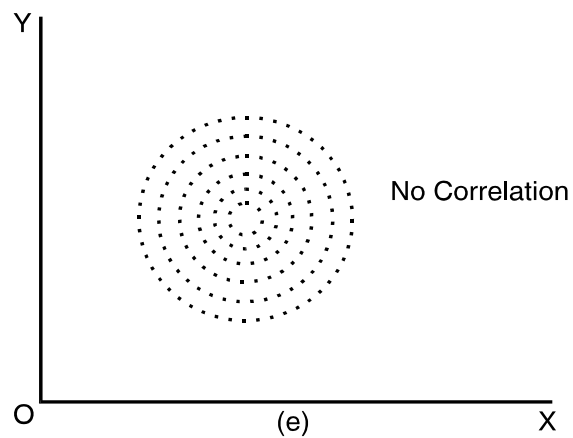
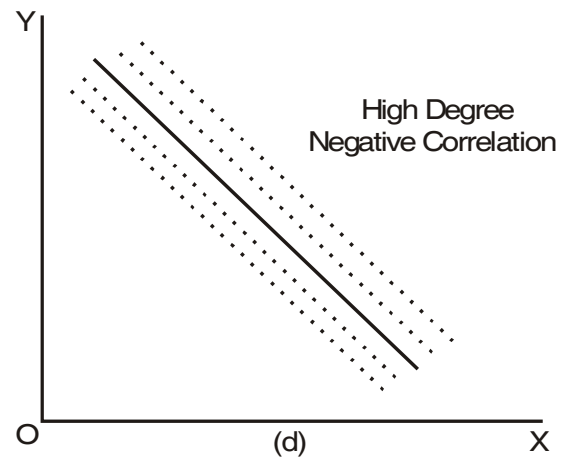
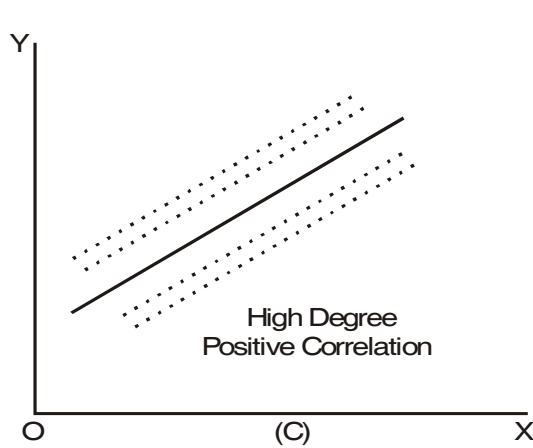
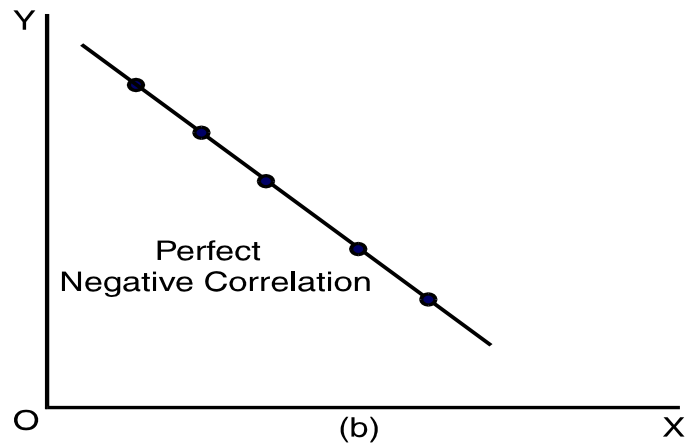
- (i) Scatter Diagram Method (ii) Karl Pearson's Coefficient of Correlation (iii) Spearman's Coefficient of Rank Correlation.

The first two methods are discussed in the present lesson while Spearman's Coefficient of rank correlation will be discussed in Lesson 13.

7.4 SCATTER DIAGRAM METHOD

A graphical representation of a set of pairs of values of two variables X and Y in a coordinate system is called a scatter diagram or simply dot diagram. By means of scatter diagram one can quickly judge the type of correlation between the variables. Scatter diagrams, as an example, showing various degrees of correlation are shown in the given figures





In figure (a) all the dots are lying on a straight line of positive slope, thus we have a perfect positive correlation between two variables and the value of correlation coefficient will be +1. Similarly, in fig. (b) all the dots in the diagram are lying on a straight line of negative slope and this situation shows

perfect negative correlation between two variables. Here its value will be -1 . In fig. (c), the dots lie close to a straight line of positive slope and this shows a high degree positive correlation. Similarly, in fig (d), the dots lie close to a straight line of negative slope which indicates that the negative correlation of high degree exists between two variables. Finally, if the dots do not follow a pattern along with a straight line as in fig. (e), we have no correlation or zero correlation and we may conclude that no linear relationship exists between the variables X and Y. In view of the above discussion, it is clear that the greater the scatter of dots from the straight line on the graph, the lesser the correlation.

This method has the following drawbacks :

- (i) It gives only a rough idea that how the two variables are related.
- (ii) It gives an idea about the direction and also whether is high or low.
- (iii) It does not indicate the degree or extent of relationship existing between the two variables.

In the following section we will discuss Karl Pearson's Coefficient of Correlation which measures the correlation numerically.

7.5 KARL PEARSON'S COEFFICIENT OF CORRELATION

To determine the degree or extent of the linear correlation between two variables, Karl Pearson, defined a numerical measure called Correlation Coefficient denoted by r , and given by

$$r = \frac{\text{Cov}(X, Y)}{\sqrt{V(X)}\sqrt{V(Y)}} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

where $\text{Cov}(X, Y)$ is the covariance between X and Y, and s_X and s_Y are the standard deviations of X and Y respectively. It is also known as Product Moment Correlation Coefficient or simply Coefficient of Correlation.

If (X_i, Y_i) , $i = 1, 2, 3, \dots, n$ be the set of values of size n from a bivariate population of (X, Y) . Let $\bar{X}$ and $\bar{Y}$ be the means of X and Y respectively, then correlation coefficient is given by

$$r = \frac{\frac{1}{n} \sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\frac{1}{n} \sum (X - \bar{X})^2} \sqrt{\frac{1}{n} \sum (Y - \bar{Y})^2}}$$

$$= \frac{\Sigma XY - n\bar{X}\bar{Y}}{\sqrt{\Sigma X^2 - n\bar{X}^2} \sqrt{\Sigma Y^2 - n\bar{Y}^2}} \quad \text{.....(1)}$$

It can also be expressed as

$$r = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{n\Sigma X^2 - (\Sigma X)^2} \sqrt{n\Sigma Y^2 - (\Sigma Y)^2}} \quad \text{.....(2)}$$

The study of the correlation is of immense use in practical life where most of the variables show some kind of relationship. With the help of correlation coefficient, we can measure the degree of relationship existing between variables.

The value of the coefficient of correlation (r) always lies between -1 and +1. Where $r = -1$ or $+1$, the correlation is said to be perfectly negative or positive. An intermediate value of r between -1 and +1 indicates the degree of linear relationship between two variables X and Y whereas its sign talks about the direction of relationship. $r = 0$ means no linear relationship between two variables. However, if covariance is zero then the variables are said to be independent and $r = 0$.

The correlation coefficient is a pure number, independent of the unit of measurement.

Limitations of Correlation Coefficient

The correlation coefficient is a measure of the relationship between two variables, say X and Y. While it generally, serves as a useful statistical tool, we should also be aware of its limitations. The correlation coefficient is a measure of statistical relationship, and not of casual relationship, between the two variables. This means that the value of r tells us whether, and with what regularity, y increases or decreases as X increase. But it cannot tell us whether that increase or decrease is due to any casual or cause-effect relationship between two variables. Further, the correlation coefficient is a measure of linear statistical relationship only, and may fail to be a proper index of statistical relationship in case it is non-linear.

Calculation of Correlation Coefficient

For ungrouped data, Karl Pearson's Coefficient of Correlation can be obtained by using any of the following three methods:

Actual Mean Method: In this method, we make use of the following formula to determine the coefficient of correlation:

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2}} \quad \text{.....(1)}$$

This method is suitable in cases where the mean values $\bar{X}$ and $\bar{Y}$ are not fractions.

Example 1. Calculate the correlation coefficient between the height of father and height of son from the given data:

Height (in inches)

Father: 70, 69, 68, 67, 66, 65, 64

Son : 72, 68, 70, 68, 65, 67, 66

Solution : In this example $\bar{X}$ and $\bar{Y}$ are not fractional values as

$\bar{X} = \frac{469}{7} = 67$ and $\bar{Y} = \frac{476}{7} = 68$, so actual mean method will be suitable.

Height of Father (X)	Height of Son (Y)	$X - \bar{X}$	$(X - \bar{X})^2$	$Y - \bar{Y}$	$(Y - \bar{Y})^2$	$(X - \bar{X})(Y - \bar{Y})$
70	72	3	9	4	16	12
69	68	2	4	0	0	0
68	70	1	1	2	4	2
67	68	0	0	0	0	0
66	65	-1	1	-3	9	3
65	67	-2	4	-1	1	2
64	66	-3	9	-2	4	6
$\Sigma X = 469$	$\Sigma Y = 476$		28		34	25

Thus

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2}}$$

$$= \frac{25}{\sqrt{28 \times 34}}$$

$$= \frac{25}{30.86} = 0.81$$

III Short-cut Method: When mean values of X and Y i.e. $\bar{X}$ and $\bar{Y}$ are in fractions and the values of paired observations is also large, then computation of the coefficient of correlation can be further simplified by using deviations of the observations from some suitable chosen values (called assumed means). The formula for computing correlation coefficient based on deviations is

$$r = \frac{n\sum d_x d_y - (\sum d_x)(\sum d_y)}{\sqrt{n\sum d_x^2 - (\sum d_x)^2} \sqrt{n\sum d_y^2 - (\sum d_y)^2}}$$

where

$d_x = X - A$, the deviations taken from assumed mean A.

$d_y = Y - B$, the deviations taken from assumed mean B.

$\sum d_x$ = Sum of deviations of X.

$\sum d_y$ = Sum of deviations of Y.

$\sum d_x^2$ = Sum of squares of deviations of X.

$\sum d_y^2$ = Sum of squares of deviations of Y.

$\sum d_x d_y$ = Sum of products of the deviations of X and Y.

The computational procedure will be clearer from the following example:

Example 3. Obtain the coefficient of correlation between X and Y from the following data :

X:	47	44	40	38	42	43	45	42	44	40	46	44
Y:	19	26	30	31	29	29	27	27	19	18	19	31

Solution :

X	Y	$d_x = X - 42$	$d_y = Y - 29$	d_x^2	d_y^2	$d_x d_y$
47	19	5	-10	25	100	-50
44	26	2	-3	4	9	-6
40	30	-2	1	4	1	-2

38	31	-4	2	16	4	-8
42	29	0	0	0	0	0
43	29	1	0	1	0	0
45	27	3	-2	9	4	-6
42	27	0	-2	0	4	0
44	19	2	-10	4	100	-20
40	18	-2	-11	4	121	22
46	19	4	-10	16	100	-40
44	31	2	2	4	4	4
Total		11	-43	87	447	-108

$$\begin{aligned}
\therefore r &= \frac{n\sum dxdy - (\sum dx)(\sum dy)}{\sqrt{n\sum dx^2 - (\sum dx)^2} \sqrt{n\sum dy^2 - (\sum dy)^2}} \\
&= \frac{12 \times (-108) - (11)(-43)}{\sqrt{12 \times 87 - (11)^2} \sqrt{12 \times 447 - (-43)^2}} \\
&= \frac{-823}{\sqrt{923} \sqrt{3515}} \\
&= \frac{-823}{1801.21} \\
&= -0.457
\end{aligned}$$

NOTE :- All the methods provide same value of coefficient of correlation.

Some Properties of Correlation Coefficient

The following are the main properties of the coefficient of correlation :

- The correlation coefficient between two variables is symmetric, i.e. correlation coefficient between X and Y (r_{xy}) is same as correlation coefficient between Y and X (r_{yx}), i.e. $r_{xy} = r_{yx}$.
- The coefficient of correlation, r , lies between -1 and +1, i.e. $-1 \leq r \leq +1$.
- The coefficient of correlation is independent of change of origin and scale.

Here change in origin means adding or subtracting some constant value from given observations on

the variable X and Y and change in scale means multiplying or dividing the observations on X and Y by same constant value. Suppose we want to obtain correlation coefficient between X and Y. Now to determine r_{xy} , let the constants 'a' and 'b' be subtracted from X and Y respectively. The resulting values of X and Y be further divided by 'h' and 'k'. Let us denote the new values by U and V respectively, i.e.,

$$U = \frac{X-a}{h} \text{ and } V = \frac{Y-b}{k}$$

Then according to this property the correlation coefficient between X and Y will be same as that of between U and V. Thus instead of finding coefficient of correlation between X and Y, we first define U and V and then find out coefficient of correlation between U and V. Thus property is very useful for reducing computational work involved in the coefficient of correlation.

NOTE :- If we take $h = k = 1$, then this property reduced to short-cut method of determining coefficient of correlation. The computational procedure will be more clear from the following example.

Example 4. Find the coefficient of correlation between X and Y from the following data:

Capital Invested ('000 Rs.): 100 90 80 70 60 50 40 30 20 10

Profit ('000 Rs.): 85 75 65 55 45 35 25 15 5 5

Solution : First we define U and V as

$$U = \frac{X-60}{10} \text{ and } V = \frac{Y-45}{10}$$

X	Y	U	V	U ²	V ²	UV
100	85	4	4	16	16	16
90	75	3	3	9	9	9
80	65	2	2	4	4	4
70	55	1	1	1	1	1
60	45	0	0	0	0	0
50	35	-1	-1	1	1	1
40	25	-2	-2	4	4	4
30	15	-3	-3	9	9	9
20	5	-4	-4	16	16	16
10	5	-5	-4	25	16	20
Total		-5	-4	85	76	80

$$\begin{aligned}
\therefore r &= \frac{n\sum UV - (\sum U)(\sum V)}{\sqrt{n\sum U^2 - (\sum U)^2} \sqrt{n\sum V^2 - (\sum V)^2}} \\
&= \frac{10 \times 80 - (-5)(-4)}{\sqrt{10 \times 85 - (-5)^2} \sqrt{10 \times 76 - (-4)^2}} \\
&= \frac{780}{\sqrt{825} \sqrt{744}} \\
&= \frac{780}{783.45} = 0.9956 \\
&= 783.45 = 0.9956
\end{aligned}$$

This means a very high degree positive correlation between capital invested and profit earned.

Coefficient of Determination

The coefficient of determination is a simple and useful way of interpreting the value of correlation coefficient. It indicates the percentage variation in the dependent variable. In other words, it explains the production of variation in the dependent variable which is explained by a change in the independent variable. The coefficient of determination, denoted by R, is defined as

$$R = r^2$$

$$\text{Also } R = \frac{\text{Explained Variation}}{\text{Total Variation}}$$

If $r = 0.80$, then R will be 0.64 and it means that 64% variation in the dependent variable has been explained by the independent variable.

CHECK YOUR PROGRESS-I

Multiple Choice questions

1. **Which of the following are types of correlation?**
 - a. Positive and Negative
 - b. Simple, Partial and Multiple
 - c. Linear and Nonlinear
 - d. All of the above
2. **Which of the following is true for the coefficient of correlation?**
 - a. The coefficient of correlation is not dependent on the change of scale
 - b. The coefficient of correlation is not dependent on the change of origin
 - c. The coefficient of correlation is not dependent on both the change of scale and change of origin
 - d. None of the above
3. **Which of the following statements is true for correlation analysis?**
 - a. It is a bivariate analysis
 - b. It is a multivariate analysis
 - c. It is a univariate analysis
 - d. Both a and c
4. **If the values of two variables move in the same direction, _____**
 - a. The correlation is said to be non-linear
 - b. The correlation is said to be linear
 - c. The correlation is said to be negative
 - d. The correlation is said to be positive
5. **If the values of two variables move in the opposite direction, _____**
 - a. The correlation is said to be linear
 - b. The correlation is said to be non-linear
 - c. The correlation is said to be positive
 - d. The correlation is said to be negative

7.6 SPEARMAN'S COEFFICIENT OF RANK CORRELATION

We have discussed Karl Pearson's Coefficient of Correlation, studied the degree of co-variability of linear relationship between two variables for which the observations are definitely measured. But often we come across situations when definite measurements on the variables are not possible. For example, if a group of n students is arranged in order of merit or proficiency in Business Statistics and Economics without any attempt to assess numerically assigning to each student a number which indicates his position in that group; the students are then said to be ranked and the number of a particular student is his rank. In such type of situations Spearman's Coefficient of rank correlation is determined.

Spearman suggested that the relationship between two ranks may be studied by calculating the Pearson's Coefficient of Correlation for numerical values that happens to rank. The Spearman's Coefficient of Rank Correlation, denoted by a Greek letter ρ , is given by

$$\rho = 1 - \frac{6\sum D^2}{n(n^2 - 1)}$$

where

D = difference between paired ranks, i.e.

$$D_i = R_{xi} - R_{yi}$$

R_{xi} = rank of i th individual of variable X

R_{yi} = rank of i th individual of variable Y

n = the number of items ranked

Remarks :

1. In fact, the coefficient of rank correlation, is nothing but Karl Pearson's coefficient of correlation between two sets of ranks.
2. In view of remark 1, its value lies between -1 and +1.
3. The value $= +1$ stands for a perfect positive agreement between two sets of ranks, while $= -1$ implies a perfect negative relationship.
4. The basic assumption in this correlation is that no two individuals be equal in either classification so that no ties in ranks exists.

Example 5. In a beauty contest two judges rank the 10 entries as follows :

Contestant :	A	B	C	D	E	F	G	H	I	J	
Judge I:	1	2	3	4	5	6	7	8	9	10	
Judge II	:	2	3	1	6	4	5	8	7	10	9

Find the degree of agreement between ranks given by two judges.

Solution : In order to find the degree of agreement between ranks, we find the coefficient of rank correlation.

Computation of Rank Correlation

Contestant	Rank by Judge I	Rank by Judge II	$D = R_1 - R_2$	D^2
A	1	2	-1	1
B	2	3	-1	1
C	3	1	2	4
D	4	6	-2	4
E	5	4	1	1
F	6	5	1	1
G	7	8	-1	1
H	8	7	1	1
I	9	10	-1	1
J	10	9	1	1
Total				16

Thus, rank correlation is given by

$$\rho = 1 - \frac{6 \sum D^2}{n(n^2 - 1)} = 1 - \frac{6 \times 16}{10(100 - 1)}$$

$$= 1 - \frac{96}{990} = \frac{894}{990} = 0.903$$

Thus, $\rho = 0.903$ shows a high degree of agreement between ranks given by two judges.

Example 6. Calculate rank correlation coefficient from the following marks given out of 200 by two judges X and Y in a music competition to 8 participants:

Participant No : 1, 2, 3, 4, 5, 6, 7, 8

Marks awarded by X : 74, 98, 110, 70, 65, 85, 88, 59

Marks awarded by Y : 121, 133, 170, 102, 90, 152, 160, 85

Solution : To determine rank correlation we first assign ranks to marks awarded by judge X and Y by allotting the first rank to the highest marks, second rank to next highest marks, and so on. The ranks so obtained for the two judges are given in the following table :

Participant No.	Marks by X	Marks by Y	Rank for Marks X R1	Rank for Marks X R2	D = R1-R2	D ²
1	74	121	5	5	0	0
2	98	133	2	4	-2	4
3	110	170	1	1	0	0
4	70	102	6	6	0	0
5	65	90	7	7	0	0
6	85	152	4	3	1	1
7	88	160	3	2	1	1
8	59	85	8	8	0	0
Total						6

Thus, Rank correlation is

$$\rho = 1 - \frac{6 \sum D^2}{n(n^2 - 1)} = 1 - \frac{6 \times 6}{8(64 - 1)}$$

$$= 1 - \frac{36}{504} = \frac{468}{504} = 0.929$$

which shows a high degree of positive correlation between the marks awarded by the two judges.

Rank Correlation for Tied Ranks

Rank Correlation For Tied Ranks

In the previous section we have assumed that no two values in either series were equal (means no tie) while discussing rank correlation. However, in some cases, we may have two or more equal observations in either of the two series or in both the series. In such cases, we assign average (mean) ranks to the set of tied observations. For example, in assigning ranks to 10 observations we may note that the third largest observation is repeating three times. These three observations (3rd, 4th and 5th) are therefore tied and each is assigned average rank $\frac{1}{3} (3+4+5) = 4$. The next individual assigned the rank 6. If we find again a tie of two observations, we assign the rank $\frac{7+8}{2} = 7.5$ each and the next individual is assigned rank 9.

Obviously the formula

$$\rho = 1 - \frac{6 \sum D^2}{n(n^2 - 1)} \quad \dots\dots\dots(1)$$

cannot be used if there are ties in either one or both series. The Spearman's Coefficient of Rank Correlation is then corrected for these tied ranks and now given as follows :

$$\rho = 1 - \frac{6 \left[\sum D^2 + \sum m(m^2 - 1)/2 \right]}{n(n^2 - 1)} \quad \dots\dots\dots(2)$$

where m be the number of tied observations with common ranks.

NOTE :- The adjustment consists of adding $m(m^2-1)/12$ to the value $\sum D^2$. If there are more than one set of tied observations, the correction factor $m(m^2-1)/12$ is to be added each time to the value of $\sum D^2$ for every value of m. This tendency of correction has been represented by $\sum m(m^2-1)/12$ in formula (2).

Example 7. Calculate the coefficient of rank correlation from the following data : Marks by

Judge I :	48	33	40	09	16	16	65	24	16	57
Judge II :	13	13	24	06	15	04	20	09	06	19

Solution :

Judge I (X)	Judge II (Y)	Rank R_1	Rank R_2	$D = R_1 - R_2$	D^2
48	13	3	5.5	-2.5	6.25
33	13	5	5.5	-0.5	0.25
40	24	4	1	-3.0	9.00
09	06	10	8.5	1.5	2.25
16	15	8	4	4	16.00
16	04	8	10	-2	4.00
65	20	1	2	-1	1.00
24	09	6	7	-1	1.00
16	06	8	8.5	-0.5	0.25
57	19	2	3	-1	1.00
Total					41

While ranking observations of judge I i.e. X, three observations of 16 are tied at rank 7 and, as such they are each marked the average rank of $(7+8+9)/3 = 8$. Similarly the observation of Judge II i.e. Y, two observations of values 13 and 6 are tied at rank 5th and 8th respectively. Therefore, these observations are assigned rank $(5+6)/2 = 5.5$ and $(8+9)/2 = 8.5$ respectively. Thus, in all there are 3 tied ranks for observation 16 of Judge I. Similarly, there are 2 tied ranks for observation 13 and 2 for observation 6 of Judge II i.e. Y. Thus, we have three sets of tied ranks with $m = 3, 2$ and 2 respectively. Therefore, on using formula (2), Spearman's Coefficient of rank correlation is given by

$$\rho = 1 - \frac{6 \left[\sum D^2 + \sum m(m^2 - 1)/12 \right]}{n(n^2 - 1)}$$

$$= 1 - \frac{6 \left[41 + \{ 3(9-1)/12 + 2(4-1)/12 + 2(4-1)/12 \} \right]}{10(100-1)}$$

$$= 1 - \frac{6 \left[41 + 2 + \frac{1}{2} + \frac{1}{2} \right]}{10 \times 99} = 1 - \frac{258}{990} = \frac{732}{990}$$

$$= 0.739$$

7.7 LET US SUM UP

The three different methods of calculating coefficient of correlation have been discussed in this lesson. The formula for coefficient of correlation under different methods are

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2}} \quad \text{(Actual Mean Method)}$$

$$= \frac{n \Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{n \Sigma X^2 - (\Sigma X)^2} \sqrt{n \Sigma Y^2 - (\Sigma Y)^2}} \quad \text{(Direct Method)}$$

$$= \frac{n \Sigma dx dy - (\Sigma dx)(\Sigma dy)}{\sqrt{n \Sigma dx^2 - (\Sigma dx)^2} \sqrt{n \Sigma dy^2 - (\Sigma dy)^2}} \quad \text{(Shortcut Method)}$$

Spearman's Coefficient of rank correlation is given by

(a) For Untied Ranks :

$$\rho = 1 - \frac{6 \Sigma D^2}{n(n^2 - 1)}$$

(b) For Tied Ranks :

$$\rho = 1 - \frac{6 \left[\Sigma D^2 + \frac{1}{12} \cdot \Sigma m(m^2 - 1) \right]}{n(n^2 - 1)}$$

7.8 GLOSSARY

- The relationship between two variables is explained by correlation.
- Correlation may be simple (between two variables) multiple (between many variables) or partial (between many variables where the effect of other Variables is neutralized and in absence the relationship of two variables is observed).
- Scatter Diagram is a graphical method to observe correlation between variables.
- The value of correlation varies between +1 to -1.

7.9 SELF-ASSESSMENT QUESTIONS

1. What do you understand by the terms:

- (i) Bivariate distribution
- (ii) Correlation
- (iii) Dot diagram

2. Discuss the meaning of correlation and distinguish between positive and negative correlations.

3. What do you mean by scatter diagram? How is scatter diagram used to determine correlation.

4. Define Karl Pearson's coefficient of correlation. What is it intending to measure.

-
-
5. Give some examples of positive and negative correlations.
-
-
6. What will be your interpretation if
(i) $r = 0$ (ii) $r = +1$ (iii) $r = -1$
7. Determine the correlation coefficient from the following data on using different method. Also find P.E., S.E., and coefficient of determination. Comment on your results.
- X : 280, 290, 290, 310, 300, 320, 330, 340
Y : 150, 160, 150, 180, 190, 210, 200, 220
8. What is rank correlation? Discuss its merits and demerits.
-
-
9. Write short note on Spearman's rank correlation coefficient.
-
-
10. The rank correlation coefficient between marks obtained by some students in two subjects is 0.80. If the sum of squares of difference of ranks is 33, then find the number of students.
-
-

11. Compute the rank correlation coefficient from the following data:

X: 115, 109, 112, 87, 98, 87, 109, 108

Y: 75, 74, 80, 76, 74, 70, 68, 70

12. Ten competitions in a beauty contest are ranked by three judges in the following order.

Judge

A: 1, 6, 5, 10, 3, 2, 4, 9, 7, 8

B: 3, 5, 8, 4, 7, 10, 2, 1, 6, 9

C: 6, 4, 9, 8, 1, 2, 3, 10, 5, 7

Use Spearman's Rank Correlation to determine which pair of judges has the nearest approach to common tastes in beauty.

7.10 LESSON END EXERCISE

1. Correlation is used to measure the qualitative relationship between variables.
(True / False)
2. We call it perfectly positive correlation when the value of correlation is
a) +1 b) -1 c) 0 d) between 0 and 1
3. When one need to check the correlation among ranking, which technique used
a) Scatter diagram b) Karl Pearson's coefficient of correlation
c) Spearman's method d) Standard deviation

7.11 SUGGESTED READINGS

1. Argyrous, George. 1997. Statistics for Social Research. New York: Mc Millan Press Ltd.
2. Goods, W.J. & Hatt, P.K. 1981. Methods in Social Research. New York: Mc Graw Hill.
3. Gupta, S.C. 1981. Fundamentals of Statistics. Bombay: Himalayan Publishing House.
4. Gupta, S.P. 2004. Statistical Methods. New Delhi: Sultan Chand and Sons.

7.12 ANSWERS TO CHECK YOUR PROGRESS

Answers: 1- d, 2-c, 3-c, 4-d, 5-d

REGRESSION ANALYSIS AND THEIR RELATIONSHIP**STRUCTURE**

- 8.0 Objectives
- 8.1 Introduction
- 8.2 Difference between Correlation and Regression
- 8.3 Regression Analysis and Lines of Regression
- 8.4 Fitting of Regression Line of Y on X
- 8.5 Fitting of Regression Line of X on Y
- 8.6 Some Properties Regression Coefficients and Lines
- 8.7 Standard Error of Estimates
- 8.8 Let us sum up
- 8.9 Glossary
- 8.10 Self-Assessment Questions
- 8.11 Lesson End Exercise
- 8.12 Suggested Readings
- 8.13 Answers to check your progress

8.0 OBJECTIVES

The main objectives of this lesson are:

- to introduce the concept of regression analysis and distinguish it with correlation coefficient
- to explain the two lines of regression

- to provide the computational procedure for determining the constants of regression line of Y on X and X on Y.
- to explain the various important properties and applications of regression coefficients, and
- to give an idea of standard error of estimates.

8.1 INTRODUCTION

In the pervious lesson, we have seen that the data giving the corresponding value of two variables can be graphically represented by a scatter diagram and a method of finding the relationship between these two variables in terms of correlation coefficient was also introduced. Very often, in the study of relationship of two variables, we come across situations where one of the two variables depends on the other. In other words, what is the possible value of the dependent variable when the value of the independent variable is known. In such situations, where one of the variables is dependent and other is independent, we can find a method of estimating the numerical relationship between two variables so that given a value of the independent variable, we can forecast the average value of the dependent variable. Regression analysis serves this purpose.

8.2 DIFFERENCE BETWEEN CORRELATION AND REGRESSION ANALYSIS

Although the two analyses are complementary to one another, yet the choice of one or the other depends upon the purpose of statistical enquiry. The following are the main differences between correlation and regression analysis:

- (i) The coefficient of correlation is used to measure the degree of covariation between the two variables, while the regression analysis provides the average relationship between these variables.
- (ii) Correlation does not necessarily establish causes and effect relationship. However, in regression analysis, there is a clear indication of cause-and-effect relationship. Here the independent variable is the cause and dependent variable is the effect.
- (iii) Whereas correlation analysis is confined only to the study of linear relationship between two variables, the regression analysis deals with linear and non-linear relationships.
- (iv) In correlation analysis, r_{xy} measures the linear relationship between the variables X and Y. Here $r_{xy} = r_{yx}$, i.e., it is immaterial which of the two variables is taken as dependent or independent. However, in regression analysis, the identity of variables, i.e., which is dependent and which one is independent, is important.

8.3 REGRESSION ANALYSIS AND LINES OF REGRESSION

The word regression was first introduced by Sir Francis Galton in the study of heredity in connection

with the study of height of parents and their offsprings. He found that the offspring of tall or short parents tend to regress to the average height. In other words, though tall fathers do tend to have tall sons, yet the average height of sons of a group of tall fathers is less than their father's height and the average height of short fathers is less than the average height of their sons. Galton termed the line describing the average relationship between the two variables as the line of regression. Thus, by regression we mean the average relationship between two variables which can be used for estimating the value of one variable from the given values of another variable. However, the dictionary meaning of regression is "Stepping Back", but nowadays it is stand for some sort of functional relationship between two or more variables. Here the variable whose value is to be predicted is called dependent or explained variable and the variable used for prediction is called independent or explanatory variable.

Lines of Regression

If the variables in a bivariate frequency distribution are correlated, we observe that the points in a scatter diagram cluster around a straight line, called the line of regression. In a bivariate study, we have two lines of regression, namely, regression of Y on X and regression of X on Y.

The line of regression of Y on X is used to predict or estimate or forecast the value of Y for the given value of the variable X. Thus, Y is the dependent variable and X is the independent variable. The regression line of Y on X is of the form:

$$Y = a + bX$$

where a and b are unknown constants to be determined by observed data on the two variables X and Y.

Similarly, the regression line of X on Y is used to predict the value of X for the given value of the variable Y. Here X is dependent variable and Y is independent. The regression line of X on Y is of the form:

$$X = a + bY$$

where a and b are unknown constants to be determined by observed data on the two variables X and Y.

8.4 FITTING OF REGRESSION LINE OF Y ON X

Suppose the regression line of Y on X is

$$Y = a + bX \quad \dots(1)$$

where a and b are unknown constants to be determined by observed data on the two variables X and Y.

The regression line of Y on X, given by (1) can be fitted by the Method of Least Squares. That is, we choose the constant a and b in the regression line $Y = a + bX$ in such a way that

$$\sum (Y_i - \hat{Y}_i)^2 \quad \dots(2)$$

is a minimum, where $\hat{Y}_i$ be the estimated value of Y for $X = X_i$ i.e. $\hat{Y}_i = a + bX_i$. Here the quantity given by (2) is called sum of squares of the residuals E_i , $E_i = y_i - \hat{Y}_i$

For obtaining a and b , we minimize

$$\begin{aligned} \sum E_i^2 &= \sum (y_i - \hat{Y}_i)^2 \\ &= \sum (y_i - a - bX_i)^2 \end{aligned} \quad \dots(3)$$

with respect to a and b . By using the Principle of Maxima and Minima, i.e., equating to zero the partial derivatives of $\sum E_i^2$ w.r.t. a and b , we get

$$\sum Y_i = na + b \sum X_i \quad \dots(4)$$

$$\sum X_i Y_i = a \sum X_i + b \sum X_i^2 \quad \dots(5)$$

Here these two equations i.e. (4) and (5) are called normal equations. Solving these equations simultaneously for a and b , we obtain.

$$\begin{aligned} b &= \frac{\frac{1}{n} \sum X_i Y_i - \bar{X} \bar{Y}}{\frac{1}{n} \sum X_i^2 - \bar{X}^2} \\ &= \frac{\sum X_i Y_i - n \bar{X} \bar{Y}}{\sum X_i^2 - n \bar{X}^2} \end{aligned} \quad \dots(6)$$

and

$$a = \bar{Y} - b \bar{X} \quad \dots(7)$$

The values of b and a given by (6) and (7) can also be expressed in terms of correlation coefficient. As we know that

$$\begin{aligned}
 r &= \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y} \\
 &= \frac{\frac{1}{n} \sum X_i Y_i - \bar{X} \bar{Y}}{\sqrt{\frac{1}{n} \sum X_i^2 - \bar{X}^2} \sqrt{\frac{1}{n} \sum Y_i^2 - \bar{Y}^2}} \quad \text{.....(8)}
 \end{aligned}$$

Thus from (6) and 8, we have

$$b = r \frac{\sigma_Y}{\sigma_X} \quad \text{.....(9)}$$

and

$$a = \bar{Y} - r \frac{\sigma_Y}{\sigma_X} \cdot \bar{X} \quad \text{.....(10)}$$

Hence, on putting the values of b and a from equations (9) and (10) in regression line of Y on X given by equation (2), we obtain the following equation of regression line of Y on X :

$$Y = \bar{y} - r \frac{\sigma_y}{\sigma_x} \cdot \bar{x} + r \cdot \frac{\sigma_y}{\sigma_x} \cdot x$$

$$= \bar{y} + r \frac{\sigma_y}{\sigma_x} (X - \bar{X})$$

or

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} \cdot (x - \bar{x}) \quad \text{.....(11)}$$

The quantity $r \cdot \frac{\sigma_y}{\sigma_x}$ is called the regression coefficient of Y on X and, in general, is denoted by b_{yx} . Thus

$$b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x}.$$

Remark

- (i) We may ignore the lower suffix i from X_i and Y_i in the formula of a and b . Thus, we may write

$$b = \frac{\Sigma XY - n \bar{X} \bar{Y}}{\Sigma X^2 - n \bar{X}^2}, \text{ and}$$

$$a = \bar{Y} - b \bar{X}$$

- (ii) If X and Y are measured from their respective means i.e.

let $x = X - \bar{X}$ and $y = Y - \bar{Y}$, then b is given by

$$b = \frac{\Sigma xy}{\Sigma x^2}$$

- (iii) To find the $\bar{Y}$, $\bar{X}$ and b_{yx} we can also use step deviation method, i.e., if we assume that $d_x = X - A$ and $d_y = Y - B$, where A and B are assumed mean of X and Y respectively, then we have

$$\bar{X} = A + \frac{1}{n} \Sigma dx$$

$$\bar{Y} = B + \frac{1}{n} \Sigma dy$$

and

$$b_{yx} = \frac{n \Sigma dxdy - (\Sigma dx)(\Sigma dy)}{n \Sigma dx^2 - (\Sigma dx)^2}$$

- (iv) The regression line of Y on X given by equation (1) is used to estimate or predict the best value of Y for a given value of the variable X .

Example 1. Find the regression equation of Y on X from the following data :

X	:	7	4	8	6	5
Y	:	6	5	9	8	2

Also estimate the value of Y when the value of X = 12.

Solution : Let the regression line of Y on X is

$$Y = a + bX$$

.....(1)

X	Y	XY	X ²
7	6	42	49
4	5	20	16
9	8	72	64
6	8	48	36
5	2	10	25
30	30	192	190

$$\text{Here } \bar{X} = \frac{\sum X}{n} = \frac{30}{5} = 6, \quad \bar{Y} = \frac{30}{5} = 6,$$

$\sum XY = 192$, and $\sum X^2 = 190$, thus

$$b(=b_{yx}) = \frac{\frac{1}{n} \sum XY - \bar{X}\bar{Y}}{\frac{1}{n} \sum X^2 - \bar{X}^2}$$

$$= \frac{\frac{192}{5} - 6 \times 6}{\frac{190}{5} - (6)^2} = \frac{2.4}{2} = 1.20$$

and

$$a = \bar{Y} - b\bar{X} = 6 - 1.20 \times 6$$

$$= -1.20$$

Thus regression equation of Y on X becomes

$$Y = -1.20 + 1.20X$$

and the estimated value of Y for X = 12 is

$$Y = -1.20 + 1.20 \times 12$$

$$= -1.20 + 14.40$$

$$= 13.20$$

Example 2. Consider the following data on heights and weights of 10 adults :

Height (cm) : 178 176 170 174 165 162 178 165 174 172

Weight (kg) : 80 75 72 74 68 64 76 66 72 70

Predict the weight of an adult whose height is 185cm.

Solution : Let Y = Weight & X = Height. First we find the regression line of Y on X

Sr. No.	X	Y	$dx = X - 174$	$dy = Y - 70$	$dx dy$	d_x^2
1	178	80	4	10	40	16
2	176	75	2	5	10	4
3	170	72	-4	2	-8	16
4	174	74	0	4	0	0
5	165	68	-9	-2	18	81
6	162	64	-12	-6	72	144
7	178	76	4	6	24	16
8	165	66	-9	-4	36	81
9	174	72	0	2	0	0
10	172	70	-2	0	0	4
Total			-26	17	192	362

$$\text{Here } \bar{X} = A + \frac{\sum dx}{n} = 174 - \frac{26}{10} = 171.40$$

$$\bar{Y} = A + \frac{\sum dy}{n} = 70 + \frac{17}{10} = 71.70$$

$$b (= b_{yx}) = \frac{n \sum dxdy - (\sum dx)(\sum dy)}{n \sum dx^2 - (\sum dx)^2}$$

$$= \frac{10 \times 192 - (-26)(17)}{10 \times 362 - (-26)^2}$$

$$= \frac{2362}{2944} = 0.802$$

Thus, regression line of Y on X is

$$Y - \bar{Y} = b_{yx} (X - \bar{X})$$

$$\text{or } Y - 71.70 = 0.802 (X - 171.40)$$

$$Y = 0.802X - 137.463 + 71.70$$

$$Y = -66.06 + 0.802X$$

Hence the weight (Y) of an adult, whose height (X) is 185, is given by

$$Y = -66.06 + 0.802 \times 185 = 82.31$$

Example 3. Fit the equation of regression line of Y on X for the following data :

X :	57	58	59	60	61	62	64
Y :	77	78	75	82	82	79	81

Solution :

Sr. No.	X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	$(X - \bar{X})^2$ $= x^2$	$(X - \bar{X})(Y - \bar{Y})$ $= xy$
1	57	77	-4	-2	16	8
2	58	78	-3	1	9	-3
3	59	75	-2	-4	4	8
4	60	81	-1	2	1	-2
5	62	82	1	3	1	3
6	65	79	4	0	16	0
7	66	81	5	2	25	10
Total	427	553	0	0	72	24

Here $\bar{X} = \frac{\Sigma X}{n} = \frac{427}{7} = 61$, $\bar{Y} = \frac{\Sigma Y}{n} = \frac{553}{7} = 79$

Thus

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{24}{72} = 0.333$$

Hence, the regression line of Y on X is

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

or $Y - 79 = 0.333(X - 61)$

$$\therefore Y = 58.687 + 0.333X$$

8.5 FITTING OF REGRESSION LINE OF X ON Y

As we have discussed in section 13.4 that the regression line of X on Y is used to

estimate or predict or forecaste the value of X for a given value of the variable Y. In this case X is called dependent variable and Y is the independent variable. The standard form of regression line of X on Y is

$$X = a + by \quad \text{..... (1)}$$

where a and b are unknown constants which are determined from the giving set of data on X and Y.

On using the procedure of least squares method similar to that of Section 14.6, Lesson 14, we obtained the following two normal equations :

$$\sum X = na + b\sum y \quad \text{..... (2)}$$

$$\sum XY = a\sum y + b\sum y^2 \quad \text{..... (3)}$$

Simplifying the equations (2) and (3) for a and b , we obtain

$$b = \frac{\frac{1}{n} \sum XY - \bar{X}\bar{Y}}{\frac{1}{n} \sum Y^2 - \bar{Y}^2} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum Y^2 - n\bar{Y}^2} \quad \text{..... (4)}$$

and

$$a = \bar{X} - b\bar{Y} \quad \text{..... (5)}$$

The value of b , given by (4), generally denoted by b_{xy} , is called regression coefficient of X on Y. It can also be expressed in terms of correlation coefficient (r) as

$$b_{xy} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum Y^2 - n\bar{Y}^2} = r \frac{\sigma_x}{\sigma_y}$$

Hence, the regression equation of X on Y becomes

$$X = (\bar{X} - b_{xy} \bar{Y}) + b_{xy} Y$$

or

$$X - \bar{X} = b_{xy} (Y - \bar{Y})$$

$$X - \bar{X} = r \frac{\sigma_x}{\sigma_y} (Y - \bar{Y}) \quad \text{..... (6)}$$

Remarks :

- (i) Assuming that X and Y each are measured about their means, then

$$b_{xy} = \frac{\sum xy}{\sum y^2}, \text{ where } x = X - \bar{X} \text{ and } y = Y - \bar{Y}.$$

- (ii) If X and Y are each measured about their assumed mean, then

$$b_{XY} = \frac{n \sum dx dy - (\sum dx)(\sum dy)}{n \sum dy^2 - (\sum dy)^2}$$

where $dx = X - A$, $dy = Y - B$, A and B are assumed means of X and Y.

CHECK YOUR PROGRESS-I

MULTIPLE CHOICE QUESTIONS

1. Which of the following best describes the purpose of regression analysis in management accounting?

- A) To identify the correlation coefficient between two variables
- B) To predict the value of a dependent variable based on one or more independent variables
- C) To calculate the break-even point
- D) To allocate overhead costs accurately

2. The coefficient of determination, R^2 , indicates:

- A) The strength and direction of a linear relationship
- B) The proportion of variance in the dependent variable explained by the independent variables
- C) The standard error of estimate
- D) The slope of the regression line

3. Which of the following is NOT an assumption of the classical linear regression model?

- A) Errors have zero mean
- B) Errors are homoscedastic (constant variance)
- C) Errors are correlated with independent variables
- D) Errors are normally distributed

4. Multicollinearity in multiple regression refers to:

- A) The correlation among dependent variables
- B) The correlation among independent variables
- C) Errors correlated across observations
- D) The assumption that errors have equal variance

5. What does a residual represent in regression analysis?

- A) Predicted value of dependent variable
- B) Difference between observed and predicted values
- C) Coefficient estimate
- D) Error term's variance

8.6 SOME PROPERTIES OF REGRESSION COEFFICIENTS AND LINES

As we know that the regression line of Y on X and X on Y are

$$Y - \bar{Y} = b_{YX} (X - \bar{X}), \text{ and}$$

$$X - \bar{X} = b_{XY} (Y - \bar{Y}),$$

where $b_{YX} = r \cdot \frac{\sigma_Y}{\sigma_X}$ and $b_{XY} = r \cdot \frac{\sigma_X}{\sigma_Y}$ are the regression coefficients of Y on X and X on Y respectively.

Keeping in view these lines and regression coefficients, we now, present some following properties which are very helpful in understanding the regression lines more clearly and to obtain some other

measures.

(i) The regression lines of Y on X and X on Y both pass through the point $(\bar{X}, \bar{Y})$. That is, they intersect each other at the point $(\bar{X}, \bar{Y})$.

This property help us in determining $\bar{X}$ and $\bar{Y}$ if regression lines are given.

(ii) If correlation coefficient, r , is zero then two regression lines are $Y = \bar{Y}$ and

$X = \bar{X}$, i.e., the two regression lines are perpendicular to each other.

(iii) If $r = \pm 1$, i.e., in case of perfect correlation, we have only one regression line which is

$$\left(\frac{Y - \bar{Y}}{\sigma_Y} \right) = \pm \left(\frac{X - \bar{X}}{\sigma_X} \right)$$

(iv) The sign of both the regression coefficients must be same.

(v) If one regression coefficient is greater than one, then other must be less than one. That is, if $b_{YX} > 1$, then $b_{XY} < 1$ or if $b_{XY} > 1$, then $b_{YX} < 1$. However, both may be less than one.

(vi) The geometric mean of the regression coefficients is equal to correlation coefficient (r). That is

$r = \sqrt{b_{YX} b_{XY}}$, but the sign of correlation coefficient is the same as that of two regression coefficients.

(vii) The arithmetic mean of the regression coefficients must be greater than the correlation coefficient.

(viii) Regression coefficients are not symmetrical functions in X and Y as correlation coefficient. That is $b_{YX} \neq b_{XY}$ as $r_{XY} = r_{YX}$.

(ix) Regression lines are not mutually reversible. That is, we cannot estimate the value of Y from the regression line of X on Y and vice-versa.

(x) Regression coefficients are independent of change of origin but not of scale.

Symbolically if $U = \frac{X - A}{h}$ and $V = \frac{Y - B}{k}$, then

$$b_{UV} = \frac{h}{k} b_{XY} \text{ and } b_{VU} = \frac{k}{h} b_{YX}.$$

If $h=k$ then $b_{UV}=b_{XY}$ and $b_{VU}=b_{YX}$.

CHECK YOUR PROGRESS-II

MULTIPLE CHOICE QUESTIONS

Q1. What is another name for a regression line?

- A. Line of the average relationship
- B. Estimating equation
- C. Prediction equation
- D. All of the above

Q2. For two random variables X and Y, how many lines of regression are possible?

- A. 1
- B. 2
- C. 3
- D. 0

Q3. If the correlation coefficient is, the percentage of variation in the dependent variable explained by the independent variable is:

- A. 0.80%
- B. 80%
- C. 0.64%
- D. 64%

Q4. What does it mean if the two regression lines are perpendicular to each other?

- A. No correlation
- B. High correlation
- C. Perfect correlation
- D. Zero correlation

8.7 STANDARD ERROR OF ESTIMATE

As discussed, the regression lines define an average relationship between two (or more) variables which can be used for estimating or forecasting the value of the dependent variable from the given values of independent variable (s). The regression lines are fitted to observed data by using the method of least squares. Although these lines are used for prediction, but it is not possible to have a perfect prediction of the values by using these lines. Thus, we need a quantitative measure which may be used to indicate how precise the prediction. The standard error of estimate provides us a measure of the scatter of the observations about an average line. For two regressions lines, we have two the following two standard error of estimates.

- (i) If the regression line of Y on X is given by $Y = a + bx$, then the standard error of estimate, denoted by $S_{y.x}$ is given by

$$S_{Y.X} = \sqrt{\frac{1}{n} \sum (Y - \hat{Y})^2}$$

where $\hat{Y}$ is the estimated value of Y.

$S_{Y.X}$ can also be expressed as

$$S_{Y.X} = \sigma_Y \sqrt{1 - r^2}$$

- (ii) If the regression line of X on Y is $X = a + bY$, then the standard error of estimate, denoted by $S_{x.y}$, is given by

$$S_{X.Y} = \sqrt{\frac{1}{n} \sum (X - \hat{X})^2}$$

where $\hat{X}$ is the estimated value of X.

The standard error of estimate $S_{X.Y}$ may also be expressed in the form

$$S_{X.Y} = \sigma_X \sqrt{1 - r^2}$$

Example 4. The following table gives the respective weights of a sample of 12 fathers and their sons:

Weight of Fathers (ks) : 65, 63, 67, 64, 68, 62, 70, 66, 68, 67, 69, 71

Weight of Sons (ks) : 68, 66, 68, 65, 69, 66, 68, 65, 71, 67, 68, 70

Obtain the two regression lines and the corresponding standard error. Also find correlation

coefficient.

Solution: Let X = Weight of Fathers

Y = Weight of Sons

X	Y	$d_x = X-68$	$d_y = Y-68$	dx^2	dy^2	$dx dy$
65	68	-3	0	9	0	0
63	66	-5	-2	25	4	10
67	68	-1	0	1	0	0
64	65	-4	-3	16	9	12
68	69	0	1	0	1	0
62	66	-6	-2	36	4	12
70	68	2	0	4	0	0
66	65	-2	-3	4	9	6
68	71	0	3	0	9	0
67	67	-1	-1	1	1	1
69	68	1	0	1	0	0
71	70	3	2	9	4	6
Total		-16	-5	106	41	47

Here $\bar{X} = A + \frac{\sum dx}{n} = 68 - \frac{16}{12} = 66.67$

$$\bar{Y} = A + \frac{\sum dy}{n} = 68 - \frac{5}{12} = 67.58$$

$$\sigma_Y = \sqrt{\frac{1}{n} \sum dy^2 - \left(\frac{\sum dy}{n} \right)^2} = \sqrt{\frac{41}{12} - \left(\frac{-5}{12} \right)^2} = 3.24$$

$$= \sqrt{3.2431} = 1.80$$

$$\sigma_X = \sqrt{\frac{1}{n} \sum dx^2 - \left(\frac{\sum dx}{n} \right)^2} = \sqrt{\frac{106}{12} - \left(\frac{-16}{12} \right)^2} = \sqrt{7.5}$$

$$= 2.7386$$

$$b_{YX} = \frac{n \sum dx dy - (\sum dx)(\sum dy)}{n \sum dx^2 - (\sum dx)^2}$$

$$= \frac{12 \times 47(-16)(-5)}{12 \times 106 - (-16)^2}$$

$$= \frac{484}{1016} = 0.4764$$

and $b_{XY} = \frac{n \sum dx dy - (\sum dx)(\sum dy)}{n \sum dy^2 - (\sum dy)^2}$

$$= \frac{12 \times 47 - (-16)(-5)}{12 \times 41 - (-5)^2}$$

$$= \frac{484}{467} = 1.0364$$

Thus, the regression line of Y on X is $Y - \bar{Y} = b_{YX}(X - \bar{X})$

or $Y - 67.58 = 0.4764(X - 66.67)$

or $Y = 67.58 - (0.4764)(66.67) + 0.4764X$

or $Y = 35.818 + 0.4764X$

Regression line of X on Y is

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$

or $X - 66.67 = 1.0364(Y - 67.58)$

$$\text{or } X = -3.369 + 1.0364Y$$

The correlation coefficient (r) between X and Y is

$$r = \sqrt{b_{YX} \cdot b_{XY}} = \sqrt{0.4764 \times 1.0364} = 0.7027$$

Hence the standard error of estimates are

$$S_{Y.X} = \sigma_Y \sqrt{1 - r^2} = 1.80 \sqrt{1 - (0.7027)^2} = 1.281$$

$$\text{and } S_{X.Y} = \sigma_X \sqrt{1 - r^2} = 2.7386 \sqrt{1 - (0.7027)^2} = 1.948$$

Example 2. Regression lines of two variables X and Y are

$$4X - 5Y + 33 = 0 \quad \text{.....(1)}$$

$$20X - 9Y - 107 = 0 \quad \text{.....(2)}$$

and variance of X is 9.

Find (i) mean values of X and Y (ii) the regression line of Y on X and X on Y (iii) the correlation coefficient between X and Y (iv) standard deviation of Y (v) standard error of the estimate.

Solution :

Mean of X and Y : As we know that $(\bar{X}, \bar{Y})$ is the intersecting point of the two lines of regression, therefore it will satisfy the two lines and thus

$$4\bar{X} - 5\bar{Y} = -33 \quad \text{.....(3)}$$

$$20\bar{X} - 9\bar{Y} = 107 \quad \text{.....(4)}$$

Multiplying the equation (3) by 5 and then subtractors from equation (4), we get

$$20\bar{X} - 9\bar{Y} = 107$$

$$20\bar{X} - 25\bar{Y} = -165$$

- + +

$$16\bar{Y} = 272 \Rightarrow \bar{Y} = \frac{272}{16} = 17$$

Thus from (3)

$$4\bar{X} = 5\bar{Y} - 33 = 5 \times 17 - 33 = 52$$

$$\text{or } \bar{X} = \frac{52}{4} = 13$$

Hence $\bar{X} = 13$, and $\bar{Y} = 17$

ii **Regression lines** : Let the equations (1) be the regression line of Y on X and (2) be the X on Y. Therefore

$$4X - 5Y + 33 = 0 \Rightarrow 5Y = 33 + 4X$$

$$\text{or } Y = \frac{33}{5} + \frac{4}{5}X = 6.5 + \frac{4}{5}X \quad \dots(5)$$

$$\text{and } 20X - 9Y - 107 = 0 \Rightarrow 20X = 107 + 9Y$$

$$\text{or } X = \frac{9}{20}Y + \frac{107}{20} = 5.35 + \frac{9}{20}Y \quad \dots(6)$$

Thus, Here we observe that the values of b_{YX} and b_{XY} satisfy the properties of regression coefficients explained in Section 15.4 of this lesson. Thus equation (5) is the regression line of Y on X and (6) is the regression line of X on Y.

Coefficient of correlation

$$r = \pm \sqrt{b_{YX} b_{XY}} = \pm \sqrt{\frac{4}{5} \cdot \frac{9}{20}}$$

$$= \pm \sqrt{0.36} = \pm 0.60 = 0.60$$

Positive sign retained as both the regression coefficient are positive.

(v) **As we know that**

$$b_{YX} = r \cdot \frac{\sigma_Y}{\sigma_X} \quad \therefore \sigma_X = 3$$

$$\therefore \frac{4}{5} = \frac{6}{10} \cdot \frac{6y}{3} \Rightarrow 6y = \frac{4 \times 3 \times 10}{5 \times 6} = 4$$

(v) Standard Error of Estimates :

Standard error of estimate of Y :

$$\begin{aligned} S_{Y.X} &= 6_Y \sqrt{1-r^2} \\ &= 4 \sqrt{1-0.36} \\ &= 4 \times 0.80 \\ &= 3.2 \end{aligned}$$

and, standard error of estimate of X :

$$\begin{aligned} S_{x.y} &= 6_X \sqrt{1-r^2} \\ &= 3 \times 0.80 \\ &= 2.40 \end{aligned}$$

8.8 LET US SUM UP

In this lesson we have discussed the regression line of X on Y, some properties of regression coefficients and standard error of estimates. Some of the important relations are :

(i) The regression line of X on Y is given by

$$X - \bar{X} = b_{XY} (Y - \bar{Y})$$

$$\text{where} \quad b_{XY} = r \cdot \frac{\sigma_X}{\sigma_Y}$$

$$= \frac{\frac{1}{n} \sum XY - \bar{X}\bar{Y}}{\frac{1}{n} \sum X^2 - \bar{X}^2}$$

- (ii) The regression on line of Y on X is given by

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

where $b_{YX} = r \cdot \frac{\sigma_Y}{\sigma_X}$

$$= \frac{\frac{1}{n} \sum XY - \bar{X}\bar{Y}}{\frac{1}{n} \sum Y^2 - \bar{Y}^2}$$

- (iii) The standard error of estimates.

If $Y = a + bx$ is the line of Y on X.

$$S_{Y.X} = \sigma_Y \sqrt{1 - r^2}$$

and, if $X = a + bY$ is the line of X on Y.

$$S_{X.Y} = \sigma_X \sqrt{1 - r^2}$$

- (iv) Relation between correlation coefficient and regression coefficients

$$r = \pm \sqrt{b_{YX} \cdot b_{XY}}$$

and $\frac{b_{YX} + b_{XY}}{2} \geq r$

8.9 GLOSSARY

- Regression is one of the most popular technique in statistics which is used to predict the Value of one or more Variables if the Value of other one variable is given.

- X on Y : It means 'X' variable is dependent on 'Y' Variable
- Y on X: It means 'Y' variable is dependant on 'X' variable
- Linear regresson is used to find the value of one dependant variable from one independent Variable
- Multiple regressio is used to predict the value of one dependant Variable from many independant Variables.
- Regression Coefficient shows the predicting power of independent Variables.
- Standard error of estimates denotes the amount of value that a regression model fail to predict

8.10 SELF ASSESSMENT QUESTIONS

1. Explain the concept of regression and point out its importance in business forecasting.

2. What do you mean by standard error of estimate? Also mention its uses.

3. Explain the regression line of X on Y. Also describe its constants and how would you obtain these constants.

4. For 10 observations on price (X) and supply (Y), the following data is given

$$\sum X=130, \sum Y=220, \sum X^2=2288$$

$$\sum Y^2 = 5506 \text{ and } \sum XY = 3467$$

Obtain the line of regression of Y on X and estimate the supply when the price is 16 unit. Also find the correlation coefficient.

5. Explain the regression line of Y on X. Also describe its constants and how would you obtain these constants.

6. From the following data on yield and rainfall, estimate the yield when the rainfall is 22cm.

	Yield	Rainfall
Mean	508.40	26.70
S.D.	36.80	4.60

Coefficient of Correlation is 0.52

7. Giving the following data

X : 20, 24, 32, 38, 45, 48, 52

Y : 17, 21, 24, 25, 28, 30, 32

Find (i) two regression lines

(ii) the correlation coefficient

(iii) standard errors of the estimates.

8. From the following regression lines

$4x - y - 35 = 0$ and $9x - 4y - 135 = 0$ obtain the following :

- (i) Mean values of X and Y.
- (ii) Regression lines of Y on X and X on Y.
- (iii) Correlation coefficient between X and Y.
- (iv) Standard error of estimates.
- (v) Variance of X if S.D. (Y) = 3.

8.11 LESSON END EXERCISE

1. Standard error denotes_____
2. Dependent Variable is also called Predictor (True/ False)
3. Independent Variable helps to predict the value of dependent Variable (True/ False)
4. In regression , there is always a cause and effect relationship (True/False)
5. _____ is the father of regression technique.

8.12 SUGGESTED READINGS

1. Shenoy, G.V., Srivastava V.K., and Sharma, S.C. (1989). Business Statistics. Willy Eastern: New Delhi.
2. Simpson, G, and Kafka, F. Basic Statistics. Oxford and IBH Publishing; New Delhi.
3. Gupta, S.P. (1998). Statistical Methods, Sultan Chand and Sons; New Delhi.

8.13 ANSWERS TO CHECK YOUR PROGRESS

Answers I: B, B, C, B, B

Answers II: 1-D, 2-B, 3-D, 4-D

ASSOCIATION OF ATTRIBUTES**STRUCTURE**

- 9.0 Objectives
- 9.1 Introduction
- 9.2 Notation and Terminology
- 9.3 Class Frequencies
- 9.4 Contingency Table
- 9.5 Relation Between Class Frequencies
- 9.6 Consistency of Data
- 9.7 Independence of Attributes
- 9.8 Association of Attributes
- 9.9 Coefficient of Association
- 9.10 Let us sum up
- 9.11 Glossary
- 9.12 Self-Assessment Questions
- 9.13 Lesson End Exercise
- 9.14 Suggested Readings
- 9.15 Answers to check your progress

9.0 OBJECTIVES

On completion of this lesson, you will be able:

- o to understand the concept of attributes,

- o to familiar with the class frequency, ultimate class frequency,
- o to develop relationship between class frequencies,
- o to understand the concept of consistency of data,
- o to learn about independence and association of attributes, and
- o to know about coefficient of association.

9.1 INTRODUCTION

Literally, an attribute means a quality or characteristic. This lesson deals with qualitative characteristics which are not amenable to quantitative measurements and hence need slightly different statistical treatment from that of the variables. Examples of attributes are drinking, blindness, health, honesty, etc. An attribute may be marked by its presence (possession) or absence (dispossession) in a member of given population. That is, the qualitative characteristics such as Deafness, Blindness, Employment, Beauty, Hair Colour, Sex etc., of an individual of universe or population are termed as Attributes. The attributes are not orderable into series from least to most or vice-versa. If we observe any attribute in the population then the whole group is divided into two complementary classes. One class contains the members who possess the attribute and another class contains the members who do not possess the attributes. For example, according to the attribute "Blindness" the people of a particular city may be classified into two classes:

- (i) The class of Blind people.
- (ii) The class of non-blind people.

The classification which divides a group into two classes according to one attribute is called classification by Dichotomy or Simple Classification. The classification which divides a group into more than two classes according to one attribute is called manifold classification. For example, according to the attribute "Hair-Colour" the population of a city may be divided into different following classes:

- (i) Fair-Haired People
- (ii) Red-Haired People
- (iii) Brown-Haired People
- (iv) Black-Haired People

If several (more than two) attributes are noted, the process of classification may however, be continued indefinitely. Such type classification may be called classification as a series of dichotomies. For example, consider the two attributes namely "Blindness and Deafness". The people of a particular city may

be first divided into two classes according to the attribute "Blindness" and then each of these two classes may further be classified according to the attribute "Deafness". And therefore, ultimately, we have following four classes:

- (i) The class of blind and deaf people.
- (ii) The class of blind and non-deaf people.
- (iii) The class of non-blind and deaf people.
- (iv) The class of non-blind and non-deaf people.

9.2 NOTATION AND TERMINOLOGY

For the sake of simplicity and convenience it is imperative to use certain symbols to represent different classes and their frequencies. It is customary to use capital letters A and B to represent the presence of the attributes and the Greek letters (α) and (β) to represent absence of the attributes. Thus α = not A and β = not B. For example, If A represents males, then (α) would represent females. Similarly, if B represents literates then β would denote illiterates. The combination of the different attributes is denoted by (AB), ($A\beta$), (αB) and ($\alpha\beta$). Thus in this example, (AB) would mean number of literate males and ($\alpha\beta$) illiterate females. The number of observations in different classes is called, ‘‘class frequencies’’. Thus if the number of literate male is 50, the frequency of class (AB) is 50. Class frequencies are denoted by enclosing class notation in brackets like (AB), ($\alpha\beta$) etc. Thus (A) denotes number of individuals possessing attribute A.

(AB) denotes the number of individuals possessing attributes A and B.

($\alpha\beta$) denotes number of individuals, possessing attributed α and β .

Any letter or combination of letters like A, AB, $\alpha\beta$ etc., by means of which we specify the characters of the members of a class, may be termed as class symbol.

9.3 CLASS FREQUENCIES AND ULTIMATE CLASS FREQUENCIES

Class Frequencies : The number of observations assigned to any class is termed for the sake of brevity the frequency of the class or the ‘‘class frequency’’. Class frequencies are denoted by enclosing the corresponding class symbols in brackets. Thus (B) denotes the number of B’s i.e., objects possessing attribute B. ($A\beta$) the number of $A\beta$ ’s i.e., objects possessing attribute A but not B, and so on for any number of attributes.

The order of a class depends upon the number of attributes specified. A class having one attribute is known as the class of the first order, a class having two attributes as class of the second

order, and so on. The total number of observations denoted by the symbol N is called the frequency of the Zero order since no attributes are specified. Thus we have

N : frequency of the zero order

$\left. \begin{matrix} (A) & (B) \\ (\alpha) & (\beta) \end{matrix} \right\}$: frequency of the first order

$\left. \begin{matrix} (AB) & (\alpha B) \\ (A\beta) & (\alpha\beta) \end{matrix} \right\}$: frequency of the second order

In general, the following rules are used to determine the class frequencies:

1. with n attributes there are in all 2^n positive classes.
2. with n attributes the number of classes is 2^n . i.e. for one attribute, the frequencies are $2^1 = 2$ and for two attributes, the total frequencies are $2^2 = 4$. They are in the order $1+2+2 = 5$.

Ultimate Class Frequencies

It is cleared from above that every class frequency can be expressed in terms of the frequencies of the highest order, i.e., of order n . Any frequency can be analysed into highest frequencies and the process need stop only when we have reached the frequencies of the highest order. For example, with two attributes,

$\left. \begin{matrix} (A) = (AB) + (A\beta) \\ (\alpha) = (\alpha B) + (\alpha\beta) \end{matrix} \right\}$ ultimate class frequencies

The classes specified by n attributes, i.e., those of the highest order, are termed the ultimate class frequencies. A given data can be completely specified if only the ultimate class frequencies are given.

The total number of classes of ultimate order is determined by the formulas 2^n when n stands for the number of attributes studied. If two attributes are studied then the number of classes of ultimate order shall be $2^2 = 4$. In case three attributes are studied then there would be $2^3 = 8$ classes of the ultimate order.

The frequencies of the positive, negative and ultimate classes can be known from the following table which is known as contingency table.

From this table certain relationships can be described :

	A	α	
B	(AB)	(αB)	(B)
β	(A β)	($\alpha\beta$)	(β)
	(A)	(α)	N

From these relationships if we know any of the ultimate class frequencies and any other three values, we can find out the frequencies of the remaining classes.

9.4 CONTINGENCY TABLE

A table which represents the classification according to the distinct classes of two characteristics A and B is called a two-way contingency table.

Suppose the attribute A has m distinct classes denoted by A₁, A₂,, A_m and the attribute B has n distinct classes denoted by B₁, B₂,, B_n. Then there are in all m×n distinct classes (called cells) in the contingency table.

In the contingency table, the totals of various rows A₁, A₂ etc. and totals of various columns B₁, B₂..... etc., give the first order frequencies and cells have the frequencies of second order. The grand total of all frequencies gives the total number of observations i.e. N. The contingency table can be written as :

	Characteristic B						Total
	B ₁	B ₂		B _j		B _n	
A ₁	(A ₁ B ₁)	(A ₁ B ₂)		(A ₁ B _j)		(A ₁ B _n)	(A ₁)
A ₂	(A ₂ B ₁)	(A ₂ B ₂)		(A ₂ B _j)		(A ₂ B _n)	(A ₂)
⋮	⋮	⋮		⋮		⋮	⋮
⋮	⋮	⋮		⋮		⋮	⋮
A _i	(A _i B ₁)	(A _i B ₂)		(A _i B _j)		(A _i B _n)	(A _i)
⋮	⋮	⋮		⋮		⋮	⋮
⋮	⋮	⋮		⋮		⋮	⋮
⋮	⋮	⋮		⋮		⋮	⋮
A _m	(A _m B ₁)	(A _m B ₂)		(A _m B _j)		(A _m B _n)	(A _m)
Total	(B ₁)	(B ₂)		(B _j)		(B _n)	N

The classification by dichotomies with two attributes A and B is generally known as two by two contingency table. Such as with the help of 2×2 contingency table, we can find the ultimate class frequency from the positive class frequencies with two attributes. For illustration consider the following examples :

Attribute	B	β	Total
A	(AB)	(A β)	(A)
α	(α B)	($\alpha\beta$)	(α)
Total	(B)	(β)	N

Example 1. Given

$N = 300$, $(A) = 100$, $(B) = 120$, $(AB) = 40$ find the ultimate class frequencies.

Solution : Filling the given values in the table, the others are assumed by mere addition or subtraction.

	A	α	
B	(AB) 40	(α B) 80	(B) 120
β	(A β) 60	($\alpha\beta$) 120	(β) 180
	(A) 100	(α) 200	N 300

Example 2. From the following data find out the missing frequencies :

$(AB) = 100$, $(A) = 300$, $(N) = 1,000$, $(B) = 600$

Solution. Putting these values in the contingency table missing frequencies are (A β)

	A	α	
B	(AB) 100	(α B) 500	(B) 600
β	(A β) 200	($\alpha\beta$) 200	(β) 400
	(A) 300	(α) 700	N 1,000

(α B), ($\alpha\beta$), (α) and (β) i.e.

$$(A\beta) = (A) - (AB) = 300 - 100 = 200$$

$$(\alpha B) = (B) - (AB) = 600 - 100 = 500$$

$$(\beta) = N - (B) = 1000 - 600 = 400$$

$$(\alpha\beta) = (\beta) - (A\beta) = 400 - 200 = 200$$

$$(\alpha) = N - (A) = 1000 - 300 = 700$$

9.5 RELATION BETWEEN CLASS FREQUENCIES

If the population is classified into two classes A and α according to the attribute A, the number of all the individuals must be equal to the number of A's plus number of α 's i.e.

$$N = (A) + (\alpha) \text{ or } (\alpha) = N - (A) \text{ or } (A) = N - (\alpha)$$

Obviously, it is true for any attribute B, C etc. i.e.

$$\begin{aligned} N &= (B) + (\beta) \\ &= (C) + (\gamma) = \dots \end{aligned}$$

Similarly, the number of A's should equal to the number of A's that are B plus the number of A's that are β i.e.,

$$(A) = (AB) + (A\beta)$$

Similarly, we have

$$\begin{aligned} (\alpha) &= (\alpha B) + (\alpha\beta) \\ N &= (AB) + (A\beta) + (\alpha B) + (\alpha\beta) \end{aligned}$$

In the same way, we have

$$\begin{aligned} (AB) &= (ABC) + (AB\gamma) \\ (\beta A) &= (A\beta C) + (A\beta\gamma) \\ (\alpha B) &= (\alpha BC) + (\alpha B\gamma) \\ (\alpha\beta) &= (\alpha\beta C) + (\alpha\beta\gamma) \end{aligned}$$

$$\text{Hence } N = (ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma) + (\alpha BC) + (\alpha B\gamma) + (\alpha\beta C) + (\alpha\beta\gamma)$$

Such relations exist for any order of class frequencies and thus we conclude that the class frequencies

of order zero can be expressed in terms of class frequencies of order one, of order one in terms of order two and so on unless ultimate classes are used. It means that any class frequency can be expressed in terms of higher class frequency or every possible class frequency can be expressed as the sum of the ultimate class frequencies.

Example 3. Express non-positive class frequencies in terms of positive class frequencies in case of two attributes.

Solution : Writing

$$(A) = N.A \text{ and}$$

$$(\alpha) = \alpha.N$$

$$\text{We have } (A)+(\alpha) = AN+\alpha N$$

$$\text{or } N = (A+\alpha)N$$

$$\text{or } A+\alpha = 1 \text{ or } \alpha = 1-A$$

$$\text{or } A = 1-\alpha$$

Obviously, it is also true for any other attribute.

(i) Now let us consider two attributes A and B, then we have

$$(A\beta) = A\beta N = A(1-B)N = AN-ABN$$

$$= (A)-(AB)$$

$$(\alpha B) = \alpha NB$$

$$= (1-A)BN = NB-ABN = (B)-(AB)$$

$$(\alpha\beta) = \alpha\beta N$$

$$= (1-A)(1-B)N$$

$$= (1-A-B+AB)N = 1\cdot N - A\cdot N - B\cdot N + AB\cdot N$$

$$= N-(A)-(B)+(AB)$$

Example 4. From the following data find out missing frequencies :

$$N=1500, (A)=383, (\beta)=1140 \text{ and } (\alpha\beta)=792$$

Solution. Putting these values in the following contingency table; the

	A	α	Total
B	(AB)= 35	(B α)= 325	(B)= 360
β	(A β)= 348	($\alpha\beta$)= 792	(β)= 1140
Total	(A)= 383	(α)= 1117	N= 1500

Others are :

$$(\beta) = N - (\beta) = 1500 - 1140 = 360$$

$$(\alpha) = N - (A) = 1500 - 383 = 1117$$

$$(AB) = (\beta) - (\alpha\beta) = 1140 - 792 = 348$$

$$(A\beta) = (A) - (A\beta) = 383 - 348 = 35$$

$$(\alpha\beta) = (\alpha) - (\alpha\beta) = 1117 - 792 = 325$$

CHECK YOUR PROGRESS-I

ANSWER THE BELOW QUESTIONS:

Q-1 Explain association of attributes along with examples.

Q-2 Find the missing frequencies for the following data and check its consistency. $N = 240$, $(AB) = 80$, $(?) = 112$, $(B) = 150$. Also find the type of association between the following pairs of attributes using Proportion Method:

(i) A and B

(ii) A and ?

(iii) ? and B

(iv) ? and ?.

Q-3 In an examination at which 600 candidates appeared, boys outnumbered girls by 10 percent of all candidates. Number of passed candidates exceeded the number of failed candidates by 200. Girls failing in the examination numbered 90. Construct the nine-square table and calculate the coefficient of association between boys and success in the examination.

9.6 CONSISTENCY OF DATA

In order to find out whether the given data are consistent or not we have to apply a very simple test. The test is to find out whether any one or more of the ultimate class frequencies is negative or not. If none of the class- frequencies is negative we can safely calculate that the given data are consistent (i.e., the frequencies do not conflict in any way with each other). On the other hand, if any of the ultimate class frequencies comes out to be negative the given data are inconsistent. Thus, the necessary and sufficient condition for the consistency of a set of independent class frequencies is that no ultimate class frequency is negative.

Example 6. From the following two cases find out whether the data are consistent or not

Case-I : $(A) = 100, (B) = 150, (AB) = 60, N = 500$

Case-II : $(A) = 100, (B) = 150, (AB) = 140, N = 500$

Solution : Case - I

	A	α	
B	(AB) 60	(αB)	(B) 150
β	(A β)	($\alpha\beta$)	(β)
	(A) 100	(α)	N 500

We are given

$(A) = 100, (B) = 150, (AB) = 60, N = 500$

Substitute all these values in the table, then from the table $(A) = (A) - (AB) = 100 - 60 = 40$

$(A\beta) = (A) - (AB) = 100 - 60 = 40$

$(\alpha B) = (B) - (AB) = 150 - 60 = 90$

$(\alpha\beta) = (\alpha) - (\alpha B) = 400 - 90 = 310$

Since all the ultimate class frequencies are positive, we conclude that the given data are consistent.

Case-II :

Given values are

$$(A) = 100, (B) = 150, (AB) = 140, (N) = 500$$

By putting these values in the nine-square table we can determine the missing value :

	A	α	
B	(AB) 140	(αB) 10	(B) 150
β	(A β) -40	($\alpha\beta$) 390	(β) 350
	(A) 100	(α) 400	N 500

From the table

$$\begin{aligned} (A\beta) &= (A) - (AB) \\ &= 100 - 140 = -40 \\ (\alpha B) &= (B) - (AB) \\ &= 150 - 140 = 10 \\ (\alpha\beta) &= (\alpha) - (\alpha B) \\ &= 400 - 10 = 390 \end{aligned}$$

Thus one of the ultimate class frequencies i.e., (A β) is negative and hence the given data are inconsistent.

9.7 INDEPENDENCE OF ATTRIBUTES

Two attributes A and B are said to be Independent if there exists no relationship of any kind between them and we may expect to find the same proportion of A's among B's as amongst β 's i.e.,

$$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} \quad \dots (1)$$

For example,

9.7.1 If A denotes "proficiency in Statistics" and β denotes "proficiency in cooking" then we naturally expect that proportion of "cooks among A" and should be equal i.e.,

$$\frac{(AB)}{(A)} = \frac{(\alpha\beta)}{(\alpha)}$$

9.7.2 If we consider that there is no relationship between sex of the newly born child and the waxing of the moon, we may anticipate that the proportion of male births (A) amongst the total births when the moon is waxing (B) should be equal to the proportion of male-births amongst the total births when the moon is not waxing (β).

9.7.3 Suppose $N = 100$, $(A) = 60$, $(B) = 40$ and $(AB) = 24$ where A denotes 'intelligence' and B denotes 'richness' then proportion of intelligent people amongst the rich is equal to

$$\frac{AB}{B} = \frac{24}{40} = \frac{3}{5}$$

and proportion of intelligent people amongst the non-rich is

$$\frac{A\beta}{\beta} = \frac{(A) - AB}{N - B} = \frac{60 - 24}{100 - 40} = \frac{36}{60} = \frac{3}{5}$$

Hence, we note that the two proportions are equal which indicates that intelligence does not have any relation with richness and the two qualities are independent.

9.7.4 Similarly, if blindness and deafness have nothing to do with one another, the proportion of blind people amongst the deaf and amongst the non-deaf must be equal.

9.8 ASSOCIATION OF ATTRIBUTES

If the criterion of Independence $\frac{AB}{B} = \frac{A\beta}{\beta}$ holds true then the following criterion :

(i) The proportion of α 's is the same in B's as in β ,

$$i.e. \quad \frac{\alpha B}{B} = \frac{\alpha\beta}{\beta} \quad \dots (1)$$

(ii) The proportion of B's is the same in A's as in α 's

$$i.e. \quad \frac{AB}{A} = \frac{\alpha B}{\alpha} \quad \dots (2)$$

(iii) The proportion of β 's is the same in A's as in α 's

$$\text{i.e.} \quad \frac{A\beta}{A} = \frac{\alpha\beta}{\alpha} \quad \dots (3)$$

(iv) The proportion of A's is the same in B's as in N

$$\text{i.e.} \quad \frac{AB}{B} = \frac{A}{N} \quad \dots (4)$$

This relation may also be expressed in any of the following forms :

$$\frac{AB}{A} = \frac{B}{N} \quad \dots (5)$$

$$\text{or} \quad (AB) = \quad \dots (6)$$

$$\text{or} \quad \frac{(AB)}{N} = \frac{(A)}{N} \times \frac{(B)}{N} \quad \dots (7)$$

(v) The proportion of A's is the same in β 's as in the population at large

$$\text{i.e.} \quad \frac{(A\beta)}{\beta} = \frac{(A)}{N} \quad \dots (8)$$

(vi) The proportion of α 's is the same in β 's as in N

$$\text{i.e.} \quad \frac{(\alpha\beta)}{(\beta)} = \frac{(\alpha)}{N} \quad \dots (9)$$

(vii) The proportion of α 's is the same in B's as in N

$$\frac{(\alpha B)}{(B)} = \frac{(\alpha)}{N} \quad \dots (10)$$

$$\text{(viii)} \quad (AB)(\alpha\beta) = (\alpha B)(A\beta) \quad \dots (11)$$

$$\text{or} \quad \frac{(AB)}{(\alpha B)} = \frac{(A\beta)}{(\alpha \beta)}$$

which means, in words, the ratio of A's to α 's amongst B's is equal to the ratio of A's to α 's amongst the β 's

Similarly $\frac{(AB)}{(A\beta)} = \frac{(\alpha B)}{(\alpha\beta)}$ means the ratio of B's to β 's amongst A's is equal to the ratio of B's to β 's amongst the α 's.

These relations may be understood easily with the help of the following 2x2 contingency table.

Attribute	A	α	Total
B	(AB)	(αB)	(B)
β	(A β)	($\alpha\beta$)	(β)
Total	(A)	(α)	N

When comparison of observed and expected frequencies method is applied, the actual observation is compared with the expectation. If the actual observation is equal to the expectation the attributes are said to be independent. If the actual observation is more than the expectation, the attributes are said to be positively associated and if the actual observation is less than the expectation, the attributes are said to be negatively associated.

Symbolically, Attributes A and B are :

- (i) Independent if $E(AB) = \frac{(A) \times (B)}{N}$
- (ii) Positively associated if $E(AB) > \frac{(A) \times (B)}{N}$
- (iii) Negatively associated if $E(AB) < \frac{(A) \times (B)}{N}$

The same is true for attributes α and B; α and β and A and β . Thus, attributes α and β shall be called

- (i) Independent, if $(\alpha\beta) = \frac{(\alpha) \times (\beta)}{N}$
- (ii) Positively associated, if $(\alpha\beta) > \left(\frac{(\alpha) \times (\beta)}{N}\right)$

and (iii) Negatively associated, if $(\alpha\beta) < \frac{(\alpha) \times (\beta)}{N}$

Example 7. As from the following data find out whether attributes (i) (AB), (ii) (Aβ), (iii) (αB) and (iv) (αβ) are independent, associated or disassociated

$$N = 100, (A) = 40, (B) = 80 \text{ and } (AB) = 30$$

Solution : (i) Apply the criterion of Independence, i.e. attribute (AB) shall be called independent if

$$(AB) = \frac{(A) \times (B)}{N}$$

$$\text{Positively associated if } (AB) > \frac{(A) \times (B)}{N}$$

and Negatively associated if $(AB) < \frac{(A) \times (B)}{N}$

$$\text{Expectation of } (AB) = \frac{(A) \times (B)}{N}$$

$$\text{Here } (A) = 40, (B) = 80, N = 100$$

$$\text{Expectation of } (AB) = \frac{40 \times 80}{100} = 32$$

	A	α	
B	(AB) 30	(αB) 50	(B) 80
β	(Aβ) 10	(αβ) 10	(β) 20
	(A) 40	(α) 60	N 100

The actual observations [i.e., the given value of (AB), i.e. 30] is less than the expectation and hence the attributes are disassociated or Negatively associated.

(ii) From the above table

$$(A\beta) = 10, (\alpha B) = 50, (\alpha\beta) = 10, (\alpha) = 60, (\beta) = 20$$

Attributes A and β shall be independent if

$$(A\beta) = \frac{(A) \times (\beta)}{N}$$

$$\text{Expectation of } (A\beta) = \frac{40 \times 20}{100} = 8$$

Thus the actual observation [i.e., $(A\beta) = 10$] is more than expectation and hence the attributes A and β are positively associated.

(iii) Attributes α and B shall be called independent if

$$(\alpha B) = \frac{(\alpha) \times (B)}{N}$$

$$\text{Expectation of } (\alpha B) = \frac{(\alpha) \times (B)}{N}$$

where $(\alpha) = 60, (B) = 80$ and $N = 100$

$$= \frac{60 \times 80}{100} = 48$$

Thus actual observation [$(\alpha B) = 50$] is more than the expectation and hence the attributes are positively associated.

(iv) Attributes α and β shall be called independent if

$$(\alpha\beta) = \frac{(\alpha) \times (\beta)}{N}$$

$$\text{Expectation of } (\alpha\beta) = \frac{(\alpha) \times (\beta)}{N}, \text{ where } (\alpha) = 60, (\beta) = 20, N = 100$$

$$= \frac{60 \times 20}{100} = 12$$

Thus actual observation [$(\alpha\beta) = 10$] is less than the expectation. Hence the attributes are disassociated.

9.9 COEFFICIENT OF ASSOCIATION

The most popular method of studying association is the Yule's Coefficient because here not only we can determine the nature of association i.e., whether the attributes are positively associated, negatively associated or independent, but also the degree or extent to which the two attributes are associated. The Yule's coefficient is denoted by the symbol Q and is obtained by applying the following formulae:

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

The value of this coefficient lies between -1 and $+1$. when the value of $Q=+1$ there is perfect positive association between the attributes, when $Q=-1$ there is perfect negative association (or perfect disassociation) between the attributes and when the value of Q is zero then the two attributes are independent.

The coefficients of association can be used to compare the intensity of association between two attributes with the intensity of association between two other attributes.

Example 8. Investigate the association between eye colour of husbands and eye colour of wives from the data given below:

Husbands with light eyes and wives with light eyes = 309
Husbands with light eyes and wives with not light eyes = 214
Husbands with not light eyes and wives with light eyes = 132
Husbands with not light eyes and wives with not light eyes = 119

Solution. Since we have to find out the association between eye colour of husband and that of wife, one attribute we would take as A and other as B ,

Let A denote husbands with light eyes

$\therefore \alpha$ would denote husbands with not light eyes.

Let B denote wives with light eyes.

$\therefore \beta$ denote wives with not light eyes.

$\therefore$ the given data in terms of these symbols are :

$(AB) = 309, (A\beta) = 214, (\alpha B) = 132$ and $(\alpha\beta) = 119$

Applying the Yule's method :

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

Substituting the above values in the formula we have

$$Q = \frac{(309)(119) - (214)(132)}{(309)(119) + (214)(132)} = 0.131$$

Thus, there is a very little association between the eye colour of husband and wife.

Example 9. Eighty-Eight residents of an Indian City, who were interviewed during a sample survey, are classified below according to their smoking and tea drinking habits. Calculate the Yule's Coefficient of association and comment on its value.

	Smokers	Non-Smokers
Tea drinkers	40	33
Non-tea drinkers	3	12

Solution : Let A denote smokers

$\therefore$ α would denote non-smokers

Let B denote tea-drinkers

$\therefore$ β would denote non-tea drinkers.

The given data in terms of these symbols are

(AB) i.e. Number of Smokers and tea drinkers = 40

(A β) i.e. Number of smokers and non-tea drinkers = 3

(α B) i.e. Number of non-smokers and tea drinkers = 33

($\alpha\beta$) i.e. Number of non-smokers and non tea drinkers = 12

$\therefore$ Applying Yule's method :

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

Substituting the values of (AB), (Aβ), (αB) and (αβ) in the formula

$$Q = \frac{(40 \times 12) - (3 \times 33)}{(40 \times 12) + (3 \times 33)} = 0.658$$

This shows that the attributes tea drinking and smoking are positively associated.

Example 10. Prepare a 2×2 table from the following information, calculate Yule's Coefficient of Association and interpret the result N = 1500, () = 1117, (B) = 360 (AB) = 35

Solution :

A	α		
B	(AB) 35	(αB) 325	(B) 360
β	(Aβ) 348	(αβ) 792	(β) 1140
	(A) 383	(α) 1117	N 1500

By putting the known values in the contingency table, we can find out the unknown values.

Thus (A) = N – (α) = 1500 – 1117 = 383

(Aβ) = (A) – (AB) = 383 – 35 = 348

Yule's coefficient of Association

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} = \frac{(35 \times 792) - (348 \times 325)}{(35 \times 792) + (348 \times 325)}$$

Example 11. Find the association between literacy and unemployment from the following figures :

Total adults : 10,000

Literates : 1,290

Unemployed : 1,390

Literate Unemployed : 820

Comment on the results.

Solution :

	A	α	
B	(AB)	(α B)	(B)
820	570	1390	
β	(A β)	($\alpha\beta$)	(β)
	470	8140	8610
	(A)	(α)	N
	1290	8710	10,000

Let A denotes literates

$\therefore$ α denotes illiterates

Let B denotes Unemployed

$\therefore$ β will denote employed

We are given

$$(A) = 1290, (B) = 1390$$

$$(AB) = 820, N = 10,000$$

Putting these value in the 2 $\times$ 2 contingency table and find missing frequencies.

$$\therefore Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$= \frac{(820 \times 8140) - (470 \times 570)}{(820 \times 8140) + (470 \times 570)} = 0.923$$

There is a high degree of positive association between literacy and unemployment.

9.10 LET US SUM UP

In this lesson we have discussed:

Attributes	:	Qualitative characteristic are termed as attributes.
Positive Attributes	:	The presence of attributes is called positive attributes. Positive attributes are denoted by the capital letters. Such as A, B, C etc.
Negative Attributes	:	The absence of a particular attribute is called negative attribute. These attributes are denoted by α, β, γ etc.
Class Frequency	:	The number of observations assigned to any class is called class frequency. Written by enclosing the class-symbols in brackets. Such as (A), (AB) etc.
Frequency of rth order	:	A class specified by r attributes is called the class of rth order and its frequency is called rth order frequency.
Ultimate Class Frequency	:	The class specified by n attributes i.e. those of the highest order, are ultimate class frequency.
Consistency	:	A set of class frequencies is said to be consistent if all its class frequencies confirm with one another and do not have any mutual contradiction.
Inconsistency	:	A set of class frequencies in which the given class frequencies do not confirm with another but provide contradictory statement of any form is called inconsistent.

9.11 GLOSSARY

- Attribute is a term used to indicate those items observation/data/ which is not quantified but qualitative in nature.
- There are many features of population that need this technique rather than correlation or regression analysis.

- Attributes usually measured as 'present' or 'absent' from the population.
- In case of attributes all capital letters of English are used as presence of attributes whereas Greek letters denote the absence of an attribute e.g. A for blind for not blind B for female and for non-female etc.

9.12 SELF-ASSESSMENT QUESTIONS

- Given the following ultimate class frequencies, find the frequencies of the +ve and -ve classes and total number of observations.

 $(AB) = 250, (A) = 120, (B) = 200, () = 70$
- Is there any inconsistency in the data given below:

 $= 1000, (A) = 150, (B) = 300, (AB) = 200, N = 100, (A) = 50, (B) = 60, (AB) = 20$
- Find if A and B are independent, positively associated or negatively associated from the data given below :

 $(A) = 470, (B) = 620, (AB) = 320, N = 1000$
- A teacher examined 280 students in Economics and Auditing and found that 160 failed in Economics, 140 failed in Auditing and 80 failed in both the subjects. Is there any association between failure in Economics and Auditing?
- Show by Short cut method whether there is disassociation or positive or negative association in the following attributes A and B.

 - $(A) = 470, (B) = 620, (AB) = 320$
 - $(AB) = 294, () = 570, () = 380$
 - $(B) = 768, (A) = 480, (B) = 145$
- In a group of 800 students, the number of married is 320. But of 240 students who failed, 96 belonged to the married group. Find out whether the attributes marriage and failure are independent.
- The male population of U.P is 250 lakhs. The number of literate males is 20 lakhs and the total number of criminals is 26 thousand. The number of literate male criminals is 2 thousand. Do you find any association between literacy and criminality.

9.13 LESSON END EXERCISE

1. If A denotes blind then denotes _____
2. Association Attributes is qualitative technique to find the presence or absence of qualities in data (True/ Falsa)
3. Ultimate class frequencies are: -
 - a) Highest order class frequencies.
 - b) Maximum number of frequencies
 - c) Minimum number of class frequencies
 - d) None of the these
4. The table that shows various possible combinations of two attributes is known as table.
5. $A =$
 - a) $(AB) + (A)$
 - b) $() + (AB)$
 - c) $(A+B)$
 - d) $(AB) (A)$

9.14 SUGGESTED READINGS

1. Argyrous, George. 1997. Statistics for Social Research. New York: Mc Millan Press Ltd.
2. Goods, W.J. & Hatt, P.K. 1981. Methods in Social Research. New York: Mc Graw Hill.
3. Gupta, S.C. 1981. Fundamentals of Statistics. Bombay: Himalayan Publishing House.
4. Gupta, S.P. 2004. Statistical Methods. New Delhi: Sultan Chand and Sons.

9.15 ANSWERS TO CHECK YOUR PROGRESS

Ans1: Association of Attributes:

1. It refers to the relationship between two or more qualitative variables.
2. If the presence or absence of one attribute affects the presence or absence of another attribute, the attributes are said to be associated.
3. If there is no such effect, the attributes are independent.

Example : Consider two attributes: Gender (Male, Female) and Preference for a product (Like, Dislike).

Gender	Like	Dislike	Total
Male	40	10	50
Female	30	20	50
Total	70	30	100

Here, we can study whether gender is associated with product preference.

Ans2: Total N = 240

Frequency of (AB) = 80

Frequency of (?) = 112

Frequency of (B) = 150

Assuming the attributes are A, ? (complement of A), B, and ? (complement of B).

Let the frequencies be:

$$(A) = x$$

$$(\beta) = N - (B) = 240 - 150 = 90$$

Since (AB) = 80, and (B) = 150, the frequency of $(A\beta) = (A) - (AB) = x - 80$

Similarly, $(\alpha B) = (B) - (AB) = 150 - 80 = 70$

$$\text{Also, } (\alpha\beta) = N - [(AB) + (A\beta) + (\alpha B)] = 240 - [80 + (x - 80) + 70] = 240 - (x + 70) = 170 - x$$

$$\text{Given } (\alpha) = 112, \text{ and } (\alpha) = (\alpha B) + (\alpha\beta) = 70 + (170 - x) = 240 - x$$

$$\text{So, } 240 - x = 112 \Rightarrow x = 128$$

Therefore:

$$(A) = 128$$

$$(A\beta) = 128 - 80 = 48$$

$$(\alpha\beta) = 170 - 128 = 42$$

Check consistency:

Sum of all frequencies = $(AB) + (A\beta) + (\alpha B) + (\alpha\beta) = 80 + 48 + 70 + 42 = 240$ (matches total N)

Proportion Method:

Calculate proportions for each pair:

For A and B:

$$P(AB) = 80/240 = 0.3333$$

$$P(A) = 128/240 = 0.5333$$

$$P(B) = 150/240 = 0.625$$

Check if $P(AB) = P(A) * P(B) = 0.5333 * 0.625 = 0.3333$ (approximately equal)

Hence, A and B are independent (no association).

For A and β :

$$P(A\beta) = 48/240 = 0.2$$

$$P(A) = 0.5333$$

$$P(\beta) = 90/240 = 0.375$$

$$P(A)*P(\beta) = 0.5333 * 0.375 = 0.2 \text{ (equal)}$$

Hence, A and β are independent.

For α and B:

$$P(\alpha B) = 70/240 = 0.2917$$

$$P(\alpha) = 112/240 = 0.4667$$

$$P(B) = 0.625$$

$$P(\alpha)*P(B) = 0.4667 * 0.625 = 0.2917 \text{ (equal)}$$

Hence, α and B are independent.

For α and β :

$$P(\alpha\beta) = 42/240 = 0.175$$

$$P(\alpha) = 0.4667$$

$$P(\beta) = 0.375$$

$$P(\alpha)*P(\beta) = 0.4667 * 0.375 = 0.175 \text{ (equal)}$$

Hence, α and β are independent.

Conclusion: All pairs are independent; no association exists.

Ans3: Total candidates, $N = 600$

Boys outnumber girls by 10% of total candidates

Passed candidates exceed failed candidates by 200

Girls failing = 90

Step 1: Find number of boys and girls

Let number of girls = G

Then number of boys = $B = G + 10\% \text{ of } 600 = G + 60$

Total candidates: $B + G = 600$

$$\Rightarrow G + 60 + G = 600$$

$$\Rightarrow 2G = 540$$

$$\Rightarrow G = 270$$

$$\Rightarrow B = 270 + 60 = 330$$

Step 2: Find number of passed and failed candidates

Let number of passed candidates = P

Number of failed candidates = F

Given: $P - F = 200$

Also, $P + F = 600$

Solving:

$$P - F = 200$$

$$P + F = 600$$

$$\text{Adding: } 2P = 800 \Rightarrow P = 400$$

$$\text{Then } F = 600 - 400 = 200$$

Step 3: Construct nine-square table

We have two attributes: Gender (Boys, Girls) and Result (Pass, Fail)

Pass (P)	Fail (F)	Total
Boys (B)	x y	330
Girls (G)	400 - x 90	270
Total	400 200	600

$$\text{Since girls failing} = 90, \text{ girls passing} = 270 - 90 = 180$$

$$\text{Let boys passing} = x$$

$$\text{Then boys failing} = 330 - x$$

$$\text{Total passing} = 400$$

$$\Rightarrow x + 180 = 400$$

$$\Rightarrow x = 220$$

$$\text{Boys failing} = 330 - 220 = 110$$

Final nine-square table:

Pass (P)	Fail (F)	Total
Boys (B)	220 110	330
Girls (G)	180 90	270
Total	400 200	600

Step 4: Calculate coefficient of association (Q) between Boys and Success

Coefficient of association formula:

$$Q = \frac{ad - bc}{ad + bc}$$

Where,

a = number of boys passed = 220

b = number of boys failed = 110

c = number of girls passed = 180

d = number of girls failed = 90

Calculate numerator:

$$ad - bc = (220)(90) - (110)(180) = 19800 - 19800 = 0$$

Calculate denominator:

$$ad + bc = 19800 + 19800 = 39600$$

Therefore,

$$Q = \frac{0}{39600}$$

$$Q = 0$$

Interpretation: Coefficient of association is 0, indicating no association between gender (boys/girls) and success (pass/fail) in this examination.

MEANING AND METHODS OF SAMPLING**STRUCTURE**

- 10.0 Objectives
- 10.1 Introduction
- 10.2 Meaning of Sampling
- 10.3 Purposes of Sampling
- 10.4 Principles of Sampling
- 10.5 Method of Sampling
 - 10.5.1 Probability Sampling
 - 10.5.2 Non-Probability Sampling
 - 10.5.3 Types of Probability Sampling
 - 10.5.4 Types of Non -Probability Sampling
- 10.6 Bias in selecting informants in Non-Probability Sampling
- 10.7 Sample Size
- 10.8 Let us sum up
- 10.9 Glossary
- 10.10 Self-Assessment Questions
- 10.11 Lesson End Exercise
- 10.12 Suggested Readings
- 10.13 Answers to check your progress

10.0 OBJECTIVES

After going thoroughly this chapter, you should be able to: -

- Understand the meaning of Sampling
- Need of Sampling
- Principles of sampling
- Methods of sampling
- Limitation of Non probability Sampling
- Understand Sample size.

10.1 INTRODUCTION

In the last few chapters various quantitative and qualitative techniques of statistics have been discussed. In this chapter we will discuss the concept of sampling and various techniques of sampling. Sampling is backbone of any statistical research study. Because the quality of information in handle purely depends upon the appropriate sampling technique. Besides sampling, census is also a data collection technique but this technique looks little impossible to apply in all kind of research works due to various constraints.

10.2 MEANING OF SAMPLING

While conducting a survey, a question is usually asked: “Should all people be studied or only a limited number of persons drawn from the total population be studied and then extend our findings about the sample to the entire population? ‘Population’ refers to, all those people with the characteristics which the researcher wants to study within the context of a particular research problem.” A population could be all students in the college, all patients in the hospital, all prisoners in the prison, all customers in a big departmental store, all users of a particular model of car, all households in the village, all workers in the factory, all cultivators using the water of a particular canal in the settlement area for irrigational purposes, all victims of a natural disaster in a particular area and so on. When the population is relatively large and is physically not accessible, researchers survey only a sample A sample is a portion of people drawn from a larger population. It will be representative of the population only if it has same basic characteristics of the population from which it is drawn. Thus, our concern in sampling is not about what types of units (persons) will be interviewed/observed but with how many units of what particular description and by what method should be chosen. Suppose a large number of thefts are reported in one week in one area three kilometers long in a city. The area consists of seven sectors, each sector consisting seven lanes, each lane having 15

houses on the front side and 15 on the back side. Thus, the city will have 1500 households. It is planned to find out if all households in this area will support a community watch programme in which each household would take the responsibility for deputing one male member for performing the night-watch duty. Have all 1500 households to be included in finding out whether the scheme will be acceptable to the people, or only a sample of people from each of the seven sectors will be enough to get the idea? The answer to this question, whether all people or just a sample need to be studied in a survey depends on five factors: -

1. How quickly are data needed?
2. What type of survey is planned? Will it be a telephone survey, or a self-administered questionnaire sent by post or through an investigator or will it be a schedule in which answers to questions one to be filled in by the investigator himself?
3. What one the available resources? Is there money to appoint an investigator and to get the questionnaire printed/cyclostyled? Do all people have telephone?
4. How credible will the findings be? In the above example, even if 70 to 80% households agreed to participate in the sentative of the neighbourhood. If only 30 to 40% wanted to participate, it would be preferable to scrap such survey.
5. How familiar is the researcher with sampling methods?

According to Manheim, “a sample is a part of the population which is studied in order to make inferences about the whole population.” In defining population, from which the sample is taken, it is necessary to identify ‘target population’ and ‘Sampling frame’. The target population is one which includes all the units for which the information is required e.g. drug abuser students in one university, or voters in one village/constituency and so on. In defining the population, the criteria need to be specified for explaining cases which are included or excluded. For example, for studying the level of awareness of rights among women in one village community, the target population is defined as all women—married and unmarried—in the age group of 18-50 yrs. If the unit is an institution (say a university) than the type of its structure size as measured by the number of students in school section, college section and in professional courses, the number of teachers and employees needs to be specified.

For making the target population operational, the sampling frame needs to be constructed. This denotes the set of all cases from which the sample is actually selected. It should be noted that sampling frame is not a sample; rather it is the operational definition of the population that provides the basis for sampling. For example, in the above example of university, if students studying in school upto 12th and in college up to M.A./MSc. are excluded only students of professional courses are left out from which the sample is to be drawn. Thus, the sample frame reduces the number of total population and gives us the target population.

Bailey has said that the experienced researcher's always start from the top (population and work down to bottom (sample) i.e., they get a clear picture of the population before selecting the sample. The novice researchers, on the other hand, often work from the bottom up. Instead of making the population, they wish to study explicit, they select a predetermined number of conveniently available cases and assume that the sample corresponds to the population under study. For example, in exit polls, seeking randomly the opinion of the voters as to whom they voted, soon after their casting the votes in a few selected constituencies in selected cities and villages cannot be representatives of all voters. No wonder, the predictions of such exit polls do not come true.

10.3 PURPOSES OF SAMPLING

A large population cannot be studied in its entirety for reasons of size, time, cost or inaccessibility. Limited time, lack of large amount of funds and population scattered in a very wide geographical area often make sampling necessary. Sarantakos has pointed out the following purposes of sampling: -

1. Population in many cases may be so large and scattered that a complete coverage may not be possible. Suppose, the Maruti Udyog Co. wished to find out the reactions of purchasers of five-seater and eight-seater Maruti vans.

For these thousands of van purchasers would have to be contacted in different cities. Some of these would even be inaccessible and it would be impossible to contact all the van purchasers within a short time.

2. It offers a high degree of accuracy because it deals with a small number of persons. Most of us have had blood samples taken, sometimes from the fingers and sometimes from the arm or another part of the body. The assumption is that the blood is sufficiently similar throughout the body and the characteristics of the blood are determined on the basis of sample. Singleton and Straits have also said that studying all cases will describe population less accurately than a small sample.

3. In a short period of time valid and comparable results can be obtained. A lengthy period of data collection generally renders some data obsolete by the time the information is completely in hands.

For example, collecting information on the attitudes of the military personnels about non-availability of vehicles to be used in a very cold areas during the Kargil War, or Voter's performances during election period, or demanding action against police persons in a lockup blind. Besides, opinions expressed at the time of incidence and those expressed after a few months are bound to be different. The findings are thus bound to be influenced if long period is involved in data collection i.e., not taking a small sample but studying the entire population.

4. Sampling is less demanding in terms of requirements of investigators since it requires a small portion of the target population.

5. It is economical since it contains fewer people. Large population would involve employing a large number of interviewers which will increase the total cost of the survey.
6. Many research projects, particularly those in quality control testing require the destruction of the items being tested. If the manufacturer of electric bulbs wishes to find out whether each bulb met a specific standard, there would be no product left after the testing.

One important objective of sampling is to draw inference about the universe which is unknown from the unit which is observed or measured. Such inferring generalization made in Sociology is called 'Sociological inference' while one made in Statistics is called 'statistical inference'. Generalizations based on statistical inference always are probability statements and are never statements of absolute certainty. Sociological inference may be either valid or invalid. It may involve either deduction

or induction. Induction is generalisation from individual or specific instance to general principle. Deduction is generalisation from general principles to specific or particular instances. In this process, the generalisation is from sample to universe.

Two other purposes of sampling may also be specified here: -

- (a) seeking representativeness and thereby studying a small population instead of very large population.
- (b) analysing data where (i) cross-tabulation is required (ii) certain variables are to be controlled and (iii) phenomenon is to be observed under certain specific conditions.

CHECK YOUR PROGRESS-I

Answer the below questions:

1. What is meant by Sample Method? What are its uses?

2. Define Census and Sampling methods. Or, what are Census and Sampling Methods?

3.	Name different methods of Sampling.
4.	Give three reasons why sample surveys are preferred.
5.	Discuss how you would use the lottery method to select three students out of ten in your class.
6.	Explain the procedure of selecting a random sample of 3 students out of 10 students in your class by using random number tables.

10.4 PRINCIPLES OF SAMPLING

The main principle behind sampling is that we seek knowledge about the total units (population) by observing a few units (sample) and extend our inference about the sample to the entire population. For purchasing a bag of wheat, if we take out a small sample from the middle of the bag with a cutter, it will give us the inference whether the wheat in the bag is good or not. But it is not necessary that study of sample will always give us the correct picture of the total population. If in a class of 100 students we take out any five students at random and per chance find that all the 5 students are third divisioners, it would not mean that all remaining students in the class will be third divisioners. If few people in a village are found in favour of family planning, it would not mean that all people in the village will necessarily have the same opinion. The opinion may vary in terms of religion, educational level, age, economic status and such other factors. The

wrong inference is drawn or generalisation is made from the study of few persons because they constitute inadequate sample of the total population.

The study of sample becomes necessary because study of a very large population would require a long period of time, a large number of interviewers, a large amount of money, and doubtful accuracy of data collected by numerous investigators. The planning of observation/study with a sample is more manageable.

The important principles of sampling are: -

1. Sample units must be chosen in a systematic and objective manner.
2. Sample units must be independent of each other.
3. Sample units must be clearly defined and easily identifiable.
4. Same units of sample should be used throughout the study.
5. The selection process should be based on sound criteria and should avoid errors, bias and distortions.

Advantages of Sampling. The advantages of sampling are: -

1. It is not possible to study large number of people scattered in wide geographical area. Sampling will reduce their number.
2. It saves time and money.
3. It saves destruction of units.
4. It increases accuracy of data (having control on the small number of subjects).
5. It achieves greater response rate.
6. It achieves greater cooperation from respondents.
7. It is easy to supervise few interviewers in the sample but difficult to supervise a very large number of interviewers in the study of total population.
8. The researcher can keep a low profile.

10.5 METHOD OF SAMPLING

There are basically two types of sampling.

Probability sampling and non-probability sampling. Probability sampling is one in which every unit of the population has an equal probability of being selected for the sample. It offers a high degree of

representativeness. However, this method is expensive, time consuming and relatively complicated since it requires a large sample size and the units selected are usually widely scattered. Non-probability sampling makes no claim for representativeness, as every unit does not get the chance of being selected. It is the researcher who decides which sample units should be chosen.

10.5.1 Probability Sampling

Probability sampling today remains the primary method for selecting large, representative samples for social science and business researches. According to Black and Champion, the probability sampling requires following conditions to be satisfied:

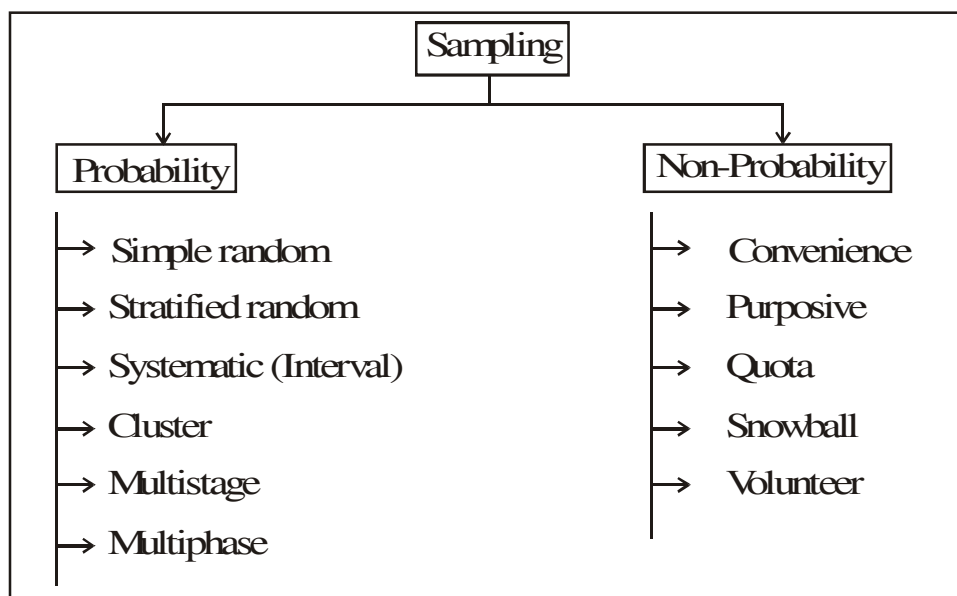
1. Complete list of subjects to be studied is available;
2. Size of the universe must be known;
3. Desired sample size must be specified, and
4. Each element must have an equal chance of being selected. The six forms of probability sampling are:

Simple random, Stratified random, Systematic (or interval), Cluster, Multi- Stage and Multi-phase.

10.5.2 Non-Probability Sampling:

In many research situations, particularly those where there is no list of persons to be studied (e.g., wife battering, widows, Maruti car owners, consumers of a particular type of detergent powder, alcoholics, students and teachers who cut classes frequently, migrant workers and so on), probability sampling is difficult and inappropriate to use. In such researches, non-probability sampling is the most appropriate one.

Non-probability sampling procedures do not employ the rules of probability theory, do not claim representativeness and are usually for qualitative exploratory analysis. The five types of non-probability sampling are: convenience, purposive, quota, snowball and volunteer.



10.5.3 Different types of Probability Sampling

I. Simple Random Sampling:

In this sampling, the sample units are selected by means of a number of methods like lottery method, pricking blind fold, Tappet's tables, computer, personal identification (PIN) or by first letter.

(a) Lottery Method:

This method involves three steps. First step is constructing the sampling frame, i.e. a list of the units of the target population, e.g. students' list, the electoral role in alphabetical order and numbered accordingly second step is writing numbers listed in the sampling frame on small pieces of paper and placing these papers in some vessel/jar etc. Third step is mixing all papers well and taking out one piece of paper from the jar. This process is continued until the required number of respondents is reached. For example, 100 houses are to be allotted to applicants out of 2,500 houses constructed. Here 2,500 pieces of papers numbered from 1 to 2,500 are put in a drum and mixed and some eminent person or some child is invited to take out 100 slips from the drum. If the number on the piece of paper is 535, the name on the list that corresponds to that number is identified and recorded. Thus, 100 numbers selected will be allottees of houses.

(b) Tippet's table or random numbers method:

Tippet has prepared a table of random numbers (of one to five digits each). These numbers are available in various forms, sizes and number combinations in the appendix of the texts on statistics. To understand this let's take an example— Two hundred teachers employed by seven English medium pre-

primary schools in the city apply for attending a two-day seminar. The sponsors, however, only had money to pay for 30 participants. The seminar director, therefore, assigned each applicant a number from 001 to 200, using a table of random number that he found in a statistics text-book. He selected 30 names by moving down columns of 3-digit random numbers and taking the first 30 numbers within the range of 001 to 200. The director decided that this method was easier than picking up number from the urn.

The advantages of Simple Random Sampling are:-

1. All elements have equal chance of being included.
2. It is the simplest of all sampling methods and easiest to conduct.
3. This method can be used in conjunction with other methods in probability sampling.
4. Researcher does not need to know the composition of the population before hand i.e. he requires minimum knowledge of population in advance.
5. Degree of sampling error is low.
6. Most statistical text books have easy to use tables for drawing a random sample.

The disadvantages of Simple Random Sampling are:-

1. It does not make use of knowledge of population which researcher may have.
2. It produces greater errors in the results than do other sampling methods.
3. It cannot be used if the researcher wants to break respondents into sub-groups or strata for comparison purpose.

(b) Stratified Random Sampling:

This is the form of sampling in which the population is divided into a number of strata or sub-groups and a sample is drawn from each stratum. These sub-samples make up the final sample of the study. It is defined as “the method involving dividing the population in homogeneous strata and then selecting simple random samples from each of the stratum.” The division of the population into homogeneous strata is based on one or more criteria, e.g. sex, age, class, educational level, residential background, family type, religion, occupation and so on. Stratification does not involve ranking.

There are two types of stratified sampling (i) proportionate and (ii) disproportionate. The former is one in which the sample unit is proportionate to the size of the sampling unit, while the latter is one in which the sample unit is not related to the units of the target population. Here is an example: Suppose population of 1,000 persons is stratified in five groups on the basis of religion and each group consists of the following number of persons: Hindu– 500, Jain-200, Sikh–150, Muslim– 100 and others–50.

Proportionate sample would be :

5	-4	-3	2	-	1	
↓	↓	↓	↓		↓	
1	2	3	4		10	= 20

Disproportionate sample would be :

5	-	4	-	3	-	2	-	1	
↓		↓		↓		↓		↓	
4		4		4		4		4	= 20

As a general rule, it is wise to use proportionate stratified sample.

As a general rule, it is wise to use proportionate stratified sample. The advantages of stratified random sampling are :

- Sample chosen can represent various groups and patterns of characteristics in the desired proportions.
- It can be used for comparing sub-categories.
- It can be more precise than simple random sampling. The disadvantages of stratified random sampling are:-
- It requires more efforts than simple random sampling.
- It needs a larger sample size than simple random sample to produce statistically meaningful results because each stratum must have at least 20 persons to make statistical comparisons meaningful.

(c) **Systematic (or Interval) Sampling:**

This sampling is obtaining a collection of elements by drawing every n^{th} person from a pre-determined list of persons. In simple words, it is randomly selecting the first respondent and then every n^{th} person after that, 'n' is a number termed as sampling interval.

When the sampling fraction method is employed, samples one drawn from a sampling frame on the basis of the sampling fraction that is equal to $\frac{N}{n}$, where N is the member of units in the target population and 'n' the number of units of the sample.

Systematic sampling differs from simple random sampling in that in the latter, the selections are independent of each other; in the former the selection of sample units is dependent on the selection of a previous one. The advantages of systematic sampling are: -

- It is easy and simple to use:
- It is rapid method and eliminates several steps otherwise taken in probability sampling, and
- Mistakes in drawing elements are relatively unimportant. The disadvantages of this sampling are:-
- It ignores all persons between two n^{th} number with the result that the possibility of over representation and under representation of several groups is greater.
- Since each element has no chance of being selected, it is not probability random sampling as has been pointed out by Black and Champion.

(b) Cluster Sampling:

This sampling implies dividing population into clusters and drawing random sample either from all clusters or selected clusters. This method is used when

- (a) cluster criteria are significant for the study, and
- (b) economic considerations are significant. Initial clusters are called primary sampling units; clusters within the primary clusters are called secondary sampling units; and clusters within the secondary clusters are called multi-stage clusters. When clusters are geographic units, it is called area sampling. For example, dividing one city into various wards, each ward into areas, each area into each neighbourhoods and each neighbourhood into lanes.

We can take an example of a hospital. The issue is to ascertain the problems faced by doctors, patients and visitors in different units and to introduce some reformative programmes. Administratively, it will not be viable to call all doctors from all units nor a large number of patients admitted in different units like cardiology, neurology, orthopedic and so on. Treating each unit as a cluster, randomly selected doctors and patients—say two doctors and three patients or about 50 people all together—from all units may be invited for discussions.

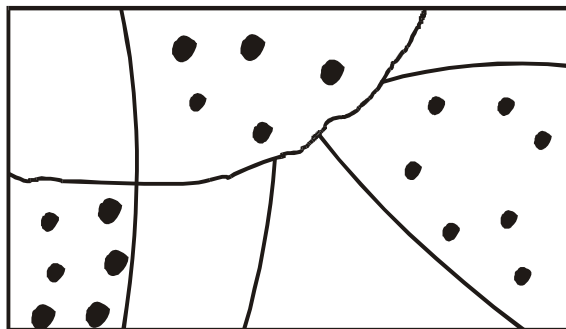
The advantages of cluster sampling are: -

- it is much easier to apply this sample when large populations are studied or when large geographical area is studied.
- Cost in this method is much less than in other methods of sampling.

- Respondents can be readily substituted for other respondents.
- Flexibility is possible.
- Characteristics of clusters can be estimated.
- It is administratively simple since no identification of individuals is necessary, and
- It can be used when it is inconvenient or unethical to randomly select individuals.

Thus, disadvantages of this sampling are: -

- Each cluster is not of equal size in selection of one district from one state, or one village from one block. The district or the village can be small, intermediate or large sized.
- Sampling error is greater
- Same individual can belong to two clusters and studied twice.
- It lacks representation; and
- There could be homogeneity in one cluster but heterogeneity in other.



(e) Multi-Stage Sampling: –

In this method, sampling is selected in various stages but only the last sample of subjects is studied. For example, for studying the Panchayat system in villages, India is divided into zones (four zones), one state is selected from each zone, one district is selected from each state, one block is selected from each district and three villages are selected from each block. This will help in comparing the functioning of Panchayats in different parts of India. Sampling in each stage will be random but it can also be deliberate or purposive. Thus, multi stage sampling according to Ackoff can be combination of (i) simple+simple sampling (ii) simple+systematic (interval) sampling, and (iii) systematic+Systematic sampling.

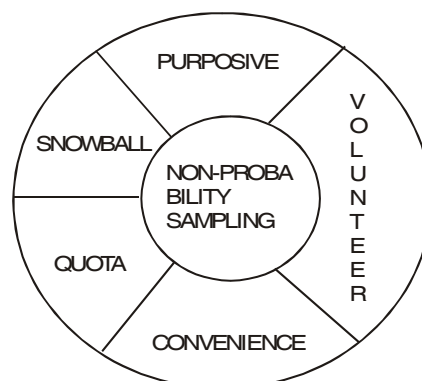
Let us take an example. Suppose bank employees are to be studied in one city for assessing their views on introducing reforms in banks, including use of computers. The names of all managers, accountants and senior clerks in all banks will be typed in the first stage. Suppose these names are typed in 100 pages, each page containing 20 names alphabetically. Out of 2,000 bank personnel, we have to take out a sample of 50 persons. We can do this first by taking out every tenth page (out of 100 pages) i.e., 10 pages, and from each page, we take out every fourth name (i.e., five bank employees from one page). This will be the example of systematic plus systematic sample. The alternative is: take first 10 pages and select any one page at random. In this way, select 10 pages out of 100 pages. From each page select any five names at random. This will be simple plus simple random sampling. The main advantage in this sampling will be that it will be more representative. other advantage is that in all cases, complete listing of population is not necessary. This saves cost.

(f) Multi-Phase Sampling:

The process in this type of sampling is same as in multi-stage sampling i.e., primary selection, secondary selection and so on. However, in a multiphase sampling procedure, each sample is adequately studied before another sample is drawn from it. Consequently, while in multi-stage sampling, only the final sample is studied, in multi-phase sampling, all samples are researched. This offers an advantage over other methods because the information garnered at each phase helps the researcher to choose a more relevant and more representative sample. We can take an example. We are interested in studying MBA students in one city. Suppose there are five institutions imparting MBA education and, in each institution, there are 30 students. Thus, firstly the sampling frame of MBA students in five institutions will be constructed. These respondents will be studied with regard to their academic background, whether they are first or second divisioners. Of these 150 students, 50 will be selected randomly: After selecting these 50 students, 25 girls and 25 boys will be chosen. This sample will be the final sample for the study.

10.5.4 TYPES OF NON-PROBABILITY SAMPLING:

As stated in the earlier lesson, sampling is mainly of two types— Probability and Non-Probability Sampling. We have already understood probability sampling and its types in detail. In this chapter, we will make an attempt to understand non-probability sampling and its various forms. The five types of non-probability sampling include: convenience, purposive, quota, snowball and volunteer.



(a) Convenience Sampling:

This is also known as ‘accidental’ or ‘haphazard’ sampling. In this sampling, the researcher studies all those persons who are most conveniently available or who accidentally come in his contact during a certain period of time in the research. For example, the researcher engaged in the study of university student might visit the university canteen, library, some departments, playgrounds, verandahs and interview certain number of students. Another example is of election study. During election times, media personnel often present man-on-the street interviews that are presumed to reflect public opinion. In such sampling, representativeness is not significant.

The most obvious advantage of convenience sample is that it is quick and economical. But it may be a very biased sample. The possible source of bias could be

1. The respondents may have a vested interest to serve in cooperating with the interviews, and
2. The respondents may be those who are vocal and want to brag

Convenience samples are best utilized for exploratory research when additional research will subsequently be conducted with a probability sample.

(b) Purposive Sampling

In this sampling, which is also known as judgmental sampling, the researcher purposely chooses persons who, in his judgement about some appropriate characteristic required of the sample members, are thought to be relevant to the research topic and are easily available to him. For example, the researcher wants to study beggars. He knows the three areas in the city where the beggars are found in abundance. He will visit only these three areas and interview beggars of his choice and convenience. The manufacturers (of cosmetics, oils garments, etc.) select test market cities because they are viewed as typical cities with demographic profiles closely matching the national profile. Popular journals conduct surveys in selected impetration cities to assess the popularity of politicians and political parties or to for cost election results. Thus, in this technique, some variables are given importance and it represents the universe but the selection of units is deliberate and based on prior judgement.

(c) Quota Sampling

This is a version stratified sampling with the difference that instead of dividing the population into strata and randomly choosing the respondents, it works on quotas fixed by the researcher. In the example of studying 50 MBA students from 150 students in five institutions, the researcher fixes the quota of 10 students from each institution, out of which five will be boys and five girls. The choice of the respondents is left to the interviewer. Determining quotas depends on a number of factors related to the nature and type of research. For instance, the researcher might decide to interview three boys out of five boys from final year and two from previous year, or two studying the morning course and three studying the evening course.

Quota can also be fixed according to their proportion in the entire population. For instance, for studying the attitudes of persons towards use of loudspeakers in religious places in one educational institution with 100 males and 50 females belonging to different religions, quota can be fixed in the ratio of one female for every two males.

Further quota may be fixed on the basis of number of persons in each of the three religious groups.

Males			Females		
Hindu	Muslim	Others	Hindu	Muslim	Others
80	10	10	35	10	5
16	2	2	7	2	1
↓			↓		
20			10		

The advantages of quota sampling are :

1. It is less costly than other techniques.
2. It does not require sampling frames.
3. It is relatively effective.
4. It can be completed in a very short period of time. Its disadvantages are:

1. It is not representative.
2. It has interviewer's bias in the selection.
3. Estimating sampling error is not possible.
4. Strict control of fieldwork is difficult (instead of 25 only 20 respondents may be available.)

(d) **Snowball Sampling:**

In this technique, the researcher begins the research with the few respondents who are known and are available to him. Subsequently, these respondents give other names who meet the criteria of research, who in turn give more new names. This process is continued until, adequate, number of persons and interviewed a until no more respondents are discovered. For instance, in studying wife battering, the researcher may first interview those cases when he knows, who may later on give additional names, and who in turn may give still more names. This method is employed when the target population is unknown or when it is difficult to approach the respondents in any other way. Reduced sample sizes and costs are a clear advantage of snowball sampling. Bias enters because a person known to someone has a higher probability of being similar to first person. If there are major differences between those who are widely known by others and those who are not, there may be serious problems with snowball sampling.

(e) Volunteer Sampling:

This is the technique in which the respondent himself volunteers to give information he holds.

10.6 BIAS IN SELECTING INFORMANTS IN NON-PROBABILITY SAMPLING

The success of the research is dependent on the ‘rich’ information given by the respondents. Many a time, the leading informants selected by the researchers are those who do not have much and appropriate information on the topic under study and who are unwilling to cooperate and respond. The researcher’s bias in selecting the leading subjects is evident in the following cases:

1. The researcher has no knowledge or little knowledge of the social setting of the research. For example, the researcher who wants to study informal social networks in a village or a factory or a university etc., has to locate individuals who could understand what he was looking for and help him in finding it. With little or no knowledge of the situation/location of research, the researcher cannot find such potential informants who have a wider range of interactions.
2. The informants do not represent the population i.e., they do not have the aggregate characteristics in the population.
3. They are not ‘typical’ in the sense that their observations and operations may be misleading. The atypical or marginal informants within their group will not provide adequate information.
4. They are unwilling to be helpful and cooperative.
5. They are activists in a ‘particular’ group because of which they do not present the viewpoints of ‘other groups’.
6. They belong to the community under investigation only marginally and this marginality is bound to bias their views.
7. Selecting informants who are convenient for study.
8. Personal leanings of the researchers of being prejudiced against certain types of persons, say, untouchables, men—Hindus, shabbily dressed persons, too fashionable women, and so forth.

10.7 SAMPLE SIZE

Considerations in Sample Size

A question is often asked: how many persons should be included in the sample, i.e., how large or small must the sample be to be representative? Some people say, the most common size is one tenth of the

total population. Some other say that a minimum of 100 subjects is required to allow statistical inferences. However, these estimates are not always correct. The sample size has to be based on the following considerations: -

1. **The size of the population**, i.e., whether the total population to be studied is very large, large or small.
2. **Nature of population**. i.e., whether the population is homogenous. In the former, a small sample may suffice but, in the latter, a larger sample is required.
3. **Purpose of study** i.e., whether the study is descriptive, exploratory or explanatory.
4. **Whether the study is qualitative or quantitative**. In qualitative studies sampling does not resort to numerical boundaries to determine the size of sample. Similarly, when purposive or accidental sampling are employed; the researcher himself, can decide the 'sufficient' number of respondents. In such cases; generalisations are concerned with quality rather than with quantity.
5. **Accessibility of the elements**: Many a time it is difficult to contact respondents at time and place convenient to the researchers.
6. **Cost of obtaining elements**: With more resources, an adequate number of investigators can be appointed and a large sample may be considered.
7. **Variability required**: Sometimes the respondents require to have persons of different groups e.g., of different age, different income, different educational background, different occupations and so on.
8. **Desired accuracy or confidence level**: For high degree of accuracy, a large sample need to be drawn. One has to think of the level at which one will be confident that his sample is representative.
9. **Sampling error or desired risk level**: The minimum sample error, maximum will be the sample's representatives.
10. **Stratification** i.e., how many times the sample has to be divided during the data analysis. This is to ensure an adequate size for each sub-division.

10.8 LET US SUM UP

After completely studying this lesson, it is clear that sampling is the important part of any research work as it allows us to choose representative items from population to study its characteristics and then generalize it.

Now it is also clear that there are two methods of sampling i.e. probability sampling and non-probability sampling. Further the probability sampling is regarded as best

sampling the data is baseless and every component of population has equal chance to be selected in study sample, Further, a proper care must be taken while determining a sample size as it must be adequate

10.9 GLOSSARY

- **Sampling** is an economical way of collecting data where time, effort and money are the constraints.
- **Probability sampling** is preferable sampling technique as compared to non-probability sampling
- Sometimes **mixed Sampling** be also used which means using multiple methods of probability or non-probability sampling together
- An adequate **sample size** must be set when Sampling technique to be applied taking into consideration various condition.

10.10 SELFASSESSMENT QUESTIONS

1. What is Sampling and why it is important.

2. What are various principles of Sampling.

3. Explain various method of Sampling? Differentiate between probability and non-probability Sampling.

4. What are various pre-requisites of determining a good sample size.

10.11 LESSON END EXERCISE

1. Census is costly and time-consuming method of data collection (True / Falsa)
2. Snowball Sampling technique i.e. _____ technique
3. Multi-phase Sampling include _____ sampling techniques

10.12 SUGGESTED READINGS

1. Argyrous, George. 1997. *Statistics for Social Research*. New York: Mc Millan Press Ltd.
2. Goods, W.J. & Hatt, P.K. 1981. *Methods in Social Research*. New York: Mc Graw Hill.
3. Gupta, S.C. 1981. *Fundamentals of Statistics*. Bombay: Himalayan Publishing House.
4. Gupta, S.P. 2004. *Statistical Methods*. New Delhi: Sultan Chand and Sons.

10.13 ANSWERS TO CHECK YOUR PROGRESS

Ans1: Sample Method:

Sampling method is a technique of collecting data from a subset (sample) of the entire population instead of the whole population.

Uses:

To save time and cost in data collection.

When it is impossible or impractical to study the whole population.

To make quick decisions based on the sample data.

Ans2: Census Method:

It is a method of data collection where information is collected from every member of the population.

Sampling Method:

It is a method where data is collected from a selected subset (sample) of the population.

Ans3: 1. Simple Random Sampling

2. Systematic Sampling

3. Stratified Sampling

4. Cluster Sampling

5. Convenience Sampling

6. Quota Sampling

Ans4: 1. Less time-consuming.

2. Less expensive.

3. Easier to manage and conduct.

Ans5: Lottery Method Procedure:

1. Write the names or numbers of all 10 students on identical slips of paper.

2. Fold the slips and put them in a box or a bag.

3. Mix the slips thoroughly.

4. Draw three slips one by one without looking.

5. The students whose names/numbers are drawn from the sample.

PROCEDURE OF TESTING A HYPOTHESIS

STRUCTURE

- 11.0 Objectives
- 11.1 Introduction
- 11.2 Testing of Hypothesis
- 11.3 Procedure of Testing of Hypothesis
- 11.4 Let us sum up
- 11.5 Glossary
- 11.6 Self-Assessment Questions
- 11.7 Lesson End Exercise
- 11.8 Suggested Readings
- 11.9 Answers to check your progress

11.0 OBJECTIVES

After successful completion of this lesson, you will be able to

- understand the concept of testing of hypothesis.
- learn about null hypothesis and alternative hypothesis,
- formulate the hypothesis,
- know about the main steps in testing of hypothesis.
- know about the concept of confidence level.

11.1 INTRODUCTION

In any research Study after reviewing the existing literature, the next step is to frame the research questions from available literature. These questions need to be answered through a successful research study, For Supporting the Study, hypothesis is framed which are actually suppositions need to be proved.

11.2 TESTING OF HYPOTHESIS

Suppose we wish to infer about population mean (μ) on the basis of sample mean. Then there will always exist a difference between the population mean (μ) and sample mean ($\bar{x}$). Now the question arises whether the difference between two means is significant or insignificant. In case, the difference is significant, we may conclude that the sample is not a true representative of the population and on the contrary, insignificant difference speaks about the representativeness of the sample. Investigations of such problems come within the scope of testing of hypothesis. The tests used to ascertain whether the differences are significant or non-significant are called tests of significance or testing of hypothesis. Before discussing the certain tests of significance let us explain the following terms which are generally use in testing of hypothesis:

- (i) **Statistical Hypothesis:** It may be defined as a statement about one or more populations, i.e., statistical is a statement about one or more parameters.
- (ii) **Null Hypothesis:** Null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true. Since this hypothesis states that there is no significance difference between two values, we call it null hypothesis. It is denoted by H_0 (read as H not).
- (iii) **Alternative Hypothesis:** Any hypothesis which complementary to the null hypothesis is called an alternative hypothesis. It is denoted by H_1 . (read as H one).
- (iv) **Errors in Hypothesis Testing:** Two type of errors may be committed in accepting or rejecting Hypothesis H_0 ; namely type I error and type II error

Type I error = Reject H_0 when it is true

Type II error = Accept H_0 when it is false.
- (v) **Level of Significance :** Probability of rejecting H_0 when it is true is called the level of significance or confidence level. It is denoted by α (read as alfa).

It is also known as the size of rejection region or critical level.

11.3 PROCEDURE OF TESTING OF HYPOTHESIS

The procedure of testing of hypothesis involves the following steps :

Step-I : Formulation of H_0 and H_1 : The very first step in hypothesis testing is to formulate the null and alternative hypothesis i.e., H_0 and H_1 .

Step-II : Selection of Test Statistic: The next step is to choose an appropriate test criterion and its sampling distribution. Which is to be used.

Step-III : Computation of Test Statistic: In third step, we compute the value of test statistic (as selected in Step-II) from the sample observations i.e. from the given data under the assumption of null hypothesis. This value is called calculated or observed value of the test statistic.

Step-IV : Selection of Suitable level of Significance: After setting H_0 and H_1 and computing value of test statistic from the given sample values, our next step is test the validity of H_0 and H_1 at ascertain level of significance. The confidence with which we reject or accept a hypothesis depends upon the significance level adopted. The significance level is expressed as percentage (such as 1 percent 5 percent, 10 percent) . It is generally denoted by α .

Step V : Choose the Critical Region: Find the rejection region at the α percent level of significance (it is also called tabulated value of test statistic).

Step VI : Decision Rule : Finally, we may draw conclusions and take decisions. A statistical decision is a decision either to reject H_0 or to accept H_0 . The decision will depend on whether the calculated value of the test statistic, as calculated in Step-III, falls in the critical region i.e. in the rejection region. We accept our H_0 if the value of computed test statistic is less than or equal to tabulated value of test statistic at a α percent level of significance. Otherwise reject H_0 .

CHECK YOUR PROGRESS-I

Select the appropriate answer:

1. A statement made about a population for testing purpose is called?

- a) Statistic
- b) Hypothesis
- c) Level of Significance
- d) Test-Statistic

2. If the assumed hypothesis is tested for rejection considering it to be true is called?

- a) Null Hypothesis
- b) Statistical Hypothesis
- c) Simple Hypothesis
- d) Composite Hypothesis

3. A statement whose validity is tested on the basis of a sample is called?

- a) Null Hypothesis
- b) Statistical Hypothesis
- c) Simple Hypothesis
- d) Composite Hypothesis

4. A hypothesis which defines the population distribution is called?

- a) NullHypothesis
- b) StatisticalHypothesis
- c) SimpleHypothesis
- d) Composite Hypothesis

5. If the null hypothesis is false then which of the following is accepted?

- a) Null Hypothesis
- b) Positive Hypothesis
- c) Negative Hypothesis
- d) Alternative Hypothesis.

7. The point where the Null Hypothesis gets rejected is called as?

- a) Significant Value
- b) Rejection Value
- c) Acceptance Value
- d) Critical Value

11.4 LET US SUM UP

After completing the lesson successfully, it is understood that hypothesis is a supposition needed to be proved. Statistical hypothesis is of two types namely null hypothesis and alternative hypothesis. These are various steps involved after choosing one of the types of hypotheses to come at decision rule i.e. draw some conclusion.

11.5 GLOSSARY

- **Hypotheses** is a statistical supposition need to be proved right or wrong.
- **Null hypotheses** assume no difference in actual value and assumed value or whatever the subject may be.
- **Alternative hypothesis** assumes difference in two values.

11.6 SELFASSESSMENT QUESTIONS

1. What do you mean by hypothesis? What are its various types.

2. What are the steps for testing hypothesis.

11.7 LESSON END EXERCISE

1. Null hypothesis is denoted by _____
2. Alternative hypothesis is denoted by _____
3. Conclusions are drawn only when hypothesis proved right or wrong (True/ False).

11.8 SUGGESTED READINGS

1. Argyrous, George. 1997. *Statistics for Social Research*. New York: Mc Millan Press Ltd.

2. Goods, W.J. & Hatt, P.K. 1981. *Methods in Social Research*. New York: Mc Graw Hill.
3. Gupta, S.C. 1981. *Fundamentals of Statistics*. Bombay: Himalayan Publishing House.
4. Gupta, S.P. 2004. *Statistical Methods*. New Delhi: Sultan Chand and Sons.

11.9 ANSWERS TO CHECK YOUR PROGRESS

1. **b) Hypothesis**
2. **a) Null Hypothesis**
3. **b) Statistical Hypothesis**
4. **c) Simple Hypothesis**
5. **.d) Alternative Hypothesis.**
6. **d) Critical Value**

TESTS OF SIGNIFICANCE-STUDENT'S T-TEST**STRUCTURE**

- 12.0 Objectives
- 12.1 Introduction
- 12.2 Test of significance
- 12.3 Students T-Test
- 12.4 Let us sum up
- 12.5 Glossary
- 12.6 Self-Assessment Questions
- 12.7 Lesson End Exercise
- 12.8 Suggested Readings
- 12.9 Answers to Check Your Progress

12.0 OBJECTIVES

After successful completion of this lesson, you shall be able to know

- Tests of Significance
- Students T- test.
- Application of T- test.

12.1 INTRODUCTION

In previous lesson we discussed the concept of hypothesis. We also studied the procedure of testing hypothesis. Now in this lesson we will discuss about students t- test which is one of the methods of testing hypothesis.

12.2 TESTS OF SIGNIFICANCE

Various tests of significance can broadly be classified into following three heads:

- (a) Test of significance for attributes.
- (b) Test of significance for large samples, and
- (c) Test of significance for small samples or exact tests.

12.3 STUDENT'S T-TEST

Some of the testing procedures can be used only when the samples are large. A sample of size 30 or more is generally considered to be a large sample. When samples are small, i.e., less than 30, the large sample results do not hold good for small samples. That is the assumption of approximate normality of the distribution is not true; in fact, another distribution (exact distribution) of the test statistic is to be used and the result modified accordingly. Generally, the students. t-test, Z-test, χ^2 test and F test are the exact tests or small tests of interest. The present lesson deals with student's t-test. Student's t-test is used to test the significance of mean, significance of difference of two mean and significance of correlation coefficient Z-test is also used to test the significance of correlation coefficient. We will discuss these tests in following sections.

These tests (based on t-test) are performed under the following assumptions/ conditions:

- (i) The observations must be drawn from normal population (s)
- (ii) The observations must be independent.
- (iii) The sample size should be small, usually not more than 30.
- (iv) The parent population (s) must have the same but unknown variance.

The student's t-test has the following applications: Testing the Significance of mean:

Suppose we wish to test the hypothesis regarding the mean of a population on the basis of a random sample of size $n(<30)$ from a population with unknown standard deviation. In this case we set up the null hypothesis against different alternatives as

$H_0 : \mu = \mu_0$ against any one of H_1

$H_1 : \mu \neq \mu_0$ (two sided alternative) or

$H_1 : \mu > \mu_0$ (right-tailed alternative) or

$H_1 : \mu < \mu_0$ (left-tailed alternative)

The test statistic for testing H_0 is

$$t = \frac{\bar{x} - \mu_0}{\text{S.E.}(\bar{x})} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where $\bar{x}$ is the sample mean and s^2 is an unbiased estimator of the unknown population variance σ^2 , which is given by

$$s^2 = \frac{1}{n-1} \sum (x - \bar{x})^2 = \frac{1}{n-1} [\sum x^2 - n\bar{x}^2]$$

The distribution of t under H_0 is t -distribution with $(n-1)$ degree of freedom. After calculating t -statistic, the decision about the acceptance or rejection of H_0 is taken in the following manner :

- (a) In case of testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ (two sided alternative) at α percent level of significance, we accept H_0 if $|t| \leq t_{n-1}(\alpha/2)$, otherwise reject H_0 .
- (b) For testing $H_0 : \mu = \mu_0$ against one sided alternative at α percent level of significance accept H_0 if $|t| \leq t_{n-1}(\alpha)$, otherwise, reject H_0 . Here $t_{n-1}(\alpha)$ is the critical value of t -statistic at α percent level of significance for $(n-1)$ degrees of freedom. These critical values are given in Table-1 of the Appendix given at the end of the study material.

Example 1. A random sample of 10 independent observations from a normal population provided the following results :

165, 160, 161, 170, 172, 160, 165, 175, 164, 168

- (a) Test the hypothesis that the population mean is 170 against the alternative that it is not 170 at 10 percent level of significance.
- (b) Find the 90 percent confidence limits of the population mean.

Solution: We want to test the hypothesis

$H_0 : \mu = 170$ against $H_1 : \mu \neq 170$

To test H_0 , we make the use of t -test

$$t = \frac{\bar{x} - \mu_0}{\text{S.E.}(\bar{x})} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

X	X ²
165	27225
160	25600
161	25921
170	28900
172	29584
160	25600
165	27225
175	30625
164	26896
168	28224
Total 1660	275800

$$\bar{x} = \frac{1660}{10} = 166$$

$$s^2 = \frac{1}{n-1} [\sum x^2 - n\bar{x}^2]$$

$$= \frac{1}{9} [275800 - 10 \times (166)^2]$$

$$= \frac{1}{9} [275800 - 275560]$$

$$= 26.67$$

$$\therefore t = \frac{166 - 170}{\sqrt{26.67/10}} = \frac{-4.0}{1.633} = -2.45$$

$$\therefore |t| = 2.45$$

For n-1=9, the tabulated value of t at 10 percent level of Significance

$t_9 \left(\frac{0.10}{2} \right) = t_9(0.05) = 1.833$. Since the observed value of $|t|$ is greater than the tabulated value of t (1.833), hence we reject H_0 , which means that the mean of population can not be regarded equal to 170.

(b) 95% confidence limits for mean are

$$\bar{x} \pm S.E.(\bar{X}) \cdot t_{n-1} \left(\frac{\alpha}{2} \right) \text{ or } \bar{x} \pm S.E.(\bar{X}) \cdot t_9(0.05)$$

$$\text{or } 166 \pm 1.633 \times 1.833$$

$$\text{or } 166 \pm 2.99$$

$$\text{or } (163.01 \text{ and } 168.99)$$

Example 2. A T.V. manufacturing company is marketing a particular type of brand through a large number of retail shops. Before a heavy advertising campaign, the average sales per shop per month was 120 T.V. After the campaign, a sample of 24 shops was taken and the average sales was found to be 130 with S.D. 12. Can you consider the advertisement campaign effective at 5 percent level of significance?

Solution : Here we are given that $n=24$, $\bar{x}=130$ and $s=12$

We wish to test

$$H_0: \mu=120 \text{ against } H_1: \mu \neq 120$$

The test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{130 - 120}{12/\sqrt{24}} = 4.082$$

For $n=24$, $\alpha=0.05$, the critical value of $t_{n-1} \left(\frac{\alpha}{2} \right)$; i.e., $t_{23}(0.025)$ is 2.069. Since the calculated value of t (4.082) is greater than the tabulated value of t (2.069) at 5 percent level of significance with 23 degrees of freedom, so we reject H_0 . Thus we may conclude that advertisement is effective for increasing the sales of T.V.

b) **Testing the Significance of Difference Between Two means (Independent Samples)**

Let two independent random samples of sizes n_1 and n_2 are drawn from two normal populations with means m_1 and m_2 and unknown but equal variances i.e. $s_1^2 = s_2^2 (=s^2)$ respectively. Here we may be interested in testing the hypothesis that both the samples come from the same normal population, i.e.,

$H_0 : m_1 = m_2$ against any one of the following alternatives

$H_1 : m_1 \neq m_2$ (two sided alternative)

$H_1 : \mu_1 > \mu_2$ (right-sided alternative)

$H_1 : \mu_1 < \mu_2$ (left-sided alternative)

To test H_0 , the test statistic is

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1 + n_2 - 2}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the means of first and second samples and s is given by

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{\sum (X_1 - \bar{X}_1)^2 + \sum (X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}}$$

The decision rule about the acceptance or rejection of H_0 is same as in previous section.

Example 3. Two horses A and B were tested according to the time (in seconds) to run a particular track with the following results :

Horse A : 28 30 32 33 33 29 34

Horse B : 29 30 30 24 27 29

Test whether can you discriminate between the two horses at 5 percent level of significance.

Solution : Let the time (in seconds), X_1 , to run a particular track by horse A is $N(\mu_1, \sigma^2)$ and X_2 that of horse B is $N(\mu_2, \sigma^2)$ where σ^2 is unknown. We wish to test the hypothesis that there is no significance difference between the performance of two horses, i.e.,

$$H_0 : \mu_1 = \mu_2 \text{ against } H_1 : \mu_1 \neq \mu_2$$

The test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

where

$$\bar{x}_1 = \frac{\sum x_1}{n} = \frac{219}{7} = 31.286$$

$$\bar{x}_2 = \frac{\sum x_2}{n} = \frac{169}{6} = 28.167$$

x_1	x_2	x_1^2	x_2^2
28	29	784	841
30	30	900	900
32	30	1024	900
33	24	1089	576
33	27	1089	729
29	29	841	841
34		1156	
Total 219	169	6883	4787

$$(n_1-1)s_1^2 = \sum x_1^2 - n_1 \bar{x}_1^2 = 6883 - 7 \times (31.286)^2$$

$$= 6883 - 6851.697$$

$$= 31.30$$

and

$$\begin{aligned}(n_2-1) s_2^2 &= \sum x_2^2 - n_2 \bar{x}_2^2 = 4787 - 6 \times (28.167)^2 \\ &= 4787 - 4760.279 \\ &= 26.721\end{aligned}$$

$$\begin{aligned}\therefore s &= \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{31 \cdot 30 + 26 \cdot 721}{7 + 6 - 2}} \\ &= \sqrt{5 \cdot 275} = 2.297\end{aligned}$$

Thus

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{31.286 - 28.167}{2.297 \sqrt{\frac{1}{7} + \frac{1}{6}}} = \frac{3.119}{1.278} = 2.441$$

The tabulated value of $t_{n_1+n_2-2} \left(\frac{\alpha}{2} \right)$, i.e., $t(0.025)$ is 2.201. Since the value of $t(2.441)$ is greater than the tabulated value of t at 5 percent level of significance with 11 degrees of freedom, so we reject H_0 . This means we can discriminate between the two horses.

Example 4. The mean life of sample of 10 electric bulbs was found to be 14500 hours with standard deviation 420 hours. A second sample of 15 bulbs chosen from a different batch showed a mean life of 14175 hours with S.D. of 380 hours. Is there a significance difference between the mean life of two batches of bulbs?

Solution : Let us formulate the hypothesis

H_0 : There is no significance between the mean life of two batches of bulbs. That is

$H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$

The test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where (given) $\bar{x}_1=15500$, $\bar{x}_2=14175$, $s_1=420$, $s_2=380$, $n_1=10$ and $n_2=15$

Thus

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{9 \times (420)^2 + 14(380)^2}{10 + 15 - 2}} = \sqrt{156921.739}$$

$$= 396.134$$

Hence

$$t = \frac{14500 - 14175}{396.134 \sqrt{\frac{1}{10} + \frac{1}{15}}} = \frac{325}{161.721} = 2.01$$

Here $n_1+n_2-2=23$ and $\alpha=0.05$ so $\alpha/2=0.025$, thus the tabulated value of $t_{23}(0.025)$ is 2.069.

Since the observed value of t (2.01) is less than $t_{23}(0.025)$ thus we accept H_0 . This means there is no significant difference between the mean life of two batches of bulbs.

(c) Testing the Difference Between Two Means (Dependent Samples or Paired Observations)

In section 19.4, we assumed that the samples have been drawn from two normal populations and they are independent. However, in situations where two samples are not independent, we use paired t -test. Here the observations in the two samples occur in pairs so that the two observations are related in respect of some characteristic. For example, if we wish to test a new diet using on some individuals, then the weight of the individuals recorded before and after completion of test will form two samples in which observations will be paired. Here our hypothesis is that there is no significance difference between the performance of before and after some test. The test statistic is

$$t = \frac{\bar{d}}{s/\sqrt{n}} \sim t_{n-1}$$

where $\bar{d}$ = the mean of difference of observations of two samples

s = standard deviation of the difference of observations.

The decision whether to accept H_0 or reject H_0 remain same as in previous sections.

Example 5. Ten college students were given a test in Business Statistics. They were given a month's coaching and a second test was held at the end of coaching. The marks recorded in both tests are :

Students :	1	2	3	4	5	6	7	8	9	10
Marks in	46	42	63	54	34	46	72	43	69	74
Test-I										
Marks in	55	40	71	59	48	41	75	50	75	76
Test-II										

Do the marks give evidence that the coaching is effective?

Solution :

Students	Marks		$d = X_1 - X_2$	d^2
	I Test	II Test		
	(X_1)	(X_2)		
1	46	55	-9	81
2	42	40	2	4
3	63	71	-8	64
4	54	59	-5	25
5	34	48	-14	196
6	46	41	5	25
7	72	75	-3	9
8	43	50	-7	49
9	69	75	-6	36
10	74	76	-2	4
Total			-47	493

$$\bar{d} = \frac{\sum d}{n} = \frac{-47}{10} = -4.70$$

$$\begin{aligned} s &= \sqrt{\frac{1}{n-1} [\sum d^2 - n\bar{d}^2]} \\ &= \sqrt{\frac{1}{9} [493 - 10 \times (-4.70)^2]} \\ &= \sqrt{30.233} = 5.498 \end{aligned}$$

Now let us formulate the hypothesis

H_0 : There is no significance between the marks of students before and after the coaching.

That is

$$H_0 : \mu_1 = \mu_2 \text{ or } H_0 : \mu_1 - \mu_2 = 0 \text{ or}$$

$$H_0 : \mu_d = 0 \text{ against } H_1 : \mu_1 < \mu_2 \text{ or } H_1 : \mu_d < 0.$$

The test statistic is

$$\begin{aligned} t &= \frac{\bar{d}}{s/\sqrt{n}} = \frac{-4.70}{5.4981/\sqrt{10}} \\ &= \frac{-4.70}{1.739} = -2.703 \end{aligned}$$

$$\therefore |t| = 2.703$$

The tabulated value of $t_{n-1}(\alpha\%)$, i.e. $t_9(0.05)$, (if $\alpha=0.05$) is 1.833. Since the observed value of t (2.703) is greater than the tabulated value of $t_9(0.05)$, thus we reject H_0 . Hence, we may conclude that there is a significance difference in the marks of students. Thus the coaching is effective.

Example 6. A certain drug was given to 12 patients and the increments in their blood pressure were recorded to be

5, 2, 8, -1, 3, 0, -2, 1, 5, 0, 4, and 6

Is it reasonable to believe that the drug has no effect on change of blood pressure?

Solution : Suppose that increase, d , in blood pressure has a normal distribution with mean μ_d and unknown S.D. σ_d . Here

$$H_0 : \mu_d=0 \text{ against } H_1 : \mu_d<0$$

Patient	d	d^2
1	5	25
2	2	4
3	8	64
4	-1	1
5	3	9
6	0	0
7	-2	4
8	1	1
9	5	25
10	0	0
11	4	16
12	6	36
Total	31	185

Here $n=12$,

$$\bar{d} = \frac{\sum d}{n} = \frac{31}{12} = 2.583$$

$$s = \sqrt{\frac{1}{n-1} [\sum d^2 - n\bar{d}^2]}$$

$$= \sqrt{\frac{1}{11} [185 - 12 \times (2.583)^2]}$$

$$= \sqrt{9.54} = 3.089$$

Hence

$$t = \frac{\bar{d}}{s/\sqrt{n}} = \frac{2.583}{3.089/\sqrt{12}} = \frac{2.583}{0.892} = 2.896$$

The tabulated value of t_{11} (0.05), if $\alpha=0.05$, is 1.796. Since the observed value of $|t|$ is greater than the tabulated value of $t_{11}(0.05)$, so we reject H_0 . This means the drug seems to have increase the blood pressure.

(d) Testing the Significance of an Observed Correlation Coefficient

Suppose ρ denote the population correlation coefficient and r denote the observed correlation coefficient from the sample values. Then to test the hypothesis

$H_0 : \rho=0$ against $H_1 : \rho \neq 0$, the test statistic used is

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

The decision rule remain same as in previous sections.

Example 7. A random sample of 11 observations from a bivariate population gave a correlation coefficient 0.239. Could the observed value have arisen from an uncorrelated population?

Solution : Let us formulate the hypothesis

$H_0 : \rho=0$ against $H_1 : \rho \neq 0$

Test statistic is

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

where $r=0.239$, $n=11$, thus

$$t = \frac{0.239 \times \sqrt{11-2}}{\sqrt{1-(0.239)^2}} = \frac{0.717}{0.971} = 0.738$$

If $\alpha=0.05$, then tabulated value of $t_{n-2}\left(\frac{\alpha}{2}\right)$, i.e. $t_9(0.025)$ is 2.262. Since observed value of $t(0.738)$ is less than the tabulated value of t at 5 percent level of significance with 9 degrees of freedom, thus we accept H_0 . This means observed value of r might have obtained from an uncorrelated population.

Example 8. How many pairs of observations must be included in a sample in order that an observed correlation coefficient of value 0.48 shall have a calculated value of t greater than 2.72?

Solution : We have given the value of r , and t and we have to find the value of n . As we know that

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

It is given that

$$t > 2.72$$

$$\text{or } \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} > 2.72$$

$$\text{or } \sqrt{n-2} > \frac{2.72 \times \sqrt{1-r^2}}{r}$$

$$\text{or } n-2 > \frac{(2.72)^2 \times (1-r^2)}{r^2}$$

$$\text{or } n > \frac{(2.72)^2 \times (1-r^2)}{r^2} + 2$$

$$\text{or } = \frac{(2.72)^2 \times (1-0.48^2)}{(0.48^2)} + 2$$

$$= \frac{5 \cdot 6938}{0 \cdot 2304} + 2$$

$$= 24 \cdot 713 + 2 = 26 \cdot 713 \approx 27$$

Hence we should include atleast 27 observations.

CHECK YOUR PROGRESS-I

MULTIPLE CHOICE QUESTIONS

1. The T-test tells you about the significant difference between the two groups.

- a. True
- b. False

2. A t-test is a significance test that assesses

- a. The means of two independent groups
- b. The medians of two dependent groups
- c. The modes of two independent variables
- d. The standard deviation of three independent variables

3. Which of these is a type of T-test?

- a. One sample t-test.
- b. Independent two-sample t-test.
- c. Paired sample t-test.
- d. All of these

4. Which of the following types of T-tests can be used to analyze the difference between the means of two groups?

- a. Only Independent T-test
- b. Only Paired T-test
- c. Only One-sample T-test
- d. All of these

5. The T-test is not a reliable test.

- a. True
- b. False

6. Statistical significance or the probability of finding statistical significance is also known as

- a. Degrees of freedom (df)
- b. P-value
- c. Standard deviation (sd)
- d. A constant source of frustration!

12.4 LET US SUM UP

In the present lesson we have discussed the student t-test. This test has the following applications.

(i) Testing of the mean of a normal distribution with unknown standard deviation

The test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

where s is the sample standard deviation which is given by

$$s = \sqrt{\frac{\sum x^2 - n\bar{x}^2}{n-1}}$$

(II) Testing the equality of means of two normal population

The test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

where

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

s_1^2 and s_2^2 are the sample variances of the first and second sample respectively.

- (iii) **Testing the equality of means of two populations when samples are dependent or paired samples.**

The test statistic is

$$t = \frac{\bar{d}}{s/\sqrt{n}} \sim t_{n-1}$$

where $\bar{d}$ is the mean of difference of observations and

$$s = \sqrt{\frac{\sum(d - \bar{d})^2}{n - 1}}$$

- (iv) **Testing the significance of observed correlation coefficient**

The test statistic is

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

where r is the sample correlation coefficient

- (v) **Z** - test is used to test whether an observed correlation coefficient differs significantly from some hypothetical value or whether two samples values of r differ significantly.

12.5 GLOSSARY

- **Test** of significance are used to test the power of hypothesis.
- **Test** is applicable when the sample size is small

- **A small sample** means the strength of data in sample must be less than or equal to 30 only.
- **T- test** is applicable when comparison is made between two populations not more than two.

12.6 SELFASSESSMENT QUESTIONS

1. Distinguish between large sample and small sample tests.

2. A certain food processing plant has prepared a product with an average water content of 37.5 percent. A sample of 10 units taken after a change in cooking procedure, shows an average water content of 39.2 percent and standard deviation of 3.4 percent. Had the procedure made a significant change in percentage of water contents.

3. Two different models are available for the same machine. The number of units produced per hour of these two models are given below:

Days :	1	2	3	4	5	6	7
Model A:	180	176	184	181	190	137	–
Model B:	195	194	190	192	187	185	187

Will you conclude that model A and model B have the same productivity.

4. A certain drug administered to each of 10 patients resulted in the following additional hours of sleep:

0.7, -1.10, -0.20, 1.20, 0.1, 3.4, 3.7, 0.8, 1.8, 2.0

Test whether these data justify the claim that drug does not produce additional sleep.

5. A random sample of 16 observations from a bivariate population gave a correlation coefficient 0.42. Could the observed value have arisen from an uncorrelated population.

6. From a sample of 20 pairs of observations the correlation is 0.56 and the corresponding population correlation is 0.42. Is the difference significant.

12.7 LESSON END EXERCISE

1. X is used for _____ mean whereas u is used for denoting .
2. T-test work on only _____ samples.
3. T-test can also be used to compare assumed results and actual results (True/ False)

12.8 SUGGESTED READINGS

1. Argyrous, George. 1997. *Statistics for Social Research*. New York: Mc Millan Press Ltd.
2. Goods, W.J. & Hatt, P.K. 1981. *Methods in Social Research*. New York: Mc Graw Hill.
3. Gupta, S.C. 1981. *Fundamentals of Statistics*. Bombay: Himalayan Publishing House.
4. Gupta, S.P. 2004. *Statistical Methods*. New Delhi: Sultan Chand and Sons.

12.9 ANSWERS TO CHECK YOUR PROGRESS

- | | |
|-----------------|--|
| 1. True | 2. The means of two independent groups |
| 3. All of these | 4. All of these |
| 5. False | 6. P-value |

CHI SQUARE TEST

STRUCTURE

- 13.0 Objectives
- 13.1 Introduction
- 13.2 Chi-square Distribution and Statistic.
- 13.3 Application of Chi-Square Test.
- 13.4 Let us sum up
- 13.5 Glossary
- 13.6 Self-Assessment Questions
- 13.7 Lesson End Exercise
- 13.8 Suggested Readings
- 13.9 Answers to Check Your Progress

13.0 OBJECTIVES

After successful completion of this lesson, you will be able to:

- o know about chi-square distribution and chi-square statistic,
- o understand the role of chi-square distribution in testing of hypothesis,
- o perform test regarding the variance of a normal distribution, and
- o understand and conduct tests of goodness of fit and testing the independence of categorized data.

13.1 INTRODUCTION

In the previous lessons we have discussed the meaning of testing of hypothesis, various tests of hypothesis and also how some of these tests concerning the means, proportions, correlation coefficient of one or two populations could be designed and conducted. But in real life, one is not always concerned with the mean and the proportions alone-nor is one always interested in only one or two populations. A business manager may want to test if there is any significant difference in the proportion of high-income households where his products is preferred in different regions. In such situations the business manager is interested in testing the quality of proportion among different regions (populations). In many of our earlier tests, we had assumed that the population distribution was normal. It should be possible for us to test if the population distribution is really normal, based on the evidence provided by a sample. Similarly, in other situations it should be possible for us to test whether the population distribution is Poisson, Binomial or any other known distribution. The methods, (based on Chi-square) that we are going to discuss in the subsequent sections of this lesson will help us in the kind of situations mentioned above as well as in many others types of situations. Before discussing these testing problems in detail first we know about chi-square distribution and its assumptions.

13.2 CHI-SQUARE DISTRIBUTION AND STATISTIC

If X is a random variable having a standard normal distribution, then X^2 will have a Chi-square distribution with one degree of freedom. Further if $X_1, X_2, \dots, X_n$ be a random sample of size n from $N(\mu, \sigma^2)$ then, $\sum \left(\frac{X - \bar{X}}{\sigma} \right)^2$ will have a Chi-square (in notation χ^2) distribution with $n-1$ degree of freedom. That is

$$\sum \left(\frac{X - \bar{X}}{\sigma} \right)^2 \sim \chi_{n-1}^2$$

$$\text{or } \frac{(n-1) s^2}{\sigma^2} \sim \chi_{n-1}^2$$

In fitting of distributions, under the assumption that two sets of frequencies are not significantly different, the Chi-square statistic is used which is defined as

$$\sum (O - E)^2/E \sim \chi_n^2$$

where O refers to the observed frequencies and E refers to the expected frequencies. However, there are certain conditions for the validity of this χ^2 -test. These are :

- (i) The sample observations should be independent and normally distributed.

- (ii) No estimated or theoretical cell frequency should be less than 5.
- (iii) The total frequency should be reasonably large, say, more than 50
- (iv) Constraints imposed upon the observations, if any, should be linear.

CHECK YOUR PROGRESS-I

MULTIPLE CHOICE QUESTIONS

- 1. Which of the following is a key characteristic of the Chi-square test?**
 - A. It is a parametric test.
 - B. It requires interval data.
 - C. It is a non-parametric test.
 - D. It assumes a normal distribution of the data.
- 2. The Chi-square test of independence is used to determine:**
 - A. If the mean of two groups is different.
 - B. If there is an association between two categorical variables.
 - C. If the variance of a population is known.
 - D. If the sample comes from a normal distribution.
- 3. The Chi-square distribution is generally:**
 - A. Left-skewed
 - B. Right-skewed
 - C. Symmetric
 - D. Bell-shaped
- 4. Which of the following is an assumption for the Chi-square test?**
 - A. The sample size should be very small.
 - B. Expected frequencies should generally be
 - C. Data must be continuous.
 - D. The population must be normally distributed.

5. For a 2x2 contingency table (e.g., gender vs. preference), what are the degrees of freedom?
- A. 4
 - B. 2
 - C. 1
 - D. 3
6. A Pearson's Chi-square test is considered a:
- A. Lower-tail test
 - B. Upper-tail test
 - C. Two-tailed test
 - D. Paired test

13.3 APPLICATION OF CHI-SQUARE TEST

As we have discussed the applications of chi-square test in section 15.1, now we will explain their procedure in the following sections:

Testing of Population Variance

Many times, we are interested in knowing if the variance of a population is different from a known value. That is, we want to test the hypothesis $H_0: \sigma^2 = \sigma_0^2$ against a suitable alternative. The test statistic is

$$\frac{\sum (X - \bar{X})^2}{\sigma_0^2} = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi_n^2 - 1$$

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

By comparing the observed value of χ^2 with tabulated value of χ^2 with (n-1) degree of freedom at appropriate level of significance we may accept or reject H_0 . The tabulated values of χ^2 are given in Table 2 of Appendix.

Example 1. A random sample of size 24 from a normal population with standard deviation 12 gave the standard deviation 8.5. Is there any significance difference between population and sample standard deviation?

Solution : Here $n = 24$, $s = 8.5$ and we want to test the hypothesis : $H_0 : \sigma^2 = 12 \cdot 0$

The test statistic is

$$\chi^2 = \frac{(n-1) s^2}{\sigma_0^2} \sim \chi_{n-1}^2$$

$$\therefore \chi^L = \frac{(24-1) \times (8.5)^2}{12.0}$$

$$= \frac{23 \times 72.25}{144}$$

$$= 11.54$$

If , $\alpha = 0.05$, then tabulated value of χ^2 with $n-1 = 24$ d.f. at 5 percent level of significance is 36.415. Since the calculated value of χ^2 is less than that of tabulated value, so we accept H_0 . Hence, we may say that population standard deviation is 12.0.

Example 2. Packages made by a standard method for a long time indicate that $\sigma_0 = 11.75$ for the variation of weight. A less expensive and less time consuming new method is tried and it shows standard deviation 12.261 on the basis of a sample of size 20. Test whether the new method results in an increase of the variability of the quality of packages in terms of their weights.

Solution : Here we wish to test the hypothesis

$$H_0 : \sigma_1 = \sigma_0 = 11.75 \quad \text{against}$$

$$H_1 : \sigma > \sigma_0 = 11.75$$

We have $n = 20$, $s = 12.261$ and $n-1 = 19$

The test statistic is

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{19 \times (12.261)^2}{(11.75)^2} = 20.69$$

The tabulated value of χ^2 for 19 d.f. at 5% level of significance is 30.14. Since the calculated χ^2 is less than tabulated χ^2 , so we accept H_0 , that the variability in terms of weight of the packages has not increased.

(II) Chi-square Test for Goodness of Fit :

Some time we are interested in knowing if it is reasonable to assume that the population is normal, Poisson, Binomial or any other known distribution. In this section we describe the procedure to test how close is the fit between observed data and distribution assumed. These tests are based on Chi-square statistic.

For explaining the goodness of fit test, let us consider a population which may be partitioned into k -classes, and let p_i be the probability that an observation belongs to the i th class ($i = 1, 2, \dots, k$) with $\sum_{i=1}^k p_i = 1$. Further let a random sample of size n be drawn from the population. Suppose O_i is the number of sample observations belonging to i th class such that $\sum_{i=1}^k O_i = n$. Further let E_i be the expected frequency of the i th class. Thus, under H_0 , expected frequency $E_i = np_i$ ($i = 1, 2, \dots, k$) and obviously $\sum_{i=1}^k E_i = n$

Now, a goodness of fit test between observed and expected frequencies is used based on the statistic

$$\begin{aligned}\chi^2 &= \sum_{i=1}^k (O_i - E_i)^2 / E_i \text{ or simply} \\ &= \sum (O - E)^2 / E\end{aligned}$$

If the total frequency n is sufficiently large and the expected frequency in none of the class is too small, χ^2 - statistic follows χ^2 distribution with ν , degrees of freedom where $\nu = k - r - 1$ and r is the number of independent parameters estimated from the sample observations. The calculated value of χ^2 is compared with the table value of χ^2 for given α . As χ^2 measures the discrepancy between the observed data and fitted distribution, the large value of χ^2 would lead to rejection of H_0 . Thus, for goodness of test, the rejection region is always the right tail.

Before calculating the value of χ^2 statistic if we observe that expected frequencies in any class is less than 5, then such frequencies are “pooled or combined with adjacent classes. Consequently, the corresponding observed frequencies too are pooled or combined with adjacent classes and in the process, the number of classes is also reduced.

Example 3. A die is thrown 90 times and the number of faces shown are as indicated below :

Face (i)	:	1	2	3	4	5	6
----------	---	---	---	---	---	---	---

Frequency (ni) : 18 14 13 15 14 16

Test whether the die is fair.

Solution : Let p_i be the probability of showing the face i , $i = 1, 2, \dots, 6$.

Here we want to test the hypothesis

$$H_0 : p_1 = p_2 = \dots = p_6 = 1/6$$

We have $E_i = n p_i$

$$= 90 \times \frac{1}{6} \text{ (for all classes)}$$

$$= 15$$

Hence

$$\chi^2 = \sum (O-E)^2/E = \frac{(18-15)^2}{15} + \frac{(14-15)^2}{15} + \frac{(13-15)^2}{15} +$$

$$\frac{(15-15)^2}{15} + \frac{(14-15)^2}{15} + \frac{(16-15)^2}{15}$$

$$= \frac{9}{15} + \frac{1}{15} + \frac{4}{15} + 0 + \frac{1}{15} + \frac{1}{15}$$

$$= \frac{16}{15} = 1.07$$

If $\alpha = 5\%$, then the tabulate value of χ^2 with $k-1$, i.e., 5 d.f. at 5 percent level is 11.08 which is greater than calculated value of χ^2 . Hence there is no reason to reject the H_0 or that the die is fair.

Example 4. Following is the frequency distribution of the number of arrivals per unit of time of patients at the OPD of a hospital. Verify whether the arrivals follow a Poisson distribution.

Number of arrivals (X_i) : 0 1 2 3 4 5 6 7

(per ten minutes)

Observed frequency (O_i) : 10 30 40 50 35 20 10 5

Solution : We want to test the hypothesis

H_0 : The number of arrivals per unit of time has a Poisson distribution.

To test this H_0 , we use χ^2 test of goodness of fit.

To compute the expected frequencies (E_i), we first estimate the parameter (λ) of the Poisson distribution from the given data as

$$\lambda = \bar{x} = \frac{\sum fx}{\sum f} = \frac{595}{200} = 2.975$$

The Poisson probabilities $p(x_i)$ can be calculated by the following relation

$$p(x+1) = \frac{\lambda}{x+1} p(x), \quad x = 0, 1, 2, \dots$$

$$\text{since } p(0) = e^{-\lambda} = e^{-2.975} = 0.0508$$

Thus

$$p(1) = \lambda p(0) = 0.1524$$

$$p(2) = \frac{\lambda}{2} p(1) = 0.2286$$

$$p(3) = \frac{\lambda}{3} p(2) = 0.2286$$

$$p(4) = \frac{\lambda}{4} p(3) = 0.1714$$

$$p(5) = \frac{\lambda}{5} p(4) = 0.1028$$

$$p(6) = \frac{\lambda}{6} p(5) = 0.0514$$

$$p(7) = \frac{\lambda}{7} p(6) = 0.022$$

Now we compute the expected frequencies (E_i) by the relation $E_i = n p(x_i)$, for $i = 0, 1, 2, \dots, 7$ and present the calculation in the following table.

No. of arrivals x	Observed Frequency 0 or f	fx	p(x)	E = n p(x)	(O-E) ²	$\frac{(O-E)^2}{E}$
0	10	0	0.0508	10.16 $\simeq$ 10	0	0
1	30	30	0.1524	30.48 $\simeq$ 30	0	0
2	40	80	0.2286	45.72 $\simeq$ 46	36	0.90
3	50	150	0.2286	45.72 $\simeq$ 46	16	0.32
4	35	140	0.1714	34.28 $\simeq$ 34	1	0.0286
5	20	100	0.1028	20.46 $\simeq$ 20	0	0
6	10	60	0.0514	10.28 $\simeq$ 10	0	0
7	5	35	0.022	4.4 $\simeq$ 4	1	0.25
Total	200	595	1.00	200		1.449

The calculated value of chi-square is given by

$$\chi_{cal}^2 = \sum \frac{(O-E)^2}{E} = 1.449$$

The degree of freedom = $v = k - r - 1 = 8 - 1 - 1 = 6$. For $\alpha = 0.05$, the tabulated value of χ^2 with 6 d.f. is 12.59 which is greater than calculated value of χ^2 , hence we accept H_0 . Thus we may conclude that number of arrivals follow Poisson distribution.

Remark : To fit the binomial distribution we first estimate the parameter p through $p = \frac{\bar{x}}{n}$, where $\bar{x} = \frac{\sum fx}{\sum f}$ and then compute $p(x_i)$ through $p(x+1) = \frac{n-x}{x+1} \frac{p}{q} p(x)$

The other procedure remains unchanged.

(III) Chi-square Test For Independence

Some times it is required to find out whether two characteristics or attributes manifest themselves independently or in some related way. For testing the independence of two attributes, we first formulate a contingency table as discussed in section 6.5 of lesson 6.

Suppose we have two attributes A and B. Further let attribute A has r mutually exclusive categories $A_1, A_2, \dots, A_r$ and B has c such categories denoted by $B_1, B_2, \dots, B_c$. A random sample of size n is drawn from the population and is classified into following (r×c) contingency table with respect to attributes A and B :

Contingency Table

A \ B	B						Total
	B_1	B_2	.	.	.	B_c	
A_1	O_{11}	O_{12}	.	.	.	O_{1c}	$O_{1.}$
A_2	O_{21}	O_{22}	.	.	.	O_{2c}	$O_{2.}$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
A_r	O_{r1}	O_{r2}	.	.	.	O_{rc}	$O_{r.}$
Total	$O_{.1}$	$O_{.2}$	.	.	.	$O_{.c}$	n

To test the hypothesis of independence of two attributes, we use the χ^2 test statistic

$$\chi^2 = \sum \sum (O_{ij} - E_{ij})^2 / E_{ij}, \quad i = 1, 2, \dots, r \text{ and } j = 1, 2, \dots, c. \text{ or simply}$$

$$\chi^2 = \sum (O - E)^2 / E$$

where $E_{ij} = \frac{O_{i.} \cdot O_{.j}}{n}$ for all i and j.

The degree of freedom would be $\nu = (r-1)(c-1)$.

If $r = c = 2$, the above contingency table reduced to 2×2 contingency table. If following is a 2×2 contingency table

A \ B	B		Total
	1	2	
1	a	b	a+b
2	c	d	c+d
Total	a+c	b+d	n

where $n = a + b + c + d$. Thus to test the independence, we use following χ^2 - statistic

$$\chi^2 = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} \sim \chi_1^2$$

Further if any cell frequency is less than 5 and total (n) is less than 50, then we use corrected (Yates correction) 2 - statistic which is given by

$$\chi^2 = \frac{n \left[|ad - bc| - \frac{n}{2} \right]^2}{(a+b)(c+d)(a+c)(b+d)}$$

Example 5. The following contingency table gives the results of the survey conducting by market researcher with respect to the performance of four competing brands of tooth paste among the users:

	Brand				
Performance	A	B	C	D	Total
No cavities	5	9	13	7	34
1-5 cavities	59	66	81	78	284
More than 5 cavities	24	33	44	33	134
Total	88	108	138	118	452

Test the hypothesis that incidences of cavities is independent of the brand of the toothpaste used.

Solution: Here we wish to test the hypothesis

H_0 : The incidence of cavities is independent of the brand of the toothpaste. against the suitable alternative hypothesis.

Now we compute E_{ij} as

$$E_{11} = \frac{88 \times 34}{452} = 6.62 \simeq 7, \quad E_{12} = \frac{108 \times 34}{452} = 8.12 \simeq 8$$

$$E_{13} = \frac{138 \times 34}{452} = 10.38 \simeq 10, \quad E_{14} = \frac{118 \times 34}{452} = 8.87 \simeq 9$$

$$E_{21} = \frac{88 \times 284}{452} = 55.29 \simeq 55, \quad E_{22} = \frac{108 \times 284}{452} = 67.85 \simeq 68$$

$$E_{23} = \frac{138 \times 284}{452} = 86.71 \simeq 87, \quad E_{24} = \frac{118 \times 284}{452} = 74.14 \simeq 74$$

$$E_{31} = \frac{88 \times 134}{452} = 26.08 \simeq 26, \quad E_{32} = \frac{108 \times 134}{452} = 32.01 \simeq 32$$

$$E_{33} = \frac{138 \times 134}{452} = 40.91 \simeq 41, \quad E_{34} = \frac{118 \times 134}{452} = 34.98 \simeq 35$$

Therefore

$$\begin{aligned} \chi^2 = & \frac{(5-7)^2}{7} + \frac{(9-8)^2}{8} + \frac{(13-10)^2}{10} + \frac{(7-9)^2}{9} + \frac{(59-55)^2}{55} \\ & + \frac{(66-68)^2}{68} + \frac{(81-87)^2}{87} + \frac{(78-74)^2}{74} + \frac{(24-26)^2}{26} \\ & + \frac{(33-32)^2}{32} + \frac{(44-41)^2}{41} + \frac{(33-35)^2}{35} \end{aligned}$$

$$= \frac{4}{7} + \frac{1}{8} + \frac{9}{10} + \frac{4}{9} + \frac{16}{55} + \frac{4}{68} + \frac{36}{87} + \frac{16}{74} + \frac{4}{26} + \frac{1}{32} + \frac{9}{41} + \frac{1}{35}$$

$$= 1.857$$

Here $r = 3$, $c = 4$, so $v = (r-1)(c-1) = 6$

For $\alpha = 0.05$, $\chi^2_{0.05, 6} = 12.59$ which is greater than calculated value of 1.857. Hence, we accept H_0 , that is cavities is independent of brand of toothpaste.

Example 6. In an experiment on the immunization of goats from anthrax, the following results were obtained. Comment on the efficiency of the vaccine.

	Died	Survived	Total
Immunized	2	10	12
Not immunised	6	6	12
Total	8	16	24

Solution : Here we wish to test the null hypothesis that vaccine is not effective in controlling the anthrax. Since one cell frequency is less 5 and total (n) < 50, so we use Yates corrected χ^2 -statistic

$$\chi^2 = \frac{n \left[|ad - bc| - \frac{n}{2} \right]^2}{(a+b)(b+d)(a+c)(c+d)}$$

$$= \frac{24 \left[\left| 12 - 60 \right| - \frac{24}{2} \right]^2}{12 \times 12 \times 8 \times 16}$$

$$= \frac{24 \times (36)^2}{12 \times 12 \times 8 \times 16}$$

$$= 1.687$$

If $\alpha = 0.05$, then $\chi^2_1 = 3.84$ which is more than calculated χ^2 . Hence we accept H_0 . That is immunization of vaccine is effective.

Example 7. The following table gives the number of accounting clerks committing errors and not committing errors among trained and untrained clerks working in a company.

	Clerks	Number of clerks	
	Committing error	Not Committing error	Total
Trained	80	540	620
Untrained	165	755	920
Total	245	1295	1540

Test whether the training is effective in preventing errors.

Solution : We wish to test the H_0 ;

H_0 : Training and committing errors are independent against suitable alternative The test statistic is

$$\chi^2 = \frac{n(ad - bc)^2}{(a + b)(b + c)(a + c)(b + d)}$$

$$= \frac{1540(80 \times 755 - 540 \times 165)^2}{620 \times 920 \times 245 \times 1295}$$

$$= 7.01$$

If $\alpha = 0.05$, then $\chi^2_1 = 3.84$, which is less than χ^2 calculated (7.01), so we reject H_0 . Hence we may conclude that training is effective in preventing the errors.

13.4 LET US SUM UP

- (i) To test the population variance, i.e., $H_0 : \sigma^2 = \sigma_0^2$

The test statistic is

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

- (ii) To test the goodness of fit the test statistic is

$$\chi^2 = \sum (O - E)^2 / E \sim \chi_{k-r-1}^2$$

where O = observed frequency

E = expected frequency

K = number of classes

r = number of parameters to be estimated.

- (iii) To test the independent of two attributes in a $r \times c$ contingency table, the test statistic is

$$\chi^2 = \sum \frac{(O - E)^2}{E} \sim \chi_{(r-1)(c-1)}^2$$

13.5 GLOSSARY

- Chi- square is a test in statistics used to measure the difference between observed and expected data.
- 'O' denotes observed frequencies.
- 'E' denotes expected frequencies.
- 'χ' Square works on a large sample only.

13.6 SELFASSESSMENT QUESTIONS

1. The number of defects per unit in a sample of 340 units of a manufactured product was found as follows :

No. of defects :	0	1	2	3	4
------------------	---	---	---	---	---

No. of units :	210	95	24	6	5
----------------	-----	----	----	---	---

Fit a Poisson distribution to the data and test the goodness of fit.

2. If 4 coins are tossed 150 times and following results were obtained.

No. of heads :	0	1	2	3	4
----------------	---	---	---	---	---

Observed freq :	6	48	60	28	8
-----------------	---	----	----	----	---

Fit the binomial distribution and test the goodness of fit.

3. An automobile company conducted a survey about liking or disliking a model of new car. The following data is obtained :

Persons	Age group			
	Below 20	Between 20-30	Between 30-40	Between 40-50
Liked the model	160	70	50	70
Disliked model	80	20	60	80

Analyse the data and give your comments.

4. A random sample of 10 students has the marks : 60, 70, 50, 40, 80, 70, 60, 90, 100 and 50. Do these data support the assumption at 5 % level that the variance of normal population is 187?
5. From the data given below about the treatment of 250 patients suffering from a disease, state whether the new treatment is superior to the conventional treatment.

	No. of Patients		
	Favourable	Not Favourable	Total
New	140	30	170
Conventional	60	50	80
Total	200	50	250

6. In a survey of 240 boys, of which 80 were intelligent, 50 had skilled fathers; while 85 of unintelligent boys had unskilled fathers. Do these figures suppose the hypothesis that skilled fathers have intelligence boys?

13.7 LESSON END EXERCISE

1. The formula for calculating χ square is χ _____
- No estimated or theoretical cell frequency must be less than 5 (True/ False)
- Chi- square test is also known as test of Goodman of fit (True/ False)

13.8 SUGGESTED READINGS

1. Gupta, S.P. (2001): Statistical Methods Sultan Chand & Sons, New Delhi.
2. Srivastava, U.K, G.V. Shenoy and S.C. Sharma. (2024). Quantitative Techniques for Managerial Decision Making. Wiley Eastern United, New Delhi, 4th Edition.

13.9 ANSWERS TO CHECK YOUR PROGRESS

1. It is a non-parametric test.
2. If there is an association between two categorical variables.
3. Right-skewed
4. Expected frequencies should generally be
5. 1
6. Upper-tail test

F-TEST**STRUCTURE**

- 14.0 Objectives
- 14.1 Introduction
- 14.2 Test of Equality of two variances
- 14.3 Analysis of Variance
- 14.4 Let us sum up
- 14.5 Glossary
- 14.6 Self-Assessment Questions
- 14.7 Lesson End Exercise
- 14.8 Suggested Readings
- 14.9 Answers to check your progress

14.0 OBJECTIVES

After successful completion of this lesson, you will be to:

- understand the assumptions and applications of F-test,
- find out whether the two population variances differ significantly or not, and
- perform the technique of analysis of variance.

14.1 INTRODUCTION

While applying student's t-test for testing equality of means, the basic assumption was that the

population variances must be same. Thus F-test is used to find out whether the two independent estimates of population variance differ significantly, or whether the two samples may be drawn from the normal populations having the equal variance.

F-test can also be used to test the equality of several population means through analysis of variance (ANOVA) technique and significance of multiple correlation coefficient. Testing of equality of two variances is discussed in the following sections.

14.2 TEST OF EQUALITY OF TWO POPULATION VARIANCES

Suppose we have two independent random samples of sizes n_1 and n_2 from normal populations and null hypothesis to be tested is that the population variances are same. Under the null hypothesis $H_0: \sigma_1^2 = \sigma_2^2$, the test statistic would be

$$F = \frac{s_1^2}{s_2^2} \quad ; \quad s_1^2 > s_2^2$$

$$\text{where } s_1^2 = \frac{1}{n_1 - 1} \sum (x_1 - \bar{x}_1)^2 \quad \text{and} \quad s_2^2 = \frac{1}{n_2 - 1} \sum (x_2 - \bar{x}_2)^2$$

The distribution of F is F with $n_1 - 1$ and $n_2 - 1$ degree of freedom.

If calculated value of F exceeds the tabulated value with $n_1 - 1$ and $n_2 - 1$ df at α percent level of significance, we reject H_0 . The table values of F are given in Table No. at the end.

Since F-test is based on the ratio of two variances, thus it is also called the Variance Ratio Test. The ratio of two variances follows, F-distribution, named after the famous statistician R.A. Fisher.

F-test is also based on, as other tests, the following assumptions: -

- 1. Normality:** The observations in each group must be normally distributed.
- 2. Homogeneity:** The variance within each group should be equal for all groups. This assumption is needed in order to combine or pool the variances within the groups into a single 'within groups' source of variation.
- 3. Independent of Error:** The error, variation of each value around its own group mean, should be independent for each value.

Now we shall illustrate F-test by taking some examples: -

Example 1. Two samples are drawn from two normal populations. From the following data test whether the two samples have been drawn from the populations with same variances at 5 percent level of significance.

Sample I: 3 5 4 7 6 5

Sample II: 4 6 7 6 8 5 3 9

Solution : Let us formulate the hypothesis

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ against } H_1 : \sigma_1^2 \neq \sigma_2^2$$

$$\text{Here } n_1 = 6, n_2 = 8, \bar{x}_1 = \frac{30}{6} = 5 \text{ and } \bar{x}_2 = \frac{48}{8} = 6$$

$$\begin{aligned} \text{so that } s_1^2 &= \frac{1}{n_1 - 1} \sum (x_1 - \bar{x}_1)^2 = \frac{1}{5} [(3 - 5)^2 + (5 - 5)^2 + (4 - 5)^2 + \\ &\hspace{25em} (7 - 5)^2 + (6 - 5)^2 + (5 - 5)^2 +] \end{aligned}$$

$$= \frac{4 + 0 + 1 + 4 + 1 + 0}{5} = \frac{10}{5} = 2$$

$$\begin{aligned} \text{and } s_2^2 &= \frac{1}{n_2 - 1} \sum (x_2 - \bar{x}_2)^2 \\ &= \frac{1}{7} [(4 - 6)^2 + (6 - 6)^2 + (7 - 6)^2 + (6 - 6)^2 + (8 - 6)^2 \end{aligned}$$

$$(5 - 6)^2 + (3 - 6)^2 + (9 - 6)^2]$$

$$= \frac{4 + 0 + 1 + 0 + 4 + 1 + 9 + 9}{7} = \frac{28}{7} = 4$$

Since $s_2^2 > s_1^2$, so our test statistic is

Since $\alpha = 0.05$, $n_1 - 1 = 5$ and $n_2 - 1 = 7$, so tabulated value of F with 7 and 5 d.f. at 5 percent level is 4.9 which is greater than calculated value of F (2), so we accept H_0 . Thus we may conclude that population variances are equal.

Example 2. A sample of 18 observations gives the sum of square of deviations from mean 96.48. Another sample of 12 observations gives the sum of square of deviations from mean 110.80. Test whether the samples are drawn from normal populations with equal variances.

Solution : We wish to test the hypothesis

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ against } H_1 : \sigma_1^2 \neq \sigma_2^2$$

$$\text{Here } n_1 = 18, n_2 = 12$$

$$\sum (x_1 - \bar{x}_1)^2 = 96.48 \text{ and } \sum (x_2 - \bar{x}_2)^2 = 110.80$$

$$s_1^2 = \frac{96.48}{17} = 5.68 \text{ and } s_2^2 = \frac{110.80}{11} = 10.07$$

since $s_2^2 > s_1^2$, hence

$$= 1.77$$

Since $\alpha = 0.05$, $n_1 - 1 = 17$ and $n_2 - 1 = 11$, thus tabulated value of F with 11 and 17 d.f. at $\alpha = 0.05$ is 2.68 which is greater than F calculated (1.77), thus we accept H_0 . That is population variances are same.

Example 3. The following figures relate to the number of limits produced per shift by two workers A and B for a number of days :

A : 19 22 24 27 24 18 20 19 and 25

B : 26 37 40 35 30 30 40 26 30 35 and 45

Can it be inferred that A is more stable worker compared to B? Answer using F-test at 5% level of significance.

Solution. As we know that a series having lesser variance is treated as more stable. Thus, for drawing the needed inference, we test the equality of the variance of the populations from which samples are observed. Let α_{A^2} and α_{B^2} the population variances of the number of units produced by workers A and B respectively.

Thus, we wish to test $H_0 : \alpha_{A^2} = \alpha_{B^2}$ against $H_1 : \alpha_{B^2} > \alpha_{A^2}$

To test H_0 , the test statistic is

$$F = \frac{S_{B^2}}{S_{A^2}} \quad (\text{As } S_{B^2} > S_{A^2}) \quad \dots (i)$$

Table Showing Calculations

Units produced by A (x)	$(x - \bar{x}) =$ (x-22)		Units produced by B (y)	$(y - \bar{y}) =$ (y-34)	$(y-34)^2$
19	-3	9	26	-8	64
22	0	0	37	+3	9
24	2	4	40	6	36
27	5	25	35	1	1
24	2	4	30	-4	16
18	-4	16	30	-4	16
20	-2	4	40	6	36
19	-3	9	26	-8	64
25	+3	9	30	-4	16
			35	1	1
			45	11	121
$\Sigma x = 198$	$\Sigma(x - \bar{x}) = 0$	$\Sigma(x - \bar{x})^2 = 80$	$\Sigma y = 374$	$\Sigma(y - \bar{y}) = 0$	$\Sigma(y - \bar{y})^2 = 380$

$$\frac{\Sigma x}{n_1} = \frac{198}{9}$$

$$S_{A^2} = \frac{1}{n_1 - 1} \Sigma(x - \bar{x})^2 = \frac{80}{9 - 1} = 10$$

$$S_{B^2} = \frac{1}{n_2 - 1} \Sigma(y - \bar{y})^2 = \frac{380}{11 - 1} = 38$$

Putting the values in (i),

$$F = \frac{38}{10} = 3.8$$

Also the critical value of F at 5% level of significance and for $\nu_1=10$ and $\nu_2=8$ degrees of freedom in $F_{10, 8}(0.05)=3.35$ [Table 3]. Since calculated value of F is greater than $F_{10, 8}(0.05)$, so H_0 is rejected at 5% level and, in the light of data information, $\alpha_{B^2} > \alpha_{A^2}$, i.e., A is more stable worker.

14.3 ANALYSIS OF VARIANCE

The student t-test is used for testing the hypothesis of equality of two normal population means when sample size is small. However, in testing of equality of more than two means, t-test cannot be used. In such situations we use analysis of variance. The analysis of variance, developed by R.A. Fisher, is one of the most powerful tools of statistical analysis. The analysis of variance is a method of splitting the total variation into different components that measure different sources of variation.

Classification of Observations: The following criteria of classification of observations are used in the analysis of variance:

One-way classification: In one-way classification, the observations are classified according to only one criterion. There are two types of variations in the data namely, the variation between samples and within samples. If the variation within the samples and between the sample do not differ from each other's, the samples are said to be belong to the same population. On the other hand, the larger variation between the samples as compared to variation within the samples indicates that the samples come from divergent populations.

Two-way classification: In a two-way classification where the observations are classified according to two factors and the number of observations in each cell is one, we partition the total variation in the data into three components namely the variation due to factor first and second and error and then test the significance of each factor with the variance due to error.

Assumptions in Analysis of Variance

The ratio of variation between samples obtained during the analysis of variance follows F-distribution. For the validity of the F-test in analysis of variance, the following assumptions are made:

- (i) The observations should be independent.
- (ii) The parent population from which observations are taken should be normal.

- (iii) The variances for all the populations from which samples are taken should be equal.
- (iv) Various treatment and environmental effects should be additive in nature.

In the following sections we will discuss the analysis and computation of this technique for one-way and two-way classification.

Analysis and Computation

New we shall present the analysis of variance separately for both one-way and two-way classifications.

(I) One-way classification : Suppose there are k normal populations with means $\mu_1, \mu_2, \dots, \mu_k$ and common variance σ^2 . Further, let k - random samples, one from each population, are drawn from these populations. Let n_i ($i = 1, 2, \dots, k$) be the size of sample from i th population. Using sample information we wish to test

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k, \text{ i.e.}$$

H_0 : There is no significance difference between the population means.

Against

H_1 : Atleast two means are not equal

Let X_{ij} ($i = 1, 2, \dots, k, j = 1, 2, \dots, n_i$) be the j th observation of i th sample, then one-way classified data can be arranged as :

Sample		Observations				Sample	Sample
No						Totals	Means
1	x_{11}	x_{12}	$\dots$		x_{1n_1}	T_1	$\overline{X}_1$
2	x_{21}	x_{22}	$\dots$		x_{2n_2}	T_2	$\overline{X}_2$
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
k	x_{k1}	x_{k2}	$\dots$		x_{kn_k}	T_k	$\overline{X}_k$
Over all total or mean				T	$\overline{X}$		

To test the above cited hypothesis, the following are the main steps :

Step-I Complete the above table

Step-II Determine the correction factor (C.F.), as

$$CF = \frac{T^2}{N} = \frac{\text{Square of grand total}}{\text{Total no. of observations}}$$

Step-III Find the total sum of squares (T.S.S.) as

$$T.S.S. = \sum_i \sum_j X_{ij}^2 - CF$$

$$= \text{Sum of squares of all the observations} - CF$$

Step-IV Find the sum of squares due to factor which is also called between sum of squares (B S S) as

$$BSS = \sum_i \frac{T_i^2}{n_i} - C.F.$$

$$= \text{Sum of squares of totals after dividing it by no. of observations in that sample} - CF$$

Step-V Find within sample sum of square which is also known error sum of squares (ESS) as $ESS = TSS - BSS$

Step-VI Determine the degree of freedom for different components as

$$\text{Degree of Freedom (d.f.) for TSS} = N-1$$

$$\text{Degree of Freedom for BSS} = k-1$$

$$\text{Degree of Freedom for ESS} = N-k$$

Step-VII Obtain the variance within samples and variance between samples by dividing the sum of squares of each by its corresponding degree of freedom, which is also called mean sum of squares (MSS).

Step-VIII Compute F-Value as

$$F = \frac{\text{Variance between samples}}{\text{Variance within samples}}$$

Step-IX Compare the value of F obtained in step VIII with the tabulated value of F under given d.f.'s at the desired level of significance (These values are given in Table - 3). If calculated value of F is greater than the tabulated value, we reject H_0 , otherwise accept H_0 .

It is customary to summarise the various steps obtained in the form of a table called analysis of variance (ANOVA) table

ANOVA TABLE

Source of Variation	Sumn of squares (SS)	d-f	Mean sum of square MSS	Calculated value of F	Table value of F
Between samples	BSS	k-1	$\frac{BSS}{k-1} = V_1$, say	$\frac{V_1}{V_2}$	$F_{k-1, N-k}(\alpha)$
Within samples or error	ESS	N-k	$\frac{ESS}{N-k} = V_2$, say		
Total	TSS	N-1	...	...	...

Example 4. Following are the yields obtained in qtls of three varieties of wheat A, B and C sown in 14 plots :

Variety of Wheat	Yields in different plots				
A	12	13	14	13	
B	10	9	10	9	9
C	13	14	13	12	14

Is there any significance difference in the production of three varieties of wheat.

Solution : First of all we formulate the null and alternative hypothesis as

H_0 : There is no significance difference in the production of three varieties i.e.

$$H_0 : \mu_A = \mu_B = \mu_C$$

Against

H_1 : Atleast two of μ_A , μ_B and μ_C differ significantly.

We shall analyse the above data using analysis of variance technique.

Variety	Yields in Plots					Total	Mean
A	12	13	14	13	—	52	13
B	10	9	10	9	9	47	9.4
C	13	14	13	12	14	66	13.2
Total	35	36	37	34	23	165	

$$CF = 1944.64$$

$$\text{Total sum of square (TSS)} = \sum \sum X_{ij}^2 - CF$$

$$\begin{aligned}
 &= 12^2 + 13^2 + 14^2 + 13^2 + 10^2 + 9^2 + 10^2 + 9^2 + 9^2 + 13^2 \\
 &\quad + 14^2 + 13^2 + 12^2 + 14^2 - 1944.64 \\
 &= 1995 - 1944.64 \\
 &= 50.36
 \end{aligned}$$

Sum of squares due to variety or between sum of squares (SSB)

$$\begin{aligned}
 SSB &= \sum \frac{T_i^2}{n_i} - CF \\
 &= \frac{(52)^2}{4} + \frac{(47)^2}{5} + \frac{(66)^2}{5} - 1944.64 \\
 &= 1989 - 1944.64 = 44.36
 \end{aligned}$$

Sum of squares due to error (SSE)

$$\begin{aligned}
 &= TSS - SSB \\
 &= 50.36 - 44.36
 \end{aligned}$$

$$= 6.00$$

Here $k = 3$, $N = 14$

Now we summarise the above calculations in the form of following ANOVA Table

ANOVA Table					
Source of Variation	Sum of Squares	d.f	MSS	F Calculated	F Tabulated
Between Variety	44.36	2	$\frac{44.36}{2} = 22.18$		$F_{2,11} (5\%)$
				$\frac{22.18}{0.546} = 40.62$	4.26
Within Varieties	6.00	11	$\frac{6.00}{11} = 0.546$		
Total	50.36	13			

Since tabulated value of F at 5% level of significance for (2, 11) d.f. is 4.26 which is less than the calculated value of F, hence we reject the H_0 . Thus the varieties differ significantly

Note : The analysis of variance technique is independent of the shift of origin. If the individual observations are large values we can subtract any constant quantity from each individual value and then perform ANOVA. The calculated value of F will remain the same and the computational work will be reduced.

Example 5. The following data shows the lives in hours of four batches of electric tubes :

	Batches				Lives in hours			
A	1700	1710	1750	1780	1800	1820	1900	
B	1680	1740	1740	1800	1850			
C	1560	1650	1700	1720	1740	1760	1840	1920
D	1610	1620	1630	1670	1700	1780		

Perform the analysis of variance and show that four batches are homogeneous.

Solution : First of all we formulate the H_0 as H_0 : Four batches are Homogenous

Since the values are large, so we shift the origin by subtracting 1740 from each observations. Thus the above data reduced to

Batches	Lives in hours								Total T_i
A	-40	-30	10	40	60	80	160	—	280
B	-60	0	0	60	110	—	—	—	110
C	-180	-90	-40	-20	0	20	100	180	- 30
D	-130	-120	-110	-70	-40	40	—	—	- 510
Total								- 150	

Total sum of squares (TSS)

$$= \sum \sum X_{ij}^2 - C.F.$$

$$= (-40)^2 + (-30)^2 + (10)^2 + \dots + (-40)^2 + (40)^2 - 865 \cdot 385$$

$$= 195900 - 865 \cdot 385 = 195034.615$$

Between (Batches) Sum of Square (BSS)

$$= \sum \frac{T_i^2}{n_i} - C.F.$$

$$= \frac{(280)^2}{7} + \frac{(110)^2}{5} + \frac{(-30)^2}{8} + \frac{(-510)^2}{6} - 865 \cdot 385$$

$$= 57082 \cdot 5 - 865 \cdot 385$$

$$= 56217 \cdot 115$$

Sum of square due error (SSE)

$$= TSS - BSS$$

$$= 195034 \cdot 615 - 56217 \cdot 115$$

$$= 138817.50$$

Here $k = 4$, $N = 26$

Now we summarise the above calculations in the form of ANOVA Table as

AVONA Table

Source of	Sum of	d.f	MSS	F-value	F-value
Variation	Squares			Calculated	Tabulated
Between Batches	56217.115	3	18739.04		$F_{3, 22} (5\%)$
				$\frac{18739.04}{6309.89}$	
Error	158817.50	22	6309.89	$= 2.97$	$= 3.05$
Total	195034.615	25	—	—	—

Since the calculated value of $F (2.97)$ is less than the tabulated value of $F (3.05)$ at 5 percent level of significance with 3 and 22 d.f., so we accept our H_0 . Hence, we may conclude that four batches are homogeneous.

Two-way Classification: As discussed earlier in two-way classification the set of observations are classified according to two factors. Thus, such a classification can be presented in the form of a rectangular array in which rows represent one factor of classification and columns represent a second factor of classification. In general, let there be r rows and c -columns. Let X_{ij} denotes the value of the i th row and j th column ($i=1, 2, \dots, r$ and $j=1, 2, 3, \dots, c$).

Thus, a general two-way classification model can be represented as:

Columns Rows							Total
	1	2	...	j	...	c	
1	X_{11}	X_{12}	...	X_{1j}	...	X_{1c}	$T_{1.}$
2	X_{21}	X_{22}	...	X_{2j}	...	X_{2c}	$T_{2.}$
.	.	.	...	...	...	.	.
.	.	.	...	...	...	.	.
.	.	.	...	...	...	.	.
i	X_{i1}	X_{i2}	...	X_{ij}	...	X_{ic}	$T_{i.}$
.	.	.	...	...	...	.	.
.	.	.	...	...	...	.	.
.	.	.	...	...	...	.	.
r	X_{r1}	X_{r2}	...	X_{rj}	...	X_{rc}	$T_{r.}$
Total	$T_{.1}$	$T_{.2}$	...	$T_{.j}$	...	$T_{.c}$	T

Now with the above sample information, two-way analysis of variance involves the following steps :

Step-I Formulate the two null hypothesis as

(a) H_0 : Row means are equal [i.e. $\mu_1 = \mu_2 = \dots = \mu_r$].

H_1 : Atleast two row means are not equal.

(b) H_0 : Column means are equal, [i.e. $\mu_1 = \mu_2 = \dots = \mu_c$]

H_1 : Atleast two column means are not equal

Step-II Compute the Correction Factor (C.F.)

$$C.F. = \frac{T^2}{rc} = \frac{(\text{Grand Total})^2}{(\text{No. of rows})(\text{No. of columns})}$$

Step-III Compute total sum of squares (TSS) as

$$\text{TSS} = \sum \sum X_{ij}^2 - \text{C.F.}$$

= Sum of squares of observations – C.F.

Step-IV Determine Row sum of squares (RSS) as

$$\text{RSS} = \sum \frac{T_i^2}{c} - \text{C.F.}$$

= Sum of squares of rows total divided by no. of columns – C.F.

Step-V Determine column sum of squares (CSS) as

$$\text{CSS} = \sum \frac{T_j^2}{r} - \text{C.F.}$$

= Sum of squares of columns total divided by no. of rows – C.F.

Step-VI Compute Error Sum of Squares (ESS) as $\text{ESS} = \text{TSS} - \text{RSS} - \text{CSS}$.

Step-VII Determine the degree of freedom (d.f.) associated with various sum of squares by using the following :

- (a) degree of freedom for TSS = $rc-1$
- (b) degree of freedom for RSS = $r-1$
- (c) degree of freedom for CSS = $c-1$
- (d) degree of freedom for ESS = $(r-1)(c-1)$

Step-VIII Compute mean sum of squares (MSS) on dividing the sum of squares by respective degree of freedoms.

Source of variation	Sum of squares	d.f.	Mean sum of squares	Calculated value of F	Tabulated value of F
Row	RSS	$r-1$	$\frac{\text{RSS}}{r-1} = V_1$	$F_1 = \frac{V_1}{V_3}$	$F_{r-1, (r-1)(c-1)}(\alpha)$

Column	CSS	c-1	$\frac{CSS}{c-1} = V_2$	$F_2 = \frac{V_2}{V_3}$	$F_{c-1, (r-1)(c-1)}(\alpha)$
Error	ESS	(r-1)(c-1)	$= V_3$		
Total	TSS	rc-1			

After forming the ANOVA table as above the decisions regarding H_0 and H'_0 are taken as

- (a) If $F_1 \leq F_{r-1, (r-1)(c-1)}(\alpha)$; Accept H_0 , otherwise reject H_0
- (b) If $F_2 \leq F_{c-1, (r-1)(c-1)}(\alpha)$; Accept H'_0 , otherwise reject H'_0 .

Finally after accepting or rejecting H_0 or H'_0 , as the case may be, the conclusion is made accordingly.

Example 6. A company appoints four Marketing Trainees T_1, T_2, T_3 and T_4 and observes their sales in three regions – North, West and South. The figures of sales (in lacs) are given below.

Trainees					
Region		T_1	T_2	T_3	T_4
North	23	26	26	24	
West	27	26	27	26	
South	23	24	25	25	

Carry out the analysis of variance. What conclusions do you draw from the analysis?

Solution : First of all we formulate the following hypothesis :

- (i) H_0 : There is no significance difference between the sales of four Trainees H_1 : Sales of atleast two trainees are not equal.
- (ii) H_0^1 : There is no significance difference between the sales of three region. H_1^1 : Sales of atleast two regions are not equal

To test these hypothesis, we complete the following table and obtain different sum of squares.

Trainees Region					
	T ₁	T ₂	T ₃	T ₄	Total
North	23	26	26	24	99
West	27	26	27	26	106
South	23	24	25	25	97
Total	73	76	78	75	302

Corrector Factor (C.F.)

$$= \frac{(\text{Grand Total})^2}{r \times c} = \frac{(320)^2}{12} = 7600.33$$

Total Sum of Squares (T.S.S)

$$\text{T.S.S.} = \sum \sum X_{ij}^2 - \text{C.F.}$$

$$\begin{aligned}
 &= (23)^2 + (26)^2 + (26)^2 + (24)^2 + (27)^2 + (26)^2 + (27)^2 + \\
 &\quad (26)^2 + (23)^2 + (24)^2 + (25)^2 + (25)^2 - 7600.33 \\
 &= 7622 - 7600.33 \\
 &= 21.67
 \end{aligned}$$

Row Sum of Squares or Sum of Squares due to Regions (R.S.S)

$$\begin{aligned}
 \text{R.S.S.} &= \frac{(99)^2}{4} + \frac{(106)^2}{4} + \frac{(97)^2}{4} - \text{C.F.} \\
 &= 7611.5 - 7600.33 \\
 &= 11.17
 \end{aligned}$$

Column Sum of Squares or sum of squares due to Trainers (C.S.S)

$$\begin{aligned}
 \text{C.S.S.} &= \frac{(73)^2}{3} + \frac{(76)^2}{3} + \frac{(78)^2}{3} + \frac{(75)^2}{3} - \text{C.F.} \\
 &= 7604.67 - 7600.33 \\
 &= 4.37
 \end{aligned}$$

Error Sum of Squares (ESS)

$$ESS = TSS - RSS - CSS$$

$$= 21.67 - 11.17 - 4.37$$

$$= 6.13$$

The following are the degree of freedoms, for different sum of squares :

$$\text{Degree of freedom for TSS} = 12 - 1 = 11$$

$$\text{Degree of freedom for RSS} = 3 - 1 = 2$$

$$\text{Degree of freedom for CSS} = 4 - 1 = 3$$

$$\text{Degree of freedom for ESS} = 11 - 3 - 2 = 6$$

Now we summarise the above calculations in the following ANOVA Table:

ANOVA TABLE					
Source of variation	Sum of squares	d.f.	M.S.S.	F-value calculated	F-value Tabulated
Rows (Regions)	11.17	2	5.59	$\frac{5.59}{1.02} = 5.48$	$F_{2, 6}(5\%) = 5.14$
Columns (Trainees)	4.37	3	1.46	$\frac{1.46}{1.02} = 1.43$	$F_{3, 6}(5\%) = 4.76$
Error	6.13	6	1.02		
Total	21.67	11	—		

The calculated value of F (5.48) is greater than the tabulated value of F (5.14) with 2 and 6 d.f. at 5 percent level of significance, hence we reject H_0 and conclude that regions have significant effect on sales. Further the calculated value of F (1.43) is less than the tabulated value of F (4.76) with 3 and 6 d.f. at 5 percent level of significance, thus we accept H_0 and conclude that there is no significant difference in trainees as far as their sales are concerned.

CHECK YOUR PROGRESS-I

Fill in the Blanks with suitable answers:

1. The F-test is used to compare theof two or more populations.
2. It is called the because it calculates the F-statistic by dividing one sample variance by another.
3. The F-test follows an (or Fisher-Snedecor distribution) under the null hypothesis.
4. In ANOVA, the F-test compares the variability..... groups to the variability groups.
5. An F-test result is always a number because variances are positive.
6. When the F-test is used in ANOVA, the test statistic is calculated as the ratio of to
7. If the calculated F-value is greater than the (or table value), we reject the null hypothesis.
8. The F-test is commonly used to test the overall significance of a model.

14.4 LET US SUM UP

In this lesson you learned the uses and application of F- test, which is used to compare the ratio of variance of two populations. F-Test is also used to test the difference between several populations through (ANOVA) Analysis of variance. ANOVA overcome the limitation of T-test where only two population are studied. In ANOVA multiple population can be considered.

14.5 SELF ASSESSMENT QUESTIONS

1. A test was given to five students taken at random from the eighth class of three schools in a town. The individual scores are:

School	Scores				
I	8	7	9	6	8
II	7	5	4	6	5
III	6	5	4	6	4

Carry out the analysis of variance and state your conclusions.

2. The following figure related to the production (in kg) of three varieties of wheat used in 15 plots :

Variety of	Yield (in kg)					
Wheat						
A	19	22	21	21	3	
B	20	16	18	20	18	19
C	23	21	23	24	20	

Test whether there is any significance difference in the production of three varieties of wheat.

3. The following data represent the number of units of production per day produced by 5 different workers using four different machines.

Workers	Machines			
	M ₁	M ₂	M ₃	M ₄
W ₁	54	48	57	46
W ₂	56	50	62	53
W ₃	44	46	54	42
W ₄	53	48	56	43
W ₅	48	52	59	49

Test whether the

- mean productivity is the same for different machines.
 - the workers differ with respect to mean productivity.
4. Using the data of (2) and (3), carry out the analysis of variance after shifting the origin and state your conclusions.

14.6 GLOSSARY

- F-test is also known as F Ratio test.
- Fisher is the founder of F-test.

- F- test is used to find the ratio between the variance of two proportions.
- ANOVA is a multiple Variable test to compare the multiple variables with the help of F- test.

14.7 LESSON END EXERCISE

1. F- test compares the _____ of two populations.
2. ANOVA is a multi-test.
3. F-test is also called _____

14.8 SUGGESTED READINGS

1. Gupta, S.P. (2001): Statistical Methods. Sultan Chand & Sons, New Delhi.
2. Srivastava, U.K., G.V. Shenoy and S.C., Sharma (1983): Quantitative Techniques for Managerial Decisions Making. Wiley Eastern limited.
3. Levin, R. (1984). Statistics for Managements. Prentice-Hall Inc., New York.

14.9 ANSWERS TO CHECK YOUR PROGRESS

1. The F-test is used to compare the variances of two or more populations.
2. It is called the Variance Ratio Test because it calculates the F-statistic by dividing one sample variance by another.
3. The F-test follows an F-distribution (or Fisher-Snedecor distribution) under the null hypothesis.
4. In ANOVA, the F-test compares the variability between groups to the variability within groups.
5. An F-test result is always a positive number because variances are positive.
6. When the F-test is used in ANOVA, the test statistic is calculated as the ratio of Mean Square Between (MSB) to Mean Square Within (MSW).
7. If the calculated F-value is greater than the critical value (or table value), we reject the null hypothesis.
8. The F-test is commonly used to test the overall significance of a regression model.

INTRODUCTION TO COMPUTER OPERATING SYSTEM

STRUCTURE

- 15.0 Objectives
- 15.1 Introduction
- 15.2 Computer Operating System
- 15.3 Let us sum up
- 15.4 Glossary
- 15.5 Self-Assessment Questions
- 15.6 Lesson End Exercise
- 15.7 Suggested Readings
- 15.8 Answers to check your progress

15.0 OBJECTIVES

After successful completion of this lesson, you shall be able to know: -

- What is operating System
- Purpose of operating System
- Input and output devices.
- Window fundamentals

15.1 INTRODUCTION

With the introduction of computer in statistics research became so easy whether we talk about data collection, writing reports, analysis or anything else. There is general application software like MS- word,

MS Excel etc. for general purpose of writing and analysis. And there are some statistical software's also like SPSS, AMOS, PASW etc. In current lesson we will discuss about operating system used in computer which is also called system software without operating system, computer cannot work at all.

15.2 COMPUTER OPERATING SYSTEM

15.2.1 What is an Operating System ?

An operating system (OS) is a collection of system programs that together control the operation of a computer system.

15.2.2 What does an operating system do ?

An operating system controls the way in which the computer system functions. In order to do this, the operating system includes programs that

- initialize the hardware of the computer system
- provide basic routines for device control
- provide for the management, scheduling and interaction of tasks
- maintain system integrity and handle errors

15.2.3 Where are operating systems found ?

There are many types of operating systems, the complexity of which varies depending upon what type of functions are provided, and what the system is being used for. Some systems are responsible for managing many users on a network. Other operating systems do not manage user programs at all. These are typically found in hardware devices like petrol pumps, airplanes, video recorders, washing machines and car engines.

15.2.4 What is a general-purpose operating system ?

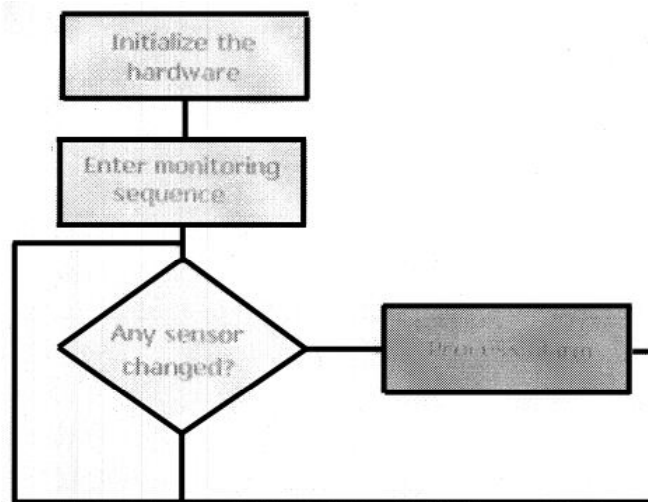
Windows NT Workstation is known as a general-purpose operating system. This is because it provides the ability to run a number of different

programs, such as games, word processing, business applications and program development tools.

15.2.5 A simple operating system for a security control system

An operating system for a security control system (such as a home alarm system) would consist of a number of programs. One of these programs would gain control of the computer system when it is powered on, and initialize the system.

The first task of this initialize program would be to reset (and probably test) the hardware sensors and alarms. Once the hardware initialization was complete, the operating system would enter a continual monitoring routine of all the input sensors. If the state of any input sensor changed, it would branch to an alarm generation routine.



15.2.6 What are Input and Output devices ?

Input and output devices are components that form part of the computer system. These devices are controlled by the operating system.

Input devices provide input signals such as commands to the operating system. These commands received from input devices instruct the operating system to perform some tasks or control its behavior. Typical input devices are a keyboard, mouse, temperature sensor, air-flow valve or door switch.

In the previous example of our simple security control system, the input devices could be door switches, alarm keypad panel and smoke detector units.

Output devices are instruments that receive commands or information from the operating system. Typical output devices are monitor screens, printers, speakers, alarm bells, fans, pumps, control valves, light bulbs and sirens.

15.2.7 What is a single-user operating system ?

Operating systems such as Windows 95, Windows NT Workstation and Windows 2000 professional are essentially single user operating systems. They provide you the capability to perform tasks on the computer system such as writing programs and documents, printing and accessing files.

Consider a typical home computer. There is a single keyboard and mouse that accept input commands,

and a single monitor to display information output. There may also be a printer for the printing of documents and images.

In essence, a single-user operating system provides access to the computer system by a single user at a time. If another user needs access to the computer system, they must wait till the current user finishes what they are doing and leaves.

Students in computer labs at colleges or University often experience this. You might also have experienced this at home, where you want to use the computer but someone else is currently using it. You have to wait for them to finish before you can use the computer system.

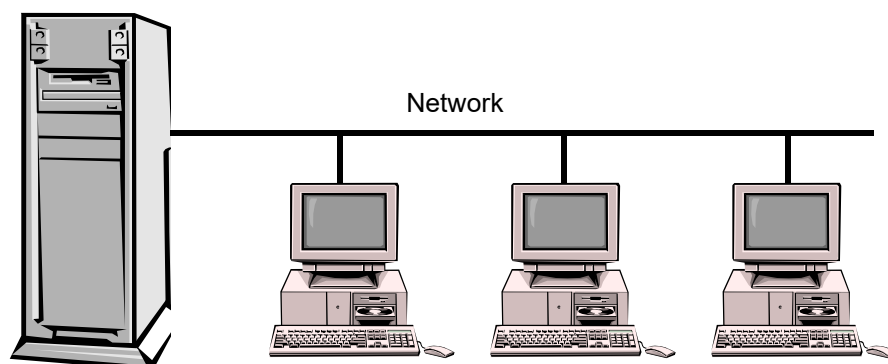
15.2.8 What is a multi-user operating system ?

A multi-user operating system lets more than one user access the computer system at one time. Access to the computer system is normally provided via a network, so that users access the computer remotely using a terminal or other computer.

In the early days of large multi-user computers, multiple terminals (keyboards and associated monitors) were provided. These terminals sent their commands to the main multi-user computer for processing, and the results were then displayed on the associated terminal monitor screen. Terminals were hard-wired directly to the multi-user computer system.

Today, these terminals are generally personal computers and use a network to send and receive information to the multi-user computer system. Example of multi-user operating systems are UNIX, Linux (a UNIX clone) and mainframes such as the IBM AS400.

Mainframe



The operating system for a large multi-user computer system with many terminals is much more complex than a single-user operating system. It must manage and run all user requests, ensuring they do not interfere with each other. Devices that are serial in nature (devices which can only be used by one user at a time, like printers and disks) must be shared amongst all those requesting them (so that all the output

documents are not jumbled up). If each user tried to send their document to the printer at the same time, the end result would be garbage. Instead, documents are sent to a queue, and each document is printed in its entirety before the next document to be printed is retrieved from the queue.

15.2.9 Operating system utilities

The operating system consists of hundreds of thousands of lines of program code and stored on hard disk. Portions of the operating system are loaded into computer system memory (RAM) when needed. Utilities are provided for

- Managing Files and Documents
- Development of Programs and Software
- Communicating between people and with other computer systems
- Managing user requirements for programs, storage space and priority

15.2.10 What is a multi-tasking operating system?

A multi-tasking operating system provides the ability to run more than one program at once. For example, a user could be running a word processing package, printing a document, copying files to the floppy disk and backing up selected files to a tape unit. Each of these tasks the user is doing appears to be running at the same time.

A multi-tasking, operating system has the advantage of letting the user run more than one task at once, so this leads to increased productivity. The disadvantage is that more programs that are run by the user, the more memory that is required.

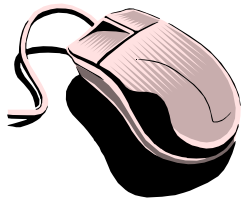
- Basic Features of Graphical Interfaces

Graphical systems use windows to display information and thus allow more than one window to be displayed at any time. Each window is associated with a running program. User input is derived from a keyboard and mouse.

15.2.11 The mouse

The mouse, invented in 1963 at the Stanford Research Institute by Douglas Engelbart, has done much to enhance the use of the personal computer. Engelbart's prototype, made of wood, with metal disks for rollers that detected the motion of the mouse, was further developed by Xerox at its Palo Alto Research Center in the early 1970's under the direction of Jack S Hawley.

Most mice have two or more buttons, which users depress to select items from a menu or click on



graphical objects on the computer screen, thus sending commands to the computer.

The mouse is held in the hand and moved across a flat surface. As the mouse is moved, its movement is detected and translated into both X and Y movements, which updates the indicated position of the mouse pointer on the computer screen accordingly

15.2.12 The mouse cursor

The position of the mouse is shown on the screen as the mouse cursor and is denoted by a number of symbols.

15.2.13 Selecting items with the Mouse

i) Single Click

A single mouse click refers to moving the mouse pointer over the desired item and quickly pressing the left mouse button once.

ii) Double Click

A double mouse click refers to moving the mouse pointer over the desired item and quickly pressing the left mouse button twice in rapid succession.

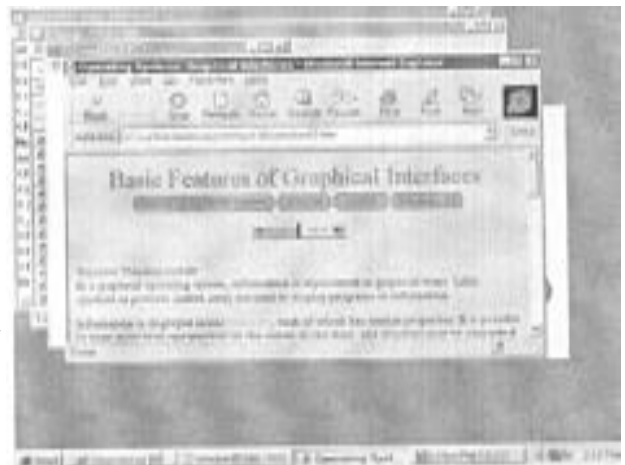
iii) Drag

A drag or move operation is performed by moving the mouse pointer over the desired item and holding the left mouse button down. The mouse is then used to move to drag the object or window to the new position, then the left mouse button is released.

15.2.14 Window Fundamentals

In a graphical operating system, information is represented in graphical ways. Little symbols or pictures (called icons) are used to display programs or information. Information is displayed inside windows, each of which has similar properties.

It is possible to have more than one window on the screen at one time, and windows may be cascaded (on top of one another) or tiled (all displayed at once and all visible).





In this picture, the windows have been cascaded. This makes each window appear on top of each other, one after the other.

The front most window is considered to be the active window, i.e., window to which the users commands will be sent.

In windows 95 or Windows NT, the titlebar of the window is shown in the default color Blue.

In this picture, the images have been **tiled**.

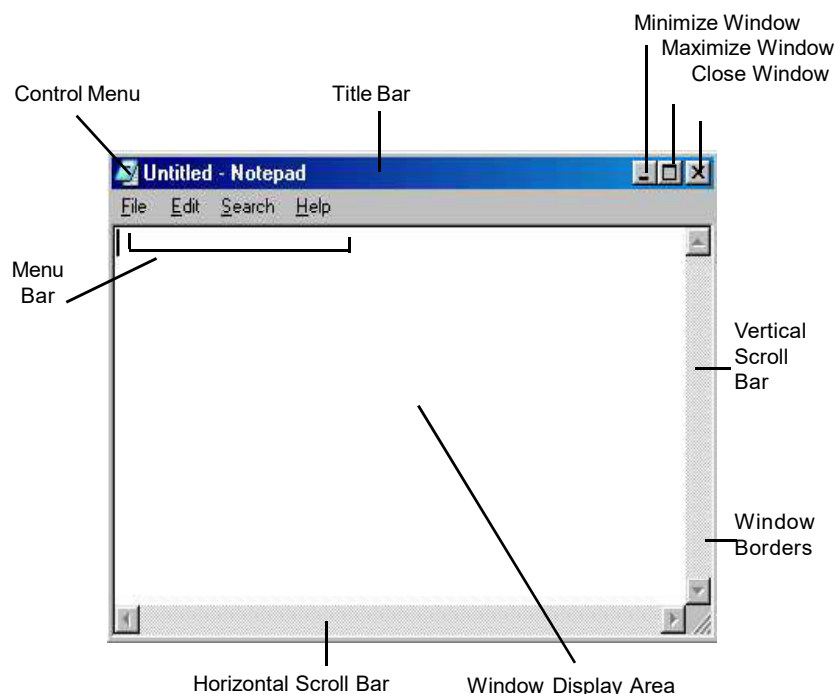
This makes all windows visible at the same time, but resizes the dimensions of each window so that they all fit on the available screen space at once.

Tip : To cascade or tile all windows on the desktop area, right mouse click on an empty portion of the taskbar and select Cascade Windows or Tile Windows from the menu.

- **Window Properties**

Each window has the same properties and behaves the same way. This provides a consistent interface to the user, as all commands are the same for each window and the operations that the user performs on each window are identical.

In the diagram below, we see the basic window as presented by Windows 95 or Windows NT. Each property is listed on the diagram, and below is an explanation for each of the window components.



The Title Bar

This normally displays the name of the program associated with the window. If the background color of the title bar is blue, the window is active and any user commands will be processed by that window. You can also toggle between a maximized window size and the windows normal size by double clicking in the title bar area.

The Control Menu

Clicking on the Control Menu pops up a small Window of selectable options, which include the operations of Restore, Move, Size, Maximize, Minimize and Close the Window.

The Horizontal and Vertical Scroll Bars

When the amount of information displayed in the window exceeds the viewing space of the window, scroll bars are automatically to the side and bottom of the window. This allows the user to scroll the contents of the window in order to view the remaining information. Arrows are used to indicate the direction of scrolling on the scroll bar, and an indicator bar represents the relative position of the viewing area compared to the total size of the information.

Clicking on the arrows associated with the scroll bar move the viewing window up or down one line, or across or back one character position. You can also click on the small indicator bar within the scroll bar and drag it with the mouse to quickly scroll the windows contents.

The Minimize Maximize Close Window Buttons

These buttons are located on the top right corner of each window. Clicking on them once performs the desired action associated with the button.

- The minimize button - reduces the window and places it on the taskbar at the bottom of the window.
- The maximize button ? expands the window to fill the entire desktop screen area.
- The close button × closes the window.

Tip : To minimize all windows on the desktop area, right mouse click on an empty portion of the taskbar and select Minimize all Windows from the menu.

The Menu Bar



The menu bar presents a number of options that the program associated with the Window supports.

Clicking on an option on the menu bar will popup a submenu of choices that you can select from.

The Windows Borders

The windows borders show the dimensions of the window. Any window can be resized, either made smaller or larger, by dragging the window border appropriately.

- **To make the window taller or shorter:**

Move the mouse pointer to either the top or bottom window border, and when it changes to a resize arrow then hold the left mouse button down and drag the window border to its new position, then let release the left mouse button.

- **To make the window narrow or wider:**

Move the mouse pointer to either the left or right window border, and when it changes to a resize arrow then hold the left mouse button down and drag the window border to its new position, then let release the left mouse button.

Moving a Window

A window can be repositioned on the desktop screen display area by moving the mouse cursor into the title bar area, then holding the left mouse button down and dragging the window to the new position then releasing the left mouse button.

Switching between Windows

When you have multiple windows displayed on the desktop screen area, you can switch between windows by clicking on the programs icon on the taskbar or pressing ALT-TAB keys on the keyboard. When you press ALT-TAB, it will pop up a window of the available programs. Hold the ALT key down, and pressing the tab key will move the selection to the next window in the list. When the desired window is highlighted, release the ALT key and that window will become active.

Clicking on the applications icon on the taskbar can also do switching to another application. The following picture shows the Windows taskbar, located at the bottom of the screen.



This picture shows the ALT-TAB pop up window, which list the available programs the user can switch to.



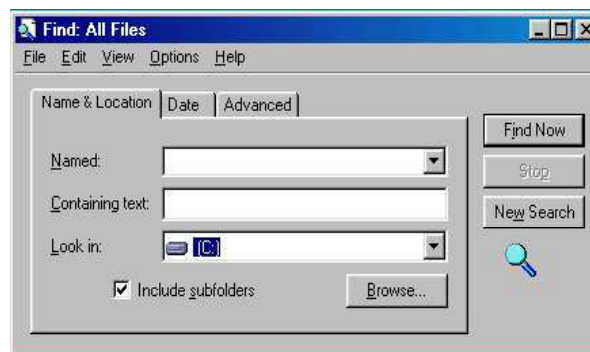
- **Previous/Home Page/ Next**
- **Window objects and components**

This section discusses window options such as buttons and dialog boxes.

- **Text Boxes**

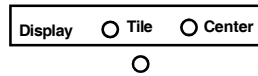
Text boxes allow you to enter text information. To enter text, first click inside the text area using the mouse, and the cursor will change to a vertical flashing bar/ showing you that text can now be entered.

In this image, a text box allows the user to specify a file to find on the computer. The name of the text box entry field is called Named :



- **Radio Buttons**

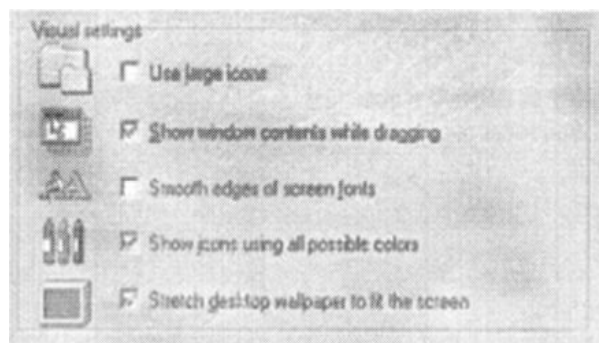
Radio buttons allow users to select one of a number of options from a selection. In the following image, a choice between Tiled and Centered is offered. A radio button is enabled when there is a black dot in its center. A radio button is enabled when it is empty. To enable a radio button, simply click once on it. To disable a radio button that is enabled, simply click once on it. It works like a toggle switch.



- **Check Boxes**

Check boxes allow users to select one or more options from a selection. In the following image, the options Show window contents while dragging, Show icons using all possible colors and Stretch desktop wallpaper to fit the screen are all enabled.

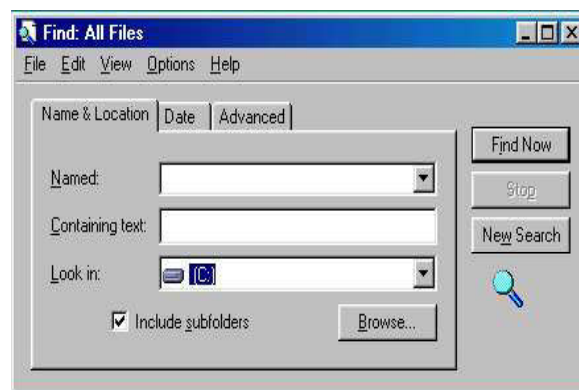
A check box is enabled when it has a tick in it, when a check box is empty, that option is not selected. To enable a check box, simply click once on it. To disable a check box that is enabled, simply click once on it. It works like a toggle switch.



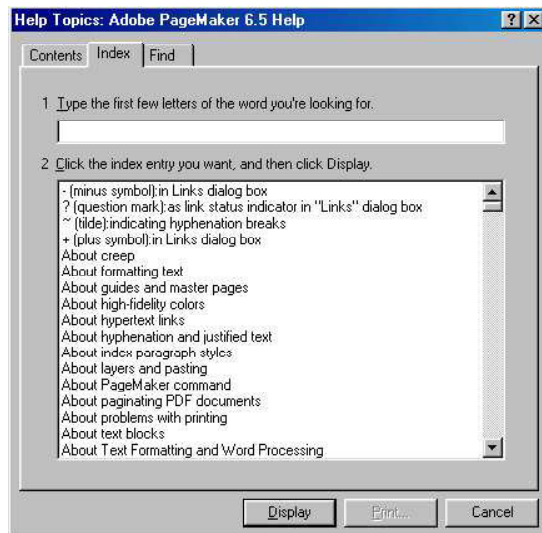
- **Dialog Boxes**

Dialog boxes allow you to make choices and enter data. They combine text boxes with radio buttons and check boxes.

To close a dialog box, press the ESC key.



- **List Boxes**



List boxes present a number of choices. You select one by double-clicking on the item you want. Often

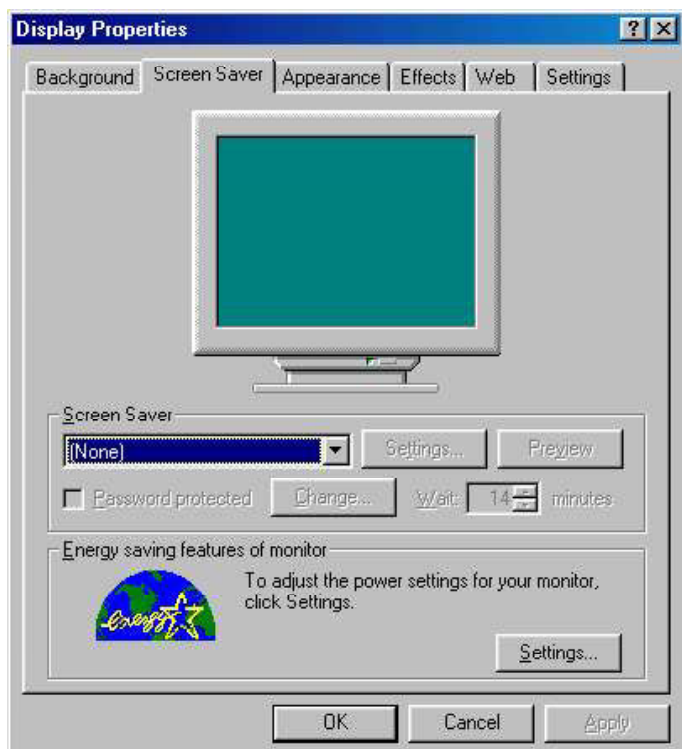
the list of choices is in a scrollable window box.

In this example, the Help dialog box of Windows lists a number of help topics that the user can double-click on to reveal the help associated with that item.

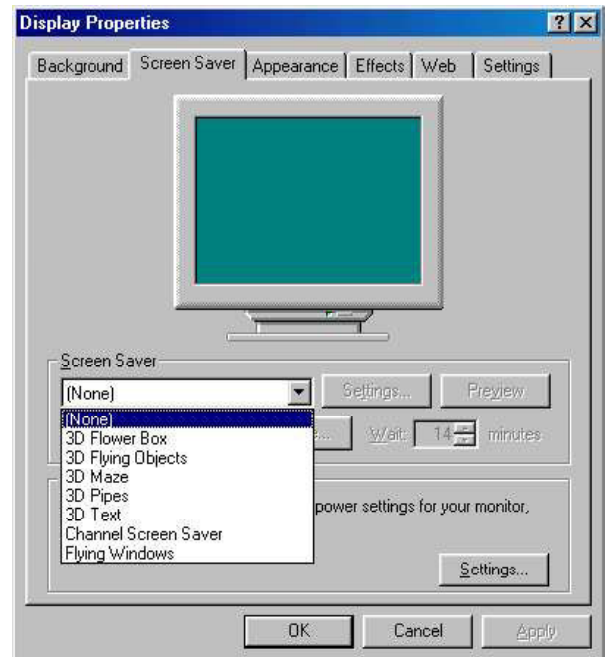
- **Drop Down List Boxes**

To minimize the amount of screen space, list boxes can sometimes be arranged as a drop-down list box. This displays a single item, but when the list box is clicked on, the range of items pops up in a secondary window.

A drop-down list box is shown below. In this example, it is part of the Dialog box associated with the Display Properties.



Notice the symbol at the end of the box. Clicking on this symbol reveals the list of options

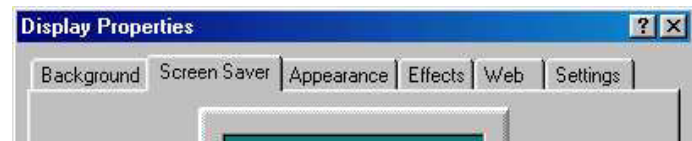


- **Tab Controls**

Tab controls allow a number of different dialog boxes associated with a device to be presented as a single combined control. For instance, if we looked at the screen display in Windows, there are so many things that can be changed, like screen saver, wall-paper, size and resolution, video display driver and so on.

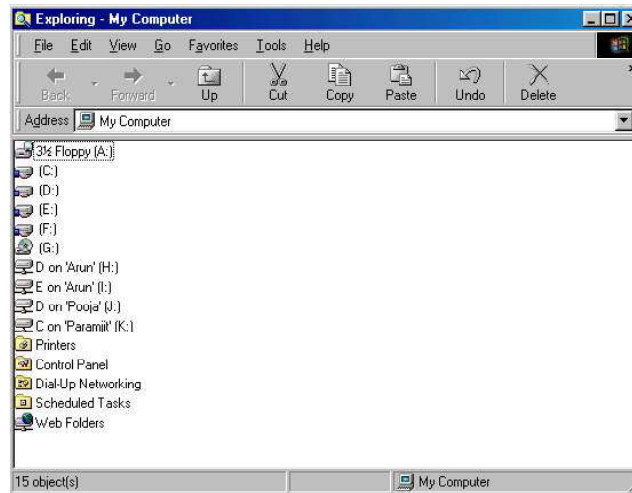
Putting all of these on a single dialog box is cumbersome and there is just not enough screen real estate. So, a number of dialog boxes are used, but they are combined using the tab control. It looks like multiple sections, and each tab has a heading. Clicking on the tab item reveals the dialog box associated with that tab.

In this example, the tab control for the Windows desktop properties is displayed. Note there are FOUR distinct dialog boxes; the current choice is Screen Saver.



- **Toolbars**

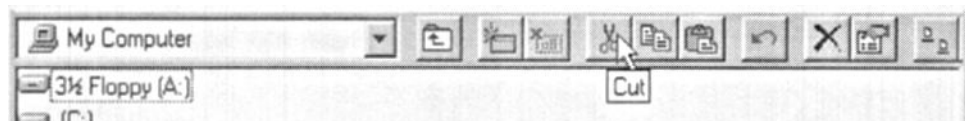
Toolbars appear on a number of windows and application programs. An example is the My Computer window.



The toolbar is displayed underneath the Menu Bar Options of the window. An expanded view looks like.



The toolbar consists of a number of icons (little pictures), each representing a command. As the mouse cursor is moved along each icon, a text description will pop up revealing the available control that is underneath the mouse cursor.



Toolbars provide shortcuts to regularly used operations like cut and paste, close, and Help.

CHECK YOUR PROGRESS-I

Very short answer type questions.

1. What is an Operating System?

2. Write a short note on 'Types of Operating Systems'.

3. What are wildcards? Give example.

Short answer type questions.

1. What are the advantages of using Proprietary operating systems?

2. Give five examples of common file extensions.

3. List the advantages and disadvantages of multitasking.

- 1. What is Windows XP?**

Windows XP is the latest version of Microsoft's graphical user interface. It has several advantages, for the user, over previous versions of Microsoft Windows, these include;

- Windows XP keeps your frequently used programmes within the start menu making it quick and easy to access programmes for a second time.
- All software installed on the computer is now listed alphabetically within the All-Programs menu, making it far easier to locate the piece of software you require.
- Windows XP is faster than previous versions of Windows. This means that, as well as the software you are using starting faster, you can switch between software applications much more quickly.
- Windows XP is more stable than previous versions of Windows making it less likely that the computer will crash completely.

2. Login Procedure

- Ignore any messages until the screen asking you to "Press Ctrl. Alt+ Delete to logon" is displayed.
- Press and hold the Control (CTRL) and ALT keys and then press Delete.
- The following dialog box will appear



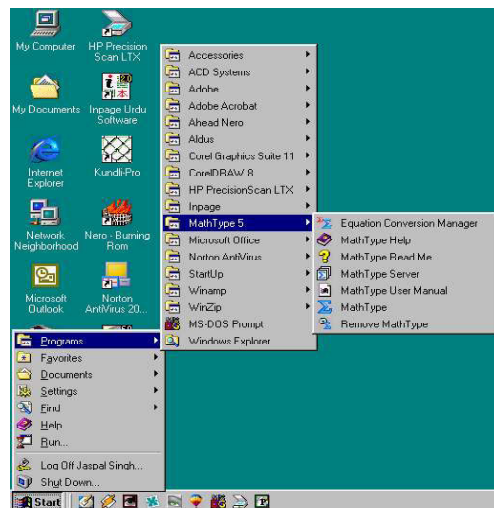
- Type in your User I.D and then use the mouse to click in box below and type in your Password.
- Click on the OK button

3. Using Software in Windows XP

Click on  and a pop up menu will appear.



- Select **All Programs**, by moving you mouse over it, this will open a sub menu containing the software installed onto the PC you are using.



4. Useful Menu Items

These features also appear when you click on the Windows XP Start button

- My Recent Documents :** This shows the latest documents you have used.
- My computer :** Gives access to the Hard Disks and other peripherals connected to your computer.
- Help and Support :** If you are not sure how to do something, the 'Help and Support' menu can be very useful. You can select a Help topic i.e., Windows Basics, Ask for Assistance or select the task you're trying to complete.
- Search :** Allows you to search for files/folders

5. My Computer

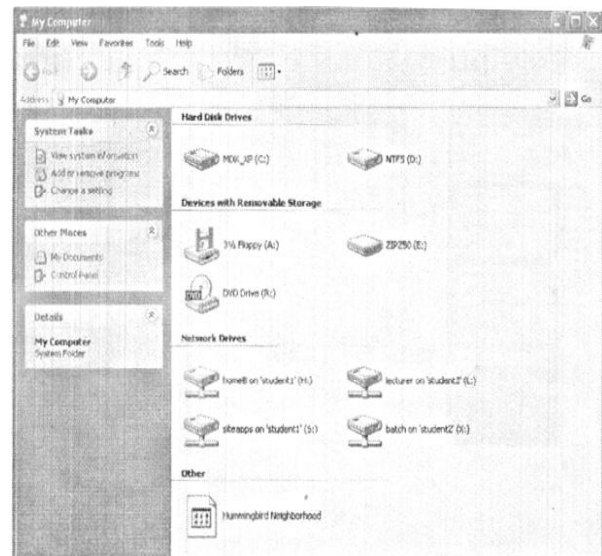
My Computer allows you to explore and maintain your computer by viewing programs, files and folders. It helps you organise files and folders so that they are easier to work with.

You can open My Computer by

Double clicking on the My Computer icon on the Desktop Clicking on the **Start** button then click on My Computer from the right pane.

Local drives

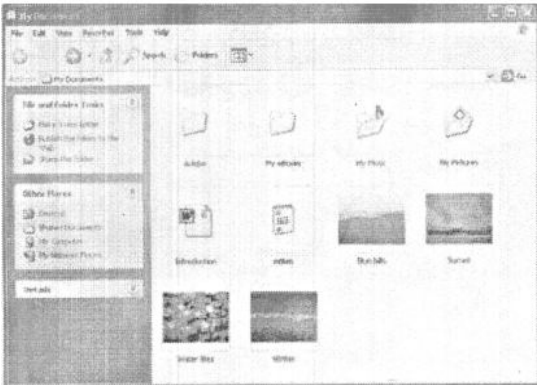
- A : Floppy
- C : Hard disk for XP Programmes
- D : Hard disk for saving files
- E : Zip drive
- R : CD ROM/ DVI



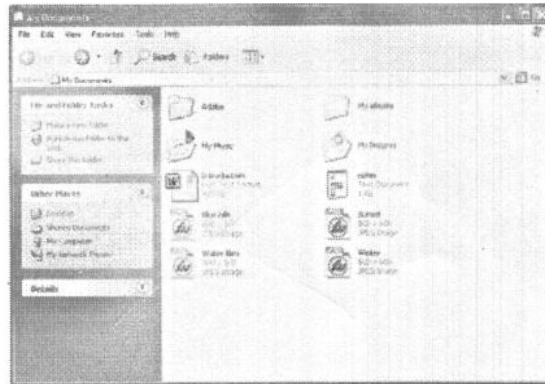
Double clicking on the drive folder will open that drive allowing the user to see a list of the files and folders saved.

You can change the way the files/folders are displayed in the View menu Click on View from the Menu Bar and choose either.

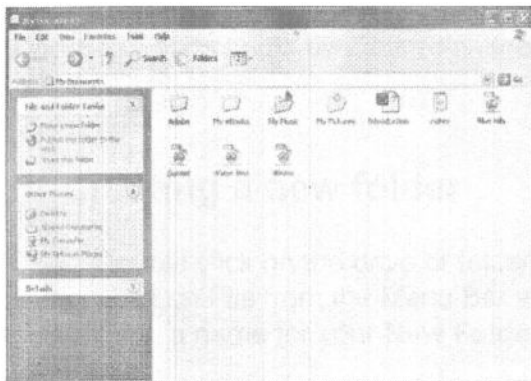
Thumbnails (display images)



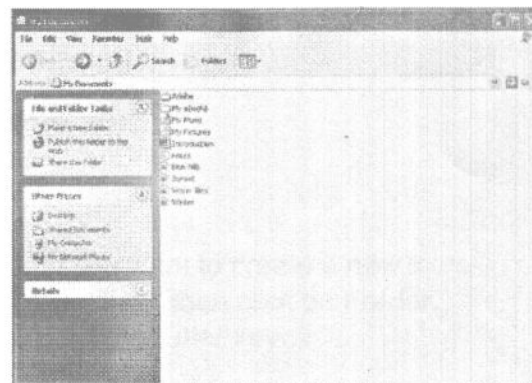
Tiles (large icons)



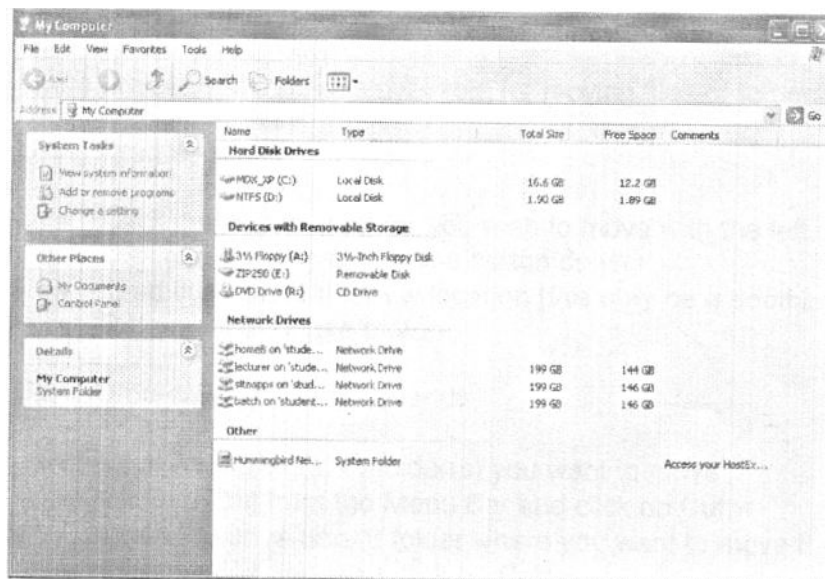
Icons



List



Details (a list containing file information inc. size)



Once you have selected the view type you can then arrange the items Click on View from the Menu Bar and choose Arrange Icons Name - sorts items alphabetically by name

Type - sorts items by type

Total Size- sorts items by size from smallest to largest

Modified- sorts items by date from oldest to most recent.

6. Creating a new folder

- Double click on the drive or folder in which you want to create a new folder.
- Click on File the Menu Bar and point to New, then click on Folder.
- Type a name for your New Folder and press the Enter key.

7. Rename a file or folder

- Click on the file or folder you want to rename.
- Click on File from the Menu Bar and then click Rename.
- Type in the new name and press the Enter key

8. Moving a file or folder

There are several methods available for moving files or folders.

Using Drag and Drop

- Click on the file or folder you wish to move with the left mouse button and continue holding the mouse button down.
- Drag the icon to the new location (this may be another folder or drive)
- Release the mouse button. Using Drop down menu commands
- Click on the file(s) or folder(s) you want to move
- Choose Edit from the Menu Bar and click on Cut.
- Open the drive and/or folder where you want to move the file.
- Click on Edit and then click on Paste.

9. Copying a file or folder

Using Drag and Drop

- Click on the file or folder you wish to move with the right mouse button and continue holding the mouse button down
- Drag the file/folder to the new
- Release the mouse button
- Select copy here from the menu that appears Using Drop down menu commands
- Click on the file(s) or folder(s) you want to move
- Choose Edit from the Menu Bar and click on Copy
- Open the drive and/or folder where you want to move the file
- Click on Edit and then click on Paste.

10. Selecting multiple files or folders

To select ALL files/folders in a location

Click on Edit from the Menu Bar and then click on Select All To select random files/folders within a location

Hold down the Control (CTRL) key and then click on each item you want to select

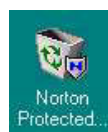
To select a group of files within a location

- Select the first file/folder in the group
- Press and Hold the Shift key
- Select the last file/folder in the group

11. Deleting a file or folder

Select the file or folder you would like to delete with the left mouse button and either; Click on File from the Menu Bar, and then click on Delete OR Press the Delete key on the keyboard

12. Recycle Bin



The Recycle Bin can be seen on the desktop and is used to store deleted files or folders. Once a file has been moved to the recycle bin it can either be; Removed permanently to create space on your Hard Disc

Or

Retrieved if you didn't mean to delete the file.

To empty the Recycle Bin

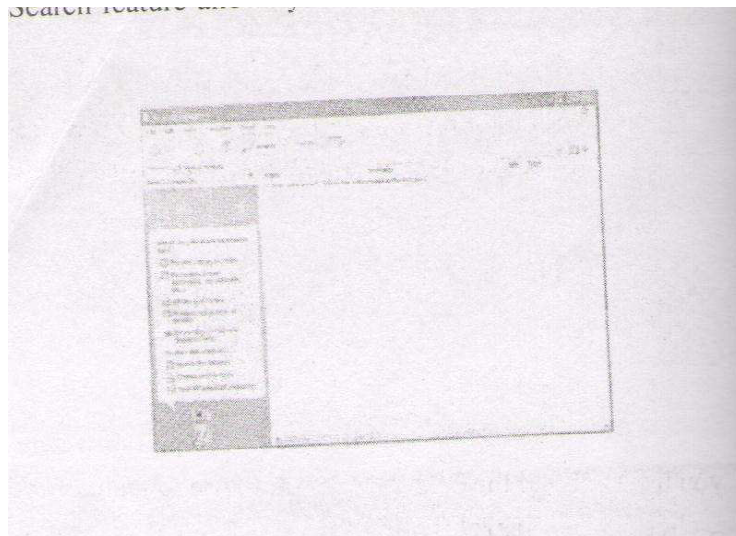
- Click on the Recycle Bin with the right mouse button.
- Select Empty Recycle Bin from the pop up menu
- Click Yes when asked "Are You Sure"

To recover a file from the Recycle Bin

- Double click on the Recycle Bin to view any files currently stored there.
- Right click on the File/Folder you want to recover
- Select Restore.

13. Search

The Search feature allows you to search your computer for Files/Folders.



To run this program

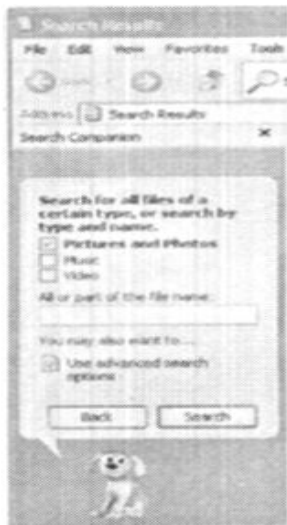
- Click on the Start button
- select Search on the main menu.

The Window is divided into 2 main sections

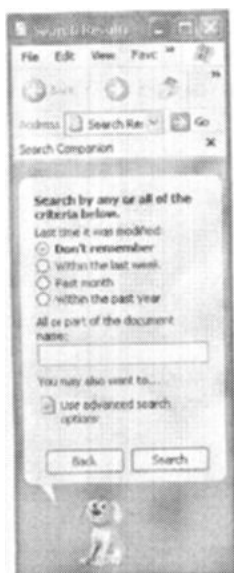
The Left hand section asks What Do You Want to Search For

Making a selection based on the type of file you are searching for

Pictures, Music or Video



Allows the user to run a search based on the type of file and/or the name of the file (whether part or the whole of the file name)



Documents

Allows you to search based on the last time that you accessed the file and/or the file name (whether part or the whole file name)

When you have entered the details you want to use to search for the file/folders you require, click on the Search button.

The search may take some time as the computer searches all drives to locate your file. Once the search has finished a list of files which match your search will appear. To open one of the displayed files simply double click on its file name.

14. Printing

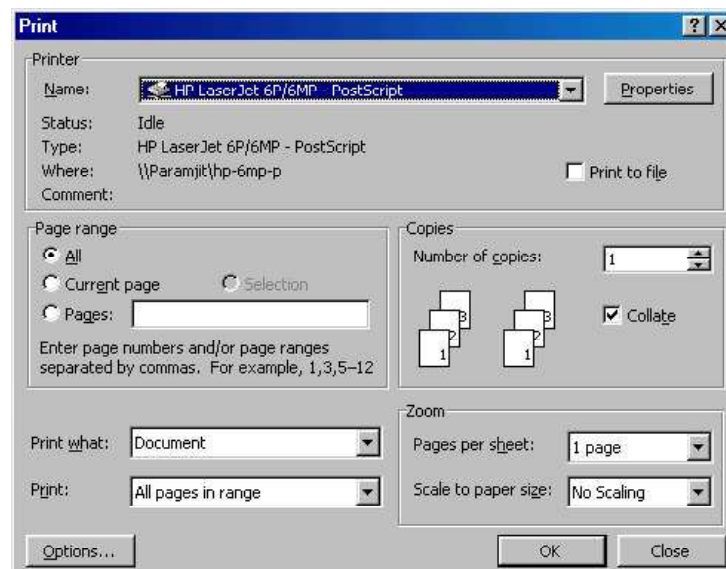
To ensure your work is printed you must

- Click on Start
- Select Middlesex Network Software
- Click on Select Printer and choose the printer you require;

To print your work from a software application i.e. Microsoft Word.

Select Print from the File menu within the application you are working.

This will cause the Print options dialog box to appear.



This box allows you to set;

- The name of the printer the work will be sent to (make sure this matches the printer you selected earlier)
- The pages of your document that will be printed
- The number of copies required

15.3 LET US SUM UP

So, in this lesson you learned that computer is a device which works with the help of various hardware devices commonly known as input and output devices. Along with hardware, computer requires an operating system also with is actually a system software. The commonly used system software is windows. The operating system makes a link among all the hardware devices to give meaningful results. Through operating system, various personalized settings can be made to computer.

15.4 GLOSSARY

- Operating System is a system software for Computer.
- Windows is commonly used operating system.
- Various input and output devices are required to make a computer set all their devices are known as hardware.

15.5 SELF-ASSESSMENT QUESTIONS

1. Which is an operating system.

2. What is a window. What are its various components.

3. Differentiate between system Software and application software with Suitable example.

15.6 LESSON END EXERCISE

1. DOS refers
2. Windows XP is a System Software (True/ False)
- 3- Windows XP is an operating System (True/ False)
- 4- RAM is required to run the programs (True/ False)

15.7 SUGGESTED READINGS

1. Argyrous, George. 1997. Statistics for Social Research. New York: Mc Millan Press Ltd.
2. Goods, W.J. & Hatt, P.K. 1981. Methods in Social Research. New York: Mc Graw Hill.
3. Gupta, S.C. 1981. Fundamentals of Statistics. Bombay: Himalayan Publishing House.
4. Gupta, S.P. 2004. Statistical Methods. New Delhi: Sultan Chand and Sons.

15.8 ANSWERS TO CHECK YOUR PROGRESS

VERY SHORT ANSWER TYPE

ANS1. An Operating System is a software that manages computer hardware and software resources and provides common services for computer programs.

ANS2. There are two main types of Operating Systems for personal computers:

Proprietary OS: Licensed OS like MS-Windows, Mac OS X, which require payment. Open-Source OS: Free OS with source code available, like Linux.

ANS3. Wildcards are special characters used to replace or represent unknown characters in search operations.

Examples: '*' (represents any number of characters), '?' (represents a single character).

SHORT ANSWER TYPE

ANS1: Advantages include:

- a. Professional support and updates from the company.
- b. User-friendly interfaces.
- c. Compatibility with many software applications.
- d. Regular security updates and patches.

ANS2: 1.docx (Microsoft Word documents)

2.. xls (Microsoft Excel spreadsheets)

3. .jpg (JPEG image files)

4. .mp3 (audio files)

5. .pdf (Portable Document Format)

ANS 3. Advantages:

- a. Multiple tasks can run at the same time, increasing productivity.
- b. Efficient use of CPU resources.

Disadvantages:

- a. Can cause slower performance if many tasks run simultaneously.
- b. May cause confusion to the user handling many tasks.

WORD PROCESSING (MS-WORD)**STRUCTURE**

- 16.0 Objectives
- 16.1 Introduction
- 16.2 All about MS word
- 16.3 Let us sum up
- 16.4 Glossary
- 16.5 Self-Assessment Questions
- 16.6 Lesson End Exercise
- 16.7 Suggested Readings
- 16.8 Answers to check your progress

16.0 OBJECTIVES

After completing this lesson, you shall be able to know: -

- How to create a MS-Word
- How to open or close a MS-Word file.
- Printing a MS-Word file.
- Performing other operations in MS-Word

16.1 INTRODUCTION

In last lesson, we discussed about operating system which is necessary to run a computer. But

merely running a computer do not serve any purpose. It demands various application software's for required purpose. M.S - Word is one of the application software's which is used to type a text document and perform some non-technical operations over it. Such as formatting, coloring, tabulating etc. In present lesson we are going to learn how to perform different operations in MS Word.

16.2 ALL ABOUT MS- WORD

In word processing, what you type on the keyboard is displayed on the screen. As well as a keyboard you also have a mouse. This is used by sliding it across a flat surface, which makes a small pointer move across the screen. The pointer changes shape as it moves. In a text area (a part of the screen in which you can type) it will be an I-shaped line. If you press (click) the left mouse button, a flashing cursor will move to the position you have chosen. The cursor shows where your text will appear when you start typing.

If you move the cursor out of the white text area it becomes a small arrow. By positioning this arrow over certain areas of the screen you can activate commands by clicking the left mouse button. These commands allow you to edit, format, save and print your text.

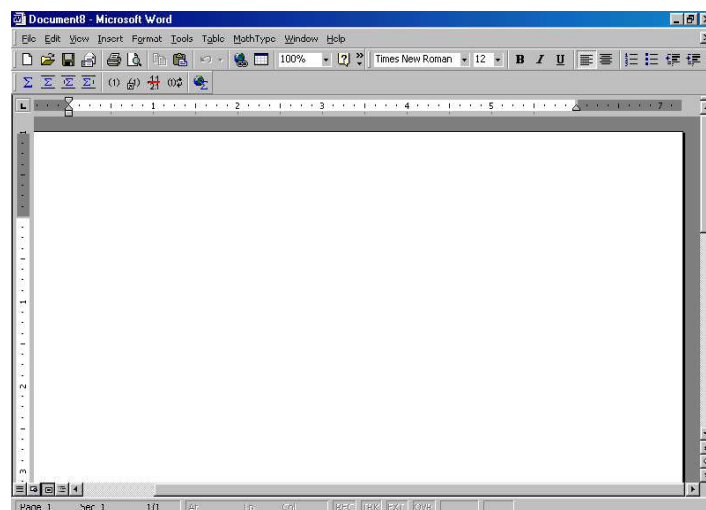
Starting Word

Click on MS Word icon available in start window programmes list.

If you can't see the Word XP icon, click on the **Microsoft Office Suite folder**. To start Word double, click on the Word 2000 icon. You may be prompted for user information: click Cancel to clear this. Finally, the Help system may be active (the small window with an animated paperclip): click the close box on this to shut it down.

- Start Word and clear any welcome message, user requests and Help windows.

The Main Document Screen



To close the New Document box at the right of the screen, click on X. To re-open select File, New.

At the top of the screen is the Title Bar, which tells you the name of the program you are using and the name of the document currently open on the screen. This document is called Document 1 as it hasn't yet been given a name.

On the right-hand side of the Title Bar are three buttons. The left-hand button (the horizontal line) is the minimise button. Clicking on this button will cause Word to be minimised to an icon on the desk top.

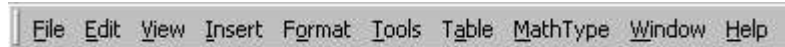
The button in the middle can take two different forms depending on the size of the Word window. A single large square is the Maximise button and causes the Word window to fill the full screen. Once the Window is maximised the button changes to two smaller overlapping squares. This is the restore button and clicking it causes the window to return to its previous size.

The right-hand button is the close button: clicking this will exit from Word.

Note: these buttons control the Word program; the second set of buttons underneath apply only to the current document.

Menus

Below the Title Bar is the Menu Bar.



This holds all the pull-down menus for Word. To see the menus, click with the left mouse button on the menu required. When a pull-down menu is displayed it will have several options which can be selected by clicking on them with the left mouse button. Some of these options have symbols to the left or right of them:

.... this option will display a dialog box.

- this option has a further sub-menu.
- this option is currently active (clicking it again will turn it off).

Ctrl+S A key combination to the right of a menu option indicates a keyboard shortcut to perform the same operation.

this option will display a menu which has more options below.

To pull open a menu using the keyboard: hold down the [Alt] key and press the underlined letter of the menu option, e.g. pressing [Alt+E] will pull down the Edit menu.

These menus have become standard for Windows products. All Windows applications (spreadsheets, databases, word processors, etc.) will have similar menus.

To select an option from a menu, point to the option and click. To close a menu without choosing a command click on a blank area of the screen. Any options displayed in light grey are currently unavailable, but may become available at a later time. For example, you will not be able to use Copy from the Edit menu unless you have already selected the text you want to copy.

Toolbars

Immediately below the menu bar are one or more rows of toolbars. The commands you need to perform most of the basic operations in Word can be found in these toolbars. By clicking on a toolbar button you can activate many different commands e.g. save, print. Some options on the toolbars are in the form of pull-down lists (at the end of the toolbar is an icon ; click here to see more toolbar buttons).



The standard and formatting toolbar looks like this:

- Move the mouse pointer over each of the buttons on the toolbar (but don't click on them).

A short description (called a Tool Tip) pops up to tell you what each button will do. All the commands found on the toolbars can also be found in the menus above.

A toolbar button which looks "pushed in" means that option is currently active: clicking it again will turn it off.

 If you need any information about what a particular command will do, or want to know more about performing certain tasks, then Word has a good on-line help system. To access this either select a command from the Help menu, or click on the Office Assistant icon.

The Ruler

Below the toolbars is the ruler.



This displays current margins, tabs and indent settings (explained below).

The Status Bar

At the bottom of the screen is the status bar.



This displays information about your document, including: Page number, Section number, the total number of pages from the beginning of the document followed by the total number of pages in the whole document, the position of the cursor from the top of the page, the line number and column number.

Typing and Editing Text Word Wrap

Word automatically wraps the lines of text to fit between the margins of the pages as you type, starting a new line when needed. If you add or delete text, change the margins, or change the format of your text, Word automatically adjusts the position of the text for you. Unlike using a typewriter, you only need to press RETURN (the large key marked with a () at the end of each paragraph).

Correcting Mistakes

If you make a mistake when typing you can delete the character to the left of your cursor by pressing BACKSPACE (the wide key marked with a just above the Return key). You can delete the character to the right of your cursor by pressing DELETE (to the right of the Return key).

Capital Letters

For single capital letters hold down SHIFT [is] key and press the letter. For all capital letters press the CAPS LOCK key.

Insert/Overtyping

By default, Word inserts characters, you type by moving existing text to the right. You can change this so that new text replaces old text, character by character (Overtyping). To turn Overtyping on press the INSERT key (to the right of the Backspace key). To turn it off press INSERT again.



Undo

This button on the toolbar allows you to undo a particular command or action. If you don't like the results of a command or accidentally delete some text, choose Undo as your next action. This command is also available in the Edit menu.

Select Edit, Undo... (where... describes the action to undo) from the menu or

press the Undo button from the standard toolbar (use the Tool Tip to find it!). Some actions, such as saving

a file, can't be undone. In this case Undo changes to Can't Undo in the menu and appears grey in the pull-down list.

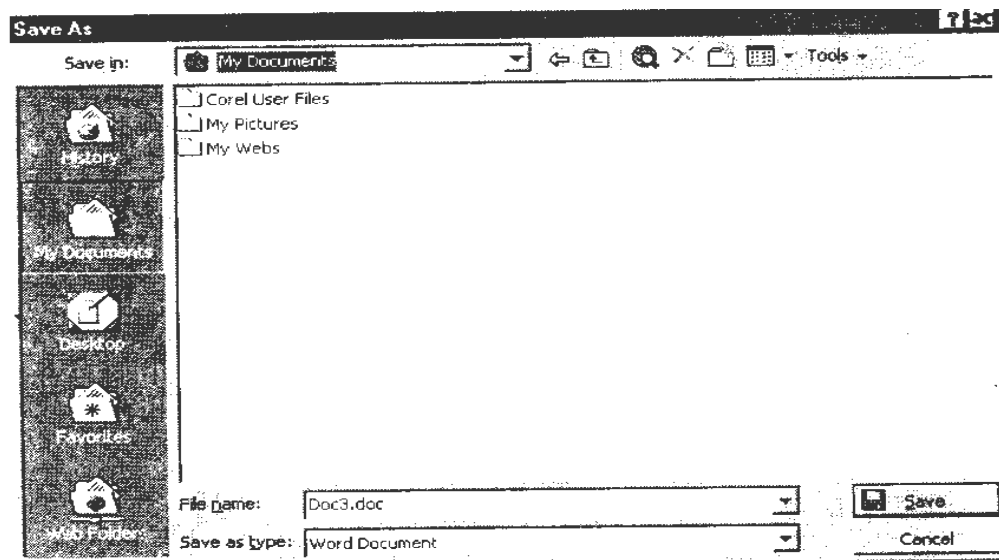
Redo

The Redo button works in the same way. Select Edit, Repeat... (If you cannot repeat the last action then the Repeat command changes to Can't Repeat).

Saving Your Work

To save a new document for the first time

Select Save from the File menu, or click on the save icon  on the Toolbar. You will see a dialog box similar to this (the folder and file names will be different) :



You can now give your document a file name. Word will suggest a name in the File Name: box. You can accept this, edit it, or replace it with a name of your choice. Then click Save. Note that your file name replaces "Document 1" (or the current default filename) in the title bar.

To save an existing document (after further edits)

Select Save from the File menu, or click on the save icon on the Toolbar.

If you choose SAVE AS from the File menu you will be prompted for a filename every time you save a file. This allows you to change the filename at a later stage (the original file will still be there).

FILE, CLOSE (i.e. pulling down the File menu then selecting the Close option) closes the current document, asking you if you want to save any changes if any have been made since you last saved the file.

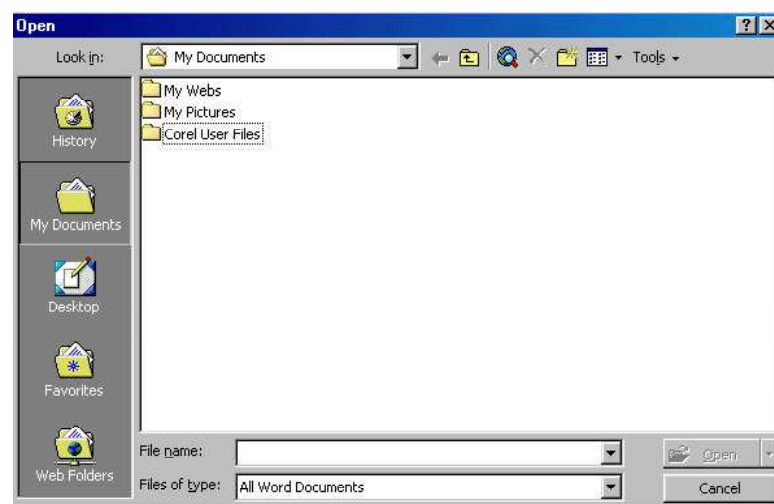
Filename

File names can be up to 255 characters long but cannot include any of the following characters: forward slash (/), backslash (\), greater than sign (>), less than sign (<), asterisk (*), period (.), question mark (?), quotation mark ("), pipe symbol (|), colon (:), or semicolon (;) When saving a file Word will automatically give each filename the extension. DOC to enable Word to recognise its own files.

Opening Files

To open an existing document you can either select Open... from the File menu or click on the open icon  on the toolbar.

Both actions will present you with a dialog box similar to this (again the folder and filenames will be different).



To open a document simply click on it and then click Open.

Clicking on the NEW file button  on the standard toolbar opens a new document window with a blank document.

FILE, NEW displays a dialog box with a choice of templates for your new document

: selecting Blank Document on the General tab will open a standard document.

Formatting Text

Once your document has been typed, it can be enhanced with effects such as bold, underline, different fonts and centering.

For new text, we can switch these options on, type the text, then switch the option off again. However, if we want to alter text we have already typed, we must SELECT it first.

Selecting Text

To select:

A range of words

Do this:

DRAG the mouse across the area to be selected. (Move the mouse pointer to the start of the text you want to select; click and hold down the left mouse button; move the pointer to the end of the text you want then release the button-the text will be highlighted as you drag the pointer across it.) Alternatively: click where you want the selection to begin. Hold down the SHIFT key. Click where you want the selection to end.

Word

Double-click on the word.

Sentence

Hold down CTRL and click anywhere in the sentence. Paragraph Triple-click anywhere in the paragraph or double-click in the left margin.

Line of text

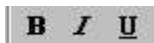
Click in the left-hand margin, next to the relevant line. Whole

document

Triple click or CTRL and click in the left margin.

Selected text is reversed (white on black). You can now alter the appearance of the selected text by choosing one or more of the commands from the Format menu. Or you can click on one of the format icons on the toolbar. For example, by clicking on the Bold icon on the toolbar the text will appear in BOLD.

Bold Underline and Italics



Select the text you want to format then click the Bold, Italic, or Underline button.

Text Alignment

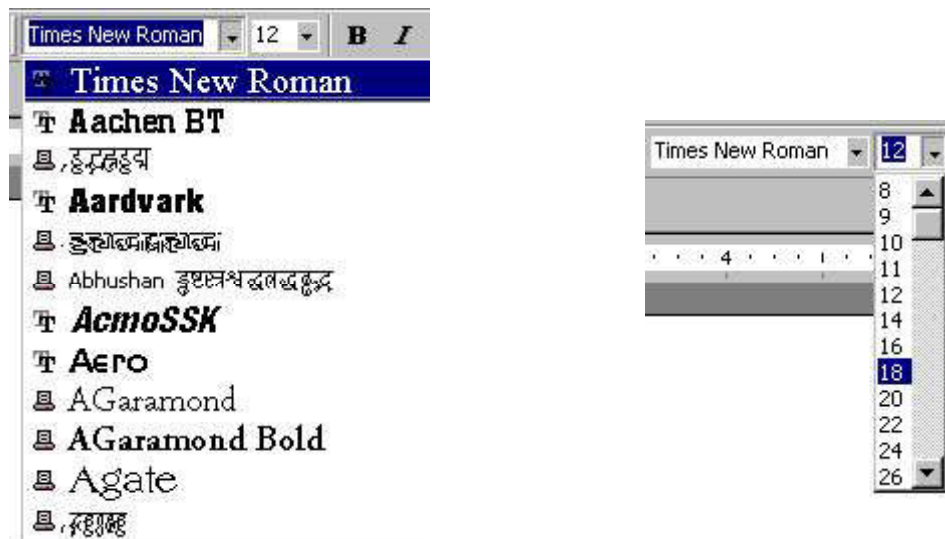


The same steps are repeated when aligning text on the screen. The two icons displayed shows text being left aligned (Align Left), and text being centred between the margins (Center). To make the text flush to the right margin (Align Right), and aligning to both margins (Justify), use the pull-down menu for more options.

Fonts and Sizes

The different typefaces used in word-processing documents are known as fonts. Fonts are usually measured in points. A point is 1/72 of an inch. Standard text is usually Times New Roman in 10pt size. Font types and sizes can be selected before typing text or can be added retrospectively to specific sections of text (by selecting the text first).

To select font types and sizes from the formatting toolbar, click on the arrow next to the current font name and a list of available fonts will be displayed. The pull-down list to the right lets you choose the font size.



Format Menu

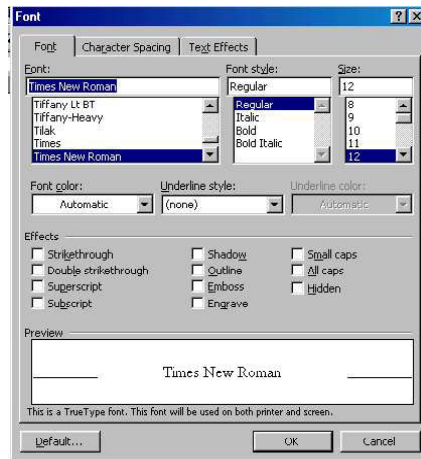
Select FORMAT, FONT.... from the menu bar.

You will be shown a dialog box where you can select from a list of fonts as well as being able to specify its style (bold, italic) and its point size, the colour of the text, plus some other effects.

Paragraph Numbering of Bulleting

Word can automatically number paragraphs or put bullet points on them. Select the paragraphs to be included then click the Bullets or Numbering button on the Toolbar.

By default, the number or bullet point will go to the left of the first line of each paragraph selected. To change the style of the bullets or numbers, select Bullets and Numbering.... from the Format menu



- Open the file.
- Using the Toolbar, add bullets to the quotations.
- Using the Format menu, change the bullets from circles to squares.
- Save the file.

Line Spacing

Standard Word documents are in single line spacing. Line spacing can be changed for individual paragraphs by selecting the paragraphs first.

Move the cursor to anywhere within the paragraph you want to change; if you want to change more than one paragraph, select them with the mouse.

Select Format Paragraph... Pull down the list box under line spacing and select one of the following options:

Single Single line spacing.

1.5 Lines One and a half line spacing. Double line spacing.

Moving Blocks of Text

Copy and Paste



This is a way of moving sections of text or copying things into your document. It is available in all Windows applications, and you can use it to copy things between different applications. There are two commands involved: Copy and Paste. Each can be selected from the EDIT menu or from buttons on the standard toolbar.

When you copy text, it is put into an area known as clipboard, which is available to all Windows programs. Copy simply puts a copy into the clipboard.

The Paste command copies whatever is in the clipboard into the document at the current cursor position. If text is selected when you paste from the clipboard, its contents replace the selected text. Text remains in the clipboard until the next Copy, so you can paste the contents of the clipboard several times.

You can paste between Windows applications. For example, you can insert a picture into a document by copying it from Paintbrush and pasting it into Word.

Cut

The Cut command is another way of moving sections of text. Unlike the Copy command, Cut deletes the text from the document and puts the text into the clipboard. You can then use the Paste command to insert the information elsewhere. The Cut icon can be found in the pull-down menu located next to the help icon on the toolbar.

Note: the DELETE or DEL key on the keyboard deletes text without copying it to the Clipboard.

Drag and Drop

Select a block of text to be moved and place the mouse pointer over the text.

Hold down the left mouse button and DRAG the pointer to the new location for the text. A grey vertical line will indicate where the text will appear in the document.

Release the left mouse button and DROP the text into its new location.

Dragging and dropping is a quick alternative to Cut and Paste but can be little tricky to control if you are new to using a mouse.

Spell Checking

Word has a built-in dictionary which can be used to check the spelling in your work. Word also checks for repeated words (the the), odd capitalisation (tHE), and proper nouns (London).

To spell check a document select the word or section of the document you want to check. If nothing is selected, Word checks the document from where the cursor is, and when it reaches the end will ask if you want to continue from the beginning, until it has checked the whole document.

Select Spelling

Word looks for words in your document that don't match those in its dictionary. It highlights those words and displays the Spelling dialog box.

To correct the misspelled word, either type it in the text area or select the correct word from the Suggestions list.

Press the Change button to replace the misspelled word with the correct one, or press Change All to change all occurrences of the misspelled word in your document.

If the word highlighted is spelt the way you want, choose Ignore to leave the words as it is, or choose Ignore All to ignore all occurrences of this word in your document.

If Word highlights a word, it does not recognise and you know that it is spelt correctly, you can add the word to the dictionary by selecting Add. The word is added to the dictionary displayed in the Add box.

Word continues checking your document. Choose Cancel at any time to stop the spell check.

Thesaurus

When you're not sure of the meaning of a word, or when you think you're offering alternatives (synonyms). To use it:

Select a word in your document then select TOOLS, LANGUAGE, THESAURUS.... from the menu bar.

The Thesaurus dialog box appears, showing the meaning of the word on the left and a list of synonyms on the right.

To see a list of other related words click on any related words in the meaning box.

To see synonyms for one of the alternative words, select it and click the lookup button.

To choose an alternative, highlight it and then click REPLACE.

Word Count

Select Tools/Word Count.

Headers and footers

Headers and footers are pieces of text that appear at the top or bottom of each page. To create a header or a footer choose Header and Footer from the View menu. A text area at the head of the page will be highlighted and you will see the following toolbar :



You can switch between the header and the footer by clicking on this icon . Text can be typed into the header or footer text area. You can also use the buttons which will automatically include page numbers and the time/date. (Page numbering is covered in more detail later.) When you have finished click on Close. The header and footer areas will disappear but you can see them easily by changing the View of your document.

Changing the View in Word

You can View a Word document in several ways: you can switch between them using the View menu or the buttons in the status bar at the bottom of the screen.



Normal View

Normal View is the default view in Word. Using Normal View it is quick and easy to edit and format text. Your document will look similar to how it will be printed but you will not see the headers and footers.

CHECK YOUR PROGRESS-I

A. Fill in the blanks.

1. Microsoft Word is a word processing _____
2. The _____ toolbar has shortcut to the most commonly used commands.
3. _____ function is performed when you want to save a copy of your file with a different name or at a different location.
4. Save your file at regular _____.
5. Ribbon is a group of _____.

B. Write true (T) or false (F) for each statement.

1. Use Enter key to move the cursor to the next line.
2. Status bar is the top most bar of any Microsoft Window.
3. Scroll bar is used to scroll up and down on a page.
4. Word processor is a software.
5. You can type your text on the title bar.

Web Layout View

Web View Layout shows how your document will look in a Web browser. MSWord saves a copy of your document and then opens it in your default browser.

Print Layout View

Print Layout View shows the layout of each page in your document exactly as it will look when printed. Headers, footers and footnotes are displayed in the correct place on the page.

Outline View

Outline View collapses the document to headings only to show the general structure of a document. This can be useful for very long documents.

Print Preview

Whichever View you are using, you can always see how your document will look when it is printed by selecting Print Preview from the File menu or clicking the . You cannot edit your document in Print Preview. Click the Close button in the Print Preview toolbar to return to your document.

Zoom

The magnification of the document in a View (or in Print Preview) can be changed by zooming in or out of the screen :

Select VIEW, Zoom... from the menu, or use the Zoom pull-down menu list from the standard toolbar and select a percentage or automatic size from the list provided. Zoom does not change the font size in your document.

Pages, Page Breaks and Pagination

As you type in Word, when you reach the end of a page you will automatically be taken to the top of the next page. The status bar will tell you which page you are on and, in Normal View, the page break will be shown as a horizontal dotted line across the page.

If you want to start a new page before you reach the end of the current one, you can force a page break. To create a page break:

Select INSERT, BREAK....from the menu.

Select PAGE BREAK from the Break dialog box and press OK. or

Press CTRL + RETURN.

To remove a page break from a document, move the cursor onto the dotted line and press DELETE.

Sections

When you start a new document the same formatting setting (margins, page numbering, etc.) apply to the whole document. To apply different settings to different parts of the document you must create a section for each part. Each section can then have its own settings.

The different types of sections are:

CONTINUOUS: Starts the new section wherever the cursor is.

NEW PAGE : Starts the new section at the top of the next page (Word automatically inserts a PAGE BREAK).

ODD/EVEN : This is used particularly for double sided printing where page PAGES: numbers, margins and headers/footers are different on odd and even pages.

To create a new section, move the cursor to where the section should start. Select INSERT, BREAK....from the menu.

Select the relevant SECTION BREAK.

You will only see section break lines in Normal view (not in Page Layout view). Make sure you are in Normal view before continuing.

To remove a section break, move the cursor to the section break line and press DELETE.

Page Numbering

Word can automatically number your pages. Page numbers appear in the header or footer area of your document.

Inserting page numbers into headers and footers

Move to page one and select View, Header and Footer Select the Header or Footer area. If the Header or Footer already has text in it, move the cursor to where the page number should appear.

Click on Insert Page Number . The page number appears in the header or footer. You can select this page number and format it as you would any other text. You can also add any additional text you want in the header or footer.

To delete the page number, select it then press the DELETE key. Click on CLOSE to leave the Header and Footer area.

Page Numbers do not display in Normal or Outline View. To see page numbers switch to either Print Preview or Page Layout View.

Adding page numbers only

If your header or footer will contain only a page number and no other text, you can use this shortcut to insert page numbers:

Put the cursor on the first page of your document. Select Insert, Page Numbers

In the position box select either Top of the Page or Bottom of the Page. In the alignment box select the desired position.

Click OK.

This header/footer can be edited in the usual way.

Margins

Margins are the spaces between the edge of your text and edge of the paper to be printed on. Default margins (top, bottom, left and right) are applied when you open a new Word document. You can change the margins for the whole document or for individual sections.

Select File, Page Setup....

Click on the Margins tab in the Page Setup dialog box.

Type in the required margin measurements (or use the scroll tabs to alter the values up and down).

Choose the area the margin changes APPLY TO :

This Section : applies the new margins to the current section only.

This Point Forward: creates a section break and applies the new margins. (Note : click on the Layout tab and make sure that the 'Section Start' setting is of the type required.)

Whole Document: applies the new margins to the whole document.

Click OK. (Note : you will have to use Print Layout View to see the effect of margin changes).

Multiple Documents

In Word you can work with several documents at the same time and copy or move text between them. Each new document you create or open has its own document window. As you open successive

documents, they appear as the topmost window, hiding any previous documents behind them. (Note that each document has its own set of Minimise/Restore/Close buttons below the set for the Word program.) To switch between documents :

Put down the Window menu: a list of open documents appears in the bottom section of the menu. Select the name of the file you want to switch to.

Indentation

Indentation is the position of the text relative to the margins.

←→ **Left Indent Only** : this paragraph is indented from the left margin and not indented from the right.

←→ **Left and Right Indent** : this paragraph on the other ←→ hand is indented from both the left and right margins.

←→ **First Line Indent** : this paragraph has no indentation from the right margin but has a first line indent from the left margin.

Hanging Indent : this paragraph has no first line indent but all subsequent ←→ lines are indented from the left margin. This is a hanging indent ←→ and is useful for any item whose first line needs to begin at the left margin.

To indent a paragraph, place the cursor in the paragraph you want to indent and then drag the indent markers on the Ruler to the required position. Any changes only affect the current (or selected) paragraphs.

The Increase/Decrease Indent buttons on the formatting toolbar are a quick way of changing the left indent. The right hand button increases the indent by one tab stop, the left hand decreases the indent.



Indents can also be changed using the Format menu. Put the cursor in the paragraph you want to indent and choose Format, Paragraph and change Left and Right indent measurements. To set first line only or hanging indents use the Special box, then set the amount in the By box

Tab Settings

Tabs are another way to indent text. Using Tab settings is the method to use if you want to align text in columns or you only want the first line moved from the left margin. Tabs cannot be used to indent whole paragraphs.

Do not use the space bar to align text in columns: the columns will almost never appear properly aligned when you print your document.

There are 4 types of tab stop:  Left justified;  Right justified;  Centre justified and  Decimal justified. Tabs are shown on the ruler. Different symbols indicate different types of tab.

[Left]

Williams

Evans

Smith

Richard

[centre]

Williams

Evans

Smith

Richard

[right]

Williams

Evans

Smith

Richard

[deci.mal]

12.433

4897.213

45.64

1200.09

Tabs can also have leaders (dots, dashes or solid lines between the margin and the tab position):

.....	12
.....	24
.....	36

Setting Tabs Using the Ruler

To ADD a tab: Click the tab button at the left-hand end of the Ruler until the symbol for the tab you want appears.

Click on the Ruler at the position you want the tab.

To MOVE a tab: DRAG it to its new position.

To DELETE a tab: click on the tab and DRAG it off the Ruler.

Setting tabs with the Format menu

Use the Format menu if you want precise control over tab positions or if you want to set tabs with leaders. Put the cursor where you want the formatting to start or select the existing text you want to add tabs stops to.

Select Format, Tabs...

Type the position you want the tab stops to be placed at in the Tab Stop Position box (zero is the left margin).

Under Alignment, select Left, Centre, Right or Decimal. Under Leader, select the option you want (if any).

Click the SET button.

Choose OK.

The tabs you have set will appear on the ruler.

Using Tabs

To use a tab press the TAB key on the keyboard and the cursor will jump to the next tab stop on the ruler.

To return text to its un-tabbed position (without deleting the tab position from the Ruler) press the BACKSPACE key.

Printing

Select FILE, PRINT from the menu The Print Dialog Box appears.

Options allow you to: print all of the document, the current page, a range of pages, or selected text; print just odd even pages; print multiple copies. Click OK when ready.

Note: if you want to do a standard print of all pages using the default settings click the Print button on the toolbar. This will send your document directly to the printer without opening the Print Dialog Box first.

Remember that you can see how your document will look when it is printed by selecting Print Preview from the File menu. You can print your document using the Print button or the File menu from the Print Preview screen as well as from the main document screen.

16.3 LET US SUM UP

After studying this lesson, it is clear that MS-word can be used to create a test document with different fonts and Styles. It offers many functions including printing mail merging, inserting tables, images, paint files etc. It is user friendly application software.

16.4 GLOSSARY

- MS -word is one of application offered by Microsoft office.
- In MS- word almost every utility is available if the job is typing the document with multimedia options too.

16.5 SELF ASSESSMENT QUESTIONS

1. What is MS- Word and What are various function that can be performed in this application software.

2. What is mail-merge and how it works.

3. What are clipboard operations.

16.6 LESSON END EXERCISE

1. Short cut key for opening a file is _____
2. Short cut key for closing a file is _____
3. Short cut key to save a word file is _____

16.7 SUGGESTED READINGS

1. Argyrous, George. 1997. Statistics for Social Research. New York: Mc Millan Press Ltd.
2. Goods, W.J. & Hatt, P.K. 1981. Methods in Social Research. New York: Mc Graw Hill.
3. Gupta, S.C. 1981. Fundamentals of Statistics. Bombay: Himalayan Publishing House.
4. Gupta, S.P. 2004. Statistical Methods. New Delhi: Sultan Chand and Sons.

16.8 ANSWERS TO CHECK YOUR PROGRESS

A. Fill in the blanks

1. Microsoft Word is a word processing software.
2. The Quick Access toolbar has shortcut to the most commonly used commands.
3. Save As function is performed when you want to save a copy of your file with a different name or at a different location.
4. Save your file at regular intervals.
5. Ribbon is a group of tabs.

B. Write true (T) or false (F) for each statement

1. True (T) - Use Enter key to move the cursor to the next line.
2. False (F) - Status bar is the bottom bar of any Microsoft window, not the top.
3. True (T) - Scroll bar is used to scroll up and down on a page.
4. True (T) - Word processor is a software.
5. False (F) - You cannot type your text on the title bar; text is typed in the document area.

INTRODUCTION TO MICROSOFT EXCEL

STRUCTURE

- 17.0 Objectives
- 17.1 Introduction
- 17.2 All about MS- Excel
- 17.3 Let us sum up
- 17.4 Glossary
- 17.5 Self-Assessment Questions
- 17.6 Lesson End Exercise
- 17.7 Suggested Readings
- 17.8 Answers to check your progress

17.0 OBJECTIVES

After successful completion of this lesson, you should be able to: -

- Create, open or close excel file
- Print excel file
- Perform clipboard operations
- Create formula.
- Use charts, graphs.
- Database operations.

17.1 INTRODUCTION

Microsoft *Excel*[®] is a piece of software which allows you to create professional spreadsheets and charts. It performs numerous functions and formulas to assist you in your projects.

17.2 ALL ABOUT MS-EXCEL

⇒Getting started

1. Click on the **Start** button towards the bottom left of the screen.  Windows operating system.
2. Click on each of the following: **Programs> Office XP[®]> Microsoft Excel.**
3. Within a few moments, Microsoft *Power Point*[®] will open.



⇒Starting a new presentation

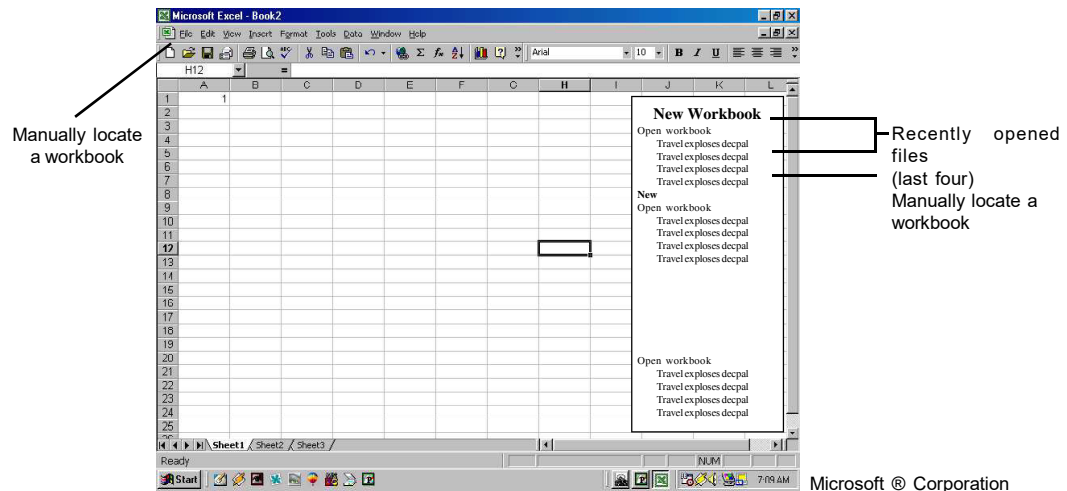
1. When you first open Excel *t*[®], a new workbook will automatically appear.
2. If at any time you wish to start a new file, click the **New** button from the toolbar.

⇒Opening an existing Workbook

There are several ways to open an existing presentation.

1. From the 'New Workbook' menu to the right, select a file name (the four most recently opened files are listed in order).

- From the 'New Workbook' menu to the right, select the 'More Work books' option to manually locate your file within a specific folder. Double click onto the file once located to open it.
- From the Menu Bar, select **FILE** then **OPEN** to manually locate your file from a folder. Double click onto it once located.



Opening an existnig workbook

Some important terms used in excel

Work book: An Excel file is called a workbook.

Worksheet: The page you work on which is made up of grid cells.

Cell/Selected Cell: Where you type data/formulae into. Cells are arranged in numbered rows and lettered columns. You have to select a cell to add data to it. data/formulae can also be typed into the formula bar.

Column/Row Heading: Use the column/row headings to identify a cells position on the worksheet i.e., A12, B6. Click heading buttons to select a whole column/ row of cells.

Name Box: Holds a cell's selected reference number—it's position on the worksheet.

Work tab Sheet: click the tabs to move between worksheets.

Scroll Bars: Use them to display hidden parts of the worksheet.

⇒Entering data into the worksheet

- Click on a cell. A thick border indicates that the cell is selected.
- Start typing. Note that data appears in the cell and in the formula bar.

⇒Editing data in a cell

1. Double click the cell and make any necessary changes.
2. Press **ENTER** to accept the changes, or **ESCAPE** to cancel.

If you make a mistake or change your mind, you can normally undo your last action by choosing **UNDO** from the **EDIT** menu.

⇒Deleting cell contents

1. Select the cell(s).
2. Press the **DELETE** key.

⇒Selecting a cell

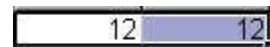
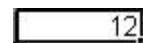
1. To select a single cell: Click onto it.
2. To select range of cells: Place the pointer on the first cell; hold down the mouse button and drag to the last cell that you want to select; release the mouse button.
3. To select a row/column: Click the appropriate row/column heading button.

⇒Copying and pasting cells

1. Select the cells you want to copy.
2. From the **Edit** menu **Copy**.
3. Click in the destination cell.
4. From the **Edit** menu, choose **Paste...** A moving border around the selected area, means that the cells may be pasted again. Press the **Escape** key to cancel the border.

Alternatively—

1. Select the cell that you want to copy.
2. Place the pointer over the black square at the bottom right of the cell; when the pointer changes to a black cross.
3. Hold down the mouse button and drag over the cell(s) that you want to copy to.
4. Release the mouse button.
5. Click the cell to cancel the selection.



Note : This method will only copy to neighboring cells in the same row or column.

⇒ **Cutting and moving cells**

1. Select the cell(s) and place the pointer over the cell's border.
2. When the pointer changes to an arrow, hold down the mouse button and drag the cell to a new location.



Inserting cells between existing cells

1. Copy/cut your cell(s) using the **Edit** menu.
2. Click on the cell that you want to move data to.
3. From the **Insert** menu, choose **Cut** or **Copied Cell**
4. Choose a direction to move existing cells; click **OK**

⇒ **Inserting new cells, rows, or columns**

1. Decide where you want to insert a cell/row/column, and click a cell.
2. From the **Insert** menu, choose **Cells... Rows** or **Columns**.

This will have the following effect on existing cells;

a new cell moves cells down or right, depending on your choice.

a new column moves column to the right.

a new row moves rows down.

⇒ **Deleting cells, rows, or columns**

1. Select the cell(s) delete.
2. From the **Edit** menu, choose **Delete...**
3. Decide what you want to delete, and click **OK**.

This effects existing cells in the opposite way to inserting new cells - see above.

⇒ **Changing column width**

* *To adjust row height, follow the same procedures but use the row heading buttons.*

If a column isn't wide enough to display numbers in a cell, '#' symbols are shown. Separating column headings, situated to the right of column letters, are border lines. When the pointer is over one of these lines, it changes to a black cross.

- double click the line to match column width to the longest data entry.
- drag the line for manual control over column width.

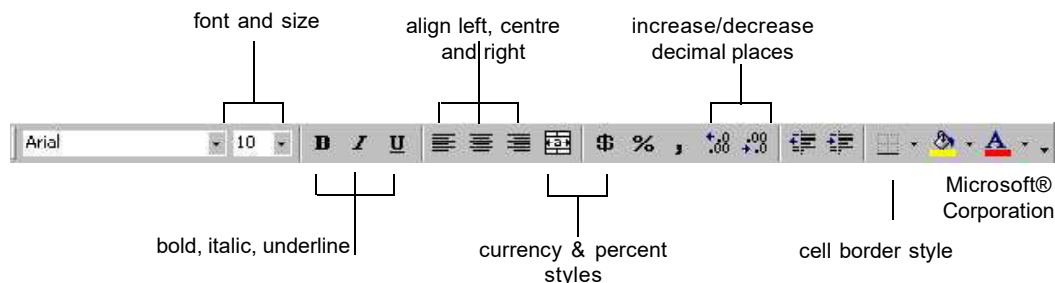
Changing the width of several columns at once

1. Decide how many columns you want to adjust, and place the pointer over the column heading of the first of these columns.
2. Hold down the mouse button and drag over the headings of the other columns.
3. Release the mouse button.
4. Drag the border line of the last column in the selection to adjust column width.
5. Click on the worksheet to cancel the selection.

⇒Formatting text, numbers and cells

General is the standard data format; it aligns numbers to the right, text to the left. Cells containing text and numbers are aligned left.

Commonly used formatting options relating to the appearance and alignment of text, can be chosen from the toolbar at the top of the screen.



To change a cells number format, use the **Cell** option in the **Format** menu.

1. Select the cells you want to change.
2. From the **Format** menu, choose **Cells...**
3. Click the **Number** tab. 4.

Applying percentages

There are two ways of applying a percentage:

- type in the actual percentage i.e. 54%
- apply the percent style button from the toolbar

The percent style can be applied before or after a number has been typed into the cell, but the number must be expressed as a decimal, as the *percent style* multiply numbers by 100. E.g. 0.54 would give 54%; 54 would give 5400%

⇒Calculations

To perform calculations, you have to type in a mathematical formula. A formula must start with = followed by a combination of cell names, numbers and operators, Example formulae:

<i>addition</i>	= A4 + C12	<i>multiplication</i>	=C1*D1	<i>subtraction</i>
	=D3-H3	<i>exponentiation</i>	=E3^2	<i>division</i>
	=B12/D18	<i>percentage</i>	=A3*20%	

1. Click on a blank cell to hold the result of the calculation.
2. Click in the formula bar or cell, type =, and then the formula.
3. Press **Enter** to perform the calculation.

To refer to a range of cells, there is no need to type out each individual reference number from A1 to A15 for example, just type A1: A15 - the colons mean 'to'.

Like data, you can copy and paste formulae. *See page 367, copying and pasting cells.*

⇒Totaling cells

1. Select the cells that you want to total – these must all be in the same column.
2. Click the *Autosume* button on the toolbar.

The result will appear in a new cell, under the totaled cells.  Microsoft® Corporation

⇒Calculating averages

1. To display the result, choose a new cell under the 'calculation' cells.
2. Click the *Function Wizard* button on the toolbar.
3. On the right side of the *Paste Function* box are a list of *Function names*; choose

Average.

4. Click **OK**.
5. A grey box will open; if necessary, move the box by dragging it.

Option **Number 1** holds the range of cells that Excel will find the average of. If the range is incorrect, highlight the cells on the worksheet that you want to find the average of - this will modify the contents of option **Number 1**.

6. Click OK.

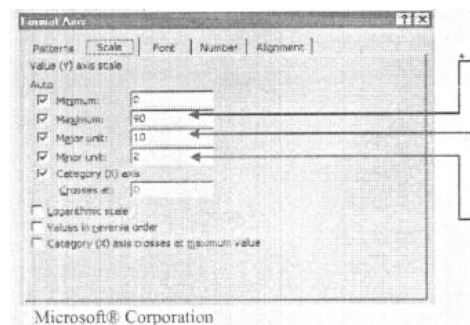
Tip: to reduce the number of decimal places, use the *decrease/increase decimal places* button in the toolbar.

⇒Creating Charts

1. Type in the data. The category axis (X-axis) labels in a single column and the value axis (Y-axis) labels at the top of each column after that.
2. Select your data, including the category and value axis labels.
3. Click the *Chart Wizard* button on the toolbar.
4. Follow the Chart Wizard's instructions. Click **Next** to move through the stages. 

Changing the Values of the Value axis (Y-axis)

1. Double click the chart's Value axis.
2. From the *Format Axis* box, click the *Scale* tab.
3. Make any necessary changes, and click **OK**.



⇒ Printing

(If you want to print the chart without data, select the chart before choosing Print.)

1. From the **File** menu, choose **Print...**
2. Under *Print what*, on the left of the print box, decide what you want to print;

Active sheet(s) Entire Workbook Selection Selected Chart

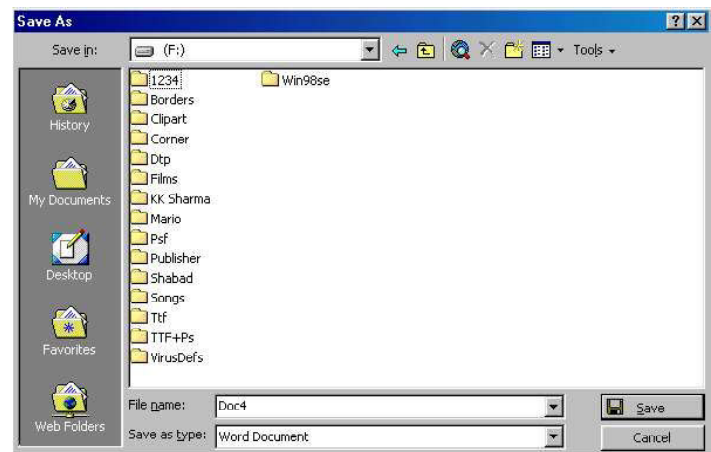
prints the current worksheet;

prints every worksheet that contains data; prints selected data;

prints the currently selected chart.

⇒ Saving your File

1. Choose **SAVE AS...**
from the **FILE** menu.
2. Click in the box next to *File Name*, and enter a name for your publication.
3. Click on the drop-down menu next to *Save In*, and choose the Home Directory Folder as



indicated by your student username e.g. aq 123456

4. Click the **SAVE** button.
5. To save again click the save button
6. It is good practice to save every 10 minutes or so and make a backup of your work to Pen drive or CD.

Formatting

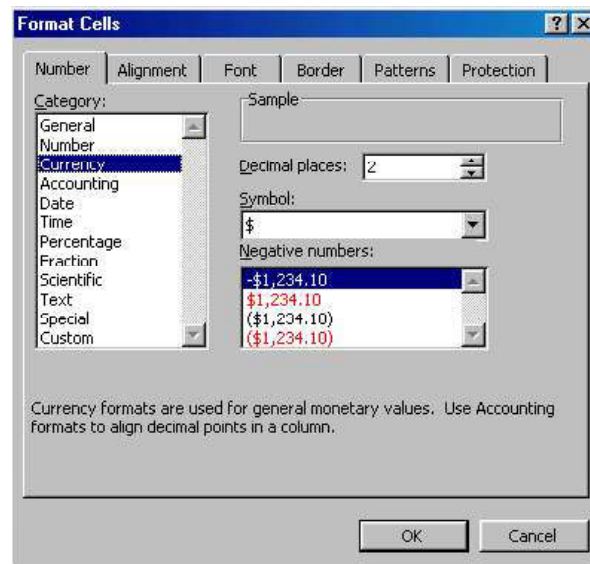
Any cell or range of cells can be given one of a number of formats, which affect the way the cell content looks and prints but do not change the actual contents. We will change the results of the Profit Forecast to show as monetary amounts with a

£ symbol in front.

Note that the precision with which numbers are stored is not altered by the process – only the appearance changes.

Highlighting a range is achieved by moving the mouse pointer to the top-left cell of the range, holding the left mouse button and dragging the pointer to the bottom right cell.

- ☐ Highlight the range.
- ☐ Chose **C**ells from the **F**ormat menu
- ☐ Ensure you have the 'number' tab selected, click on 'currency' from the category section as in the example below :



- ☐ Click on OK

The figures on the worksheet will now be prefixed by a £ symbol and should display 2 decimal places.

Erasing cell contents

Deleting a row or column causes the surrounding rows or columns to close up and fill the gap. In order to blank out the contents of a cell or range without disturbing the rest of the sheet, the **Delete** key on the keyboard is use.

- ☐ Highlight the range
- ☐ Press the **Delete** key

The contents of the cells have been erased. The formatting applied earlier will remain associated with the empty cells.

Creating a graph

Creating a graph needs several steps. To produce the graph, Excel needs to know the location of the data to be plotted and what type of graph (Column, Bar etc.) to plot. Other information such as titles and legends can be added to the graph but these are optional.

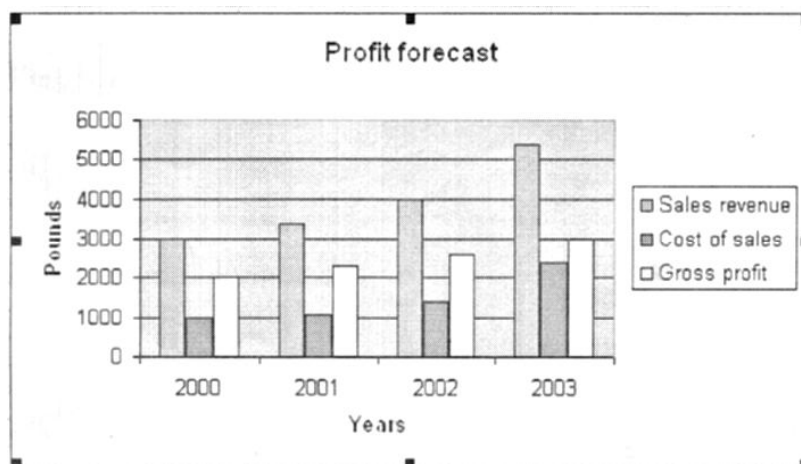
Excel does however have a **chart wizard**, which takes you through the steps needed to create a graph.

The example we will use is the Profit Forecast and it is reproduced below with the associated graph for the Sales Revenue, Cost of Sales and Gross Profit.

CHECK YOUR PROGRESS-I

Fill in the blanks with suitable answers:

- a. MS-Excel contains _____ worksheet, by default.
- b. Texts are _____ aligned automatically in the cell.
- c. _____ operators are used to perform mathematical calculations.
- d. IF is a logical function which returns the result in terms of true or _____.
- e. We can use _____ shortcut key to print the current worksheet.



Defining the plot area

As we are going to plot the Sales Revenue, Cost of Sales and Gross Profit for 2000-2003, all of this data needs to be highlighted for the **chart wizard** to create the graph.

- ☐.....Highlight A3 : E8
- ☐Hold the **Ctrl Key**
- ☐Highlight A9:E11
- ☐.... Let go of the Ctrl key and the 2 separate ranges should be highlighted :

	A	B	C	D	E
1	Profit Forecast				
2					
3		2000	2001	2002	2003
4					
5	Sales (units)	2000	2104	2335	3000
6	Price/unit	1.5	1.6	1.7	1.8
7	Cost/unit	0.5	0.5	0.6	0.8
8					
9	Sales revenue	3000	3366.4	3969.5	54000
10	Cost of sales	1000	1052	1401	2400
11	Gross profit	2000	2314.4	2568.5	3000
12					

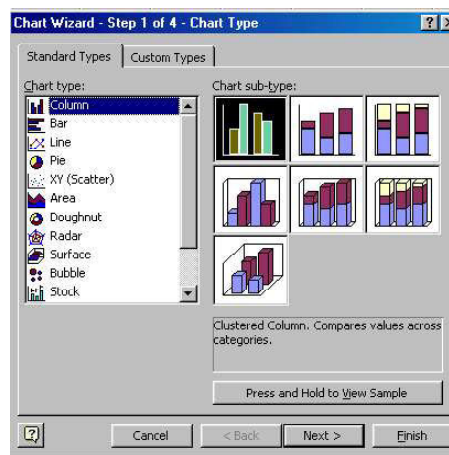
It is important that the cell A3 is highlighted, even though it contains no data as Excel creates graphs using a pattern matching process - it will match one row by five columns against three rows by five columns. Don't worry about this at the moment, just make sure the 2 ranges are highlighted as in the diagram above.

Using the chart wizard

- ☐.....Click on the Chart Wizard Button Step one of the  Chart Wizard will be displayed.

Step one - chart type

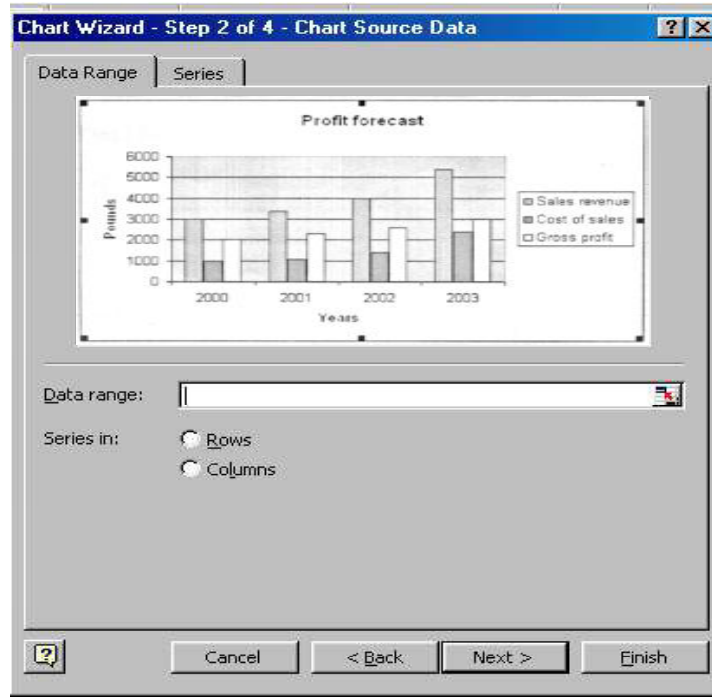
This box allows you to choose the **Chart Type**.



☐.....Select the type indicated then click on **Next**.

Step two - data definition

The data range selection box is shown below:



As you selected the data before starting, this should be correct. Click on **Next**.

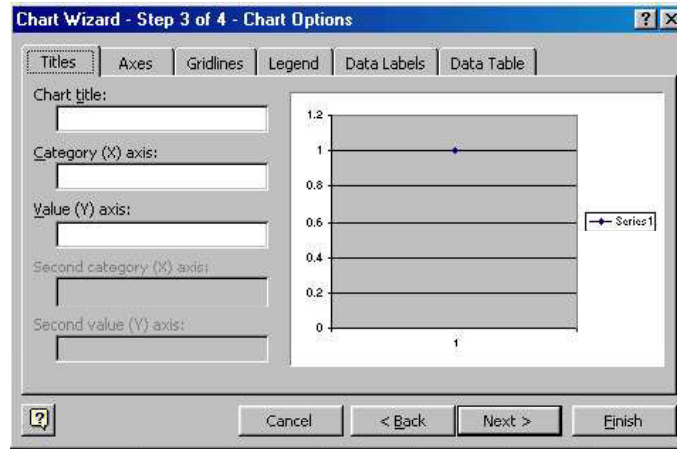
Step three - entering titles and labels

Next is the chart title and axes labels. Enter in the appropriate areas:

Profit Forecast for the title

Years for the X axes labels

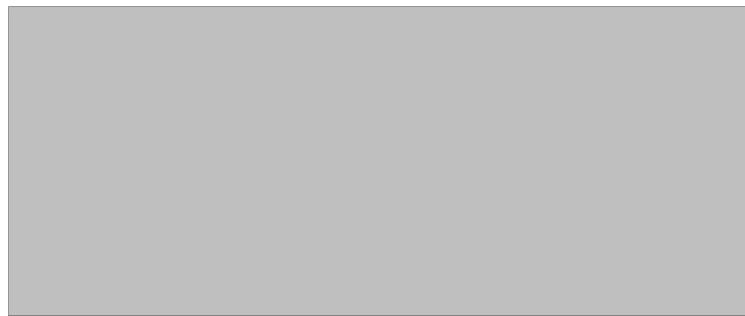
Pounds for the Y axes labels



There are lots of other options in this section, which you can look at later, for now just click on **Next**.

Step four - chart location

Next is the option to insert the chart as a new sheet or as an object on the sheet



Choose the option you want and then select **Finish**

Your graph will be inserted in the workbook either as a new sheet or as an object depending upon your selection. You have now completed your graph

17.3 LET US SUM UP

Hence in excel all the mathematical operations can be done sufficiently by creating customized formulae. This software is also thus helpful to perform statistical operation with graph and charts. It is a basic utility in the world of Statistics.

17.4 GLOSSARY

- In excel, charts and graphs can be used to show your arithmetic data.
- You can copy the work done on excel to your word document if the need arises.
- You can do advanced programming to MS- excel to customize it for any particular operation.

17.5 SELFASSESSMENT QUESTIONS

1. Give the brief description about Microsoft excel and its functions.

2. What is the main difference between MS word and MS excel.

17.6 LESSON END EXERCISE

1. MS - Excel is used to perform Various arithmetic operations (True/ False)
2. In excel simple to complex data processing is possible if you know little programming (True/ False)

17.7 SUGGESTED READINGS

1. Argyrous, George. 1997. *Statistics for Social Research*. New York: Mc Millan Press Ltd.
2. Goods, W.J. & Hatt, P.K. 1981. *Methods in Social Research*. New York: Mc Graw Hill.
3. Gupta, S.C. 1981. *Fundamentals of Statistics*. Bombay: Himalayan Publishing House.
4. Gupta, S.P. 2004. *Statistical Methods*. New Delhi: Sultan Chand and Sons.

17.8 ANSWERS TO CHECK YOUR PROGRESS

- a. MS-Excel contains 3 worksheets, by default.
- b. Texts are left aligned automatically in the cell.
- c. Arithmetic operators are used to perform mathematical calculations.
- d. IF is a logical function which returns the result in terms of true or false.
- e. We can use Ctrl + P shortcut key to print the current worksheet.

M.S. POWER POINT

STRUCTURE

- 18.0 Objectives
- 18.1 Introduction
- 18.2 Tool Bar, Icons and Commands
- 18.3 Navigating in Power Point
- 18.4 Working with Power Point
- 18.5 Let us sum up
- 18.6 Glossary
- 18.7 Self-Assessment Questions
- 18.8 Lesson End Exercise
- 18.9 Suggested Readings
- 18.10 Answers to check your progress

18.0 OBJECTIVES

The main objectives of this lesson are:

- To understand the meaning of Power-Point.
- To understand how to use power point

18.1 INTRODUCTION

Power Point is a presentation tool that helps you create eye-catching and effective presentations in a matter of minutes. A presentation comprises of individual slides arranged in a sequential manner. Normally, each slide would cover a brief topic. Once having prepared a presentation, you can ask PowerPoint to also generate hand out material and speaker's notes. Similarly, you have, the option of either printing out the slides-in case you want to use an overhead projector, or simply attach your computer to an LCD display panel that enlarges the picture several times and shows you the output on a screen.

When you create a new presentation, you have three options:

1. You can start by working with a wizard (called the AutoContent Wizard) that helps you determine

- the theme, contents and organization of your presentation by using a predefined outline, or
2. You can also start by picking out a PowerPoint Design Template, which determines the presentation's color scheme, fonts, and other design features, or
 3. You can also begin with a completely blank presentation with the color scheme, fonts, and other design features set to default values.

Should you decide to choose the third option, PowerPoint designers have provided a wide assortment of predefined slide formats and Clip Art graphics libraries. Through these predefined slide formats, you can quickly create slides based upon standard layouts and attributes.

PowerPoint shares a common look and feels with other MS Office components, and once having mastered Word and Excel, learning PowerPoint is almost like playing a game.

And of course, it is easy to pick up data from Word and Excel directly into a PowerPoint presentation and vice versa.

The next page shows you the various parts of a PowerPoint screen. Also, it is important to familiarize yourself with the PowerPoint tool bars. All these are presented in the following pages. Do not worry if everything does not make complete sense to you at this stage. Once you do the hands-on exercise, you will understand these options much better.

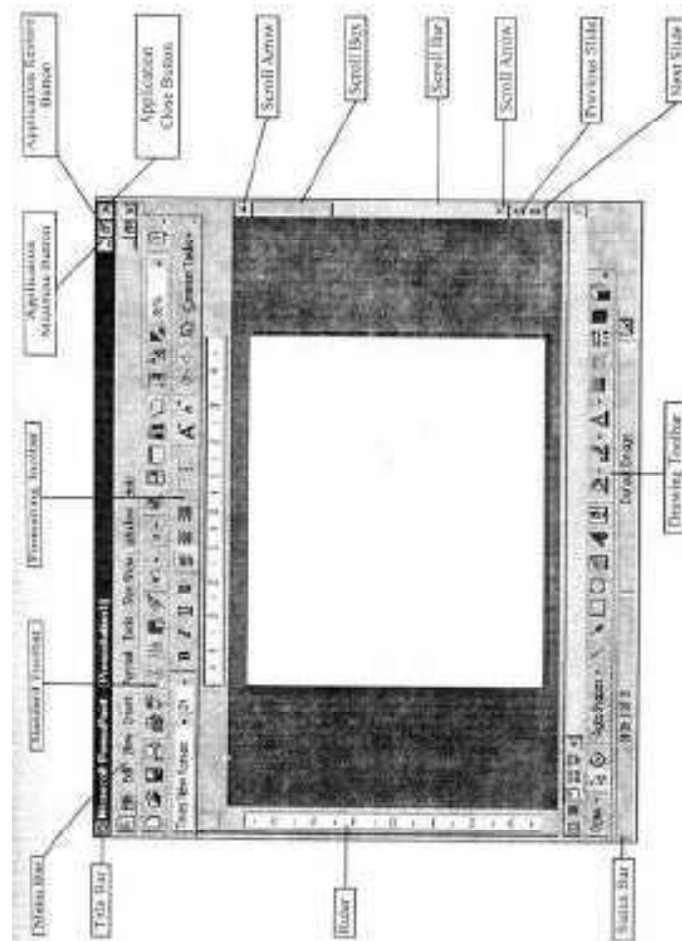
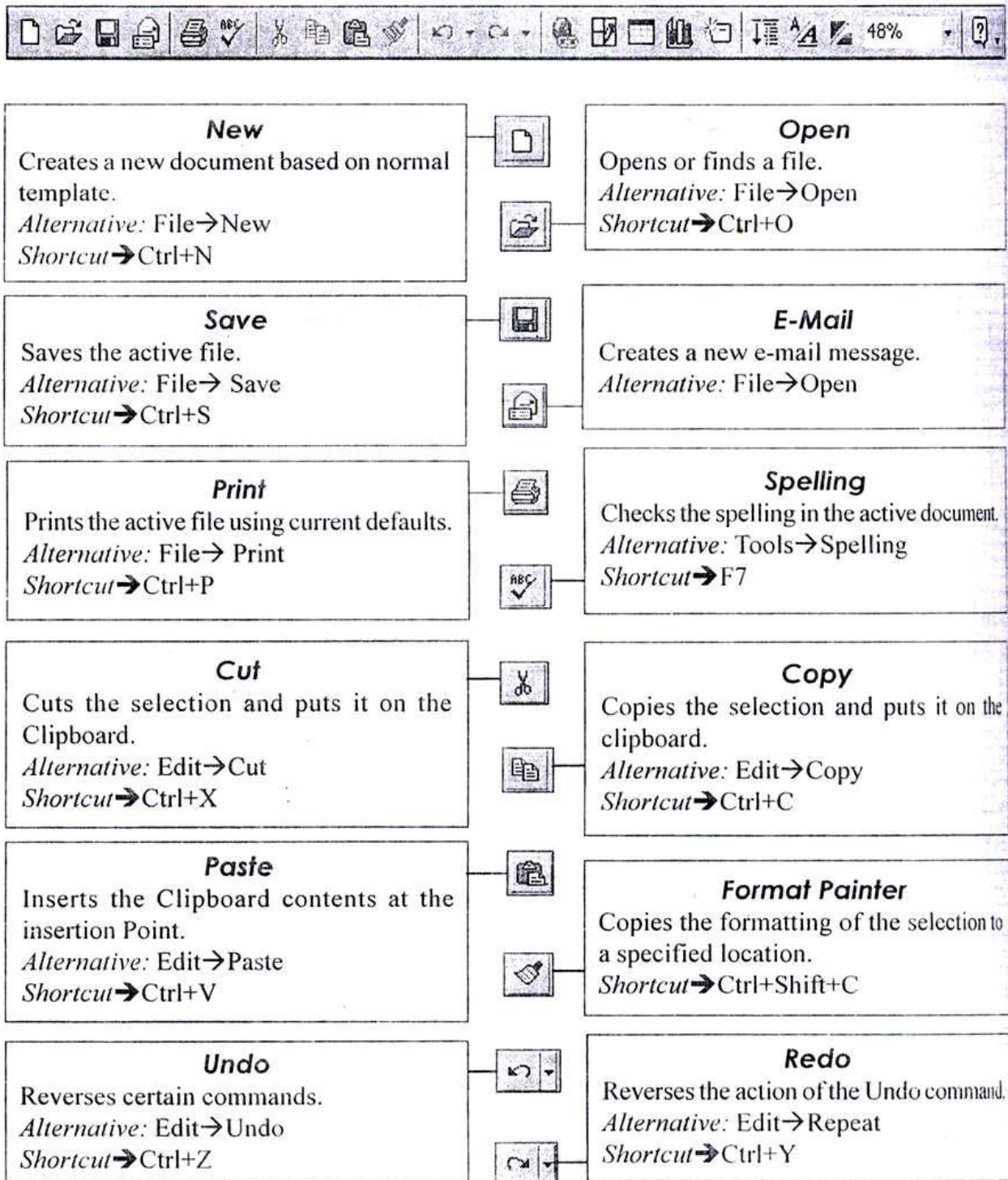
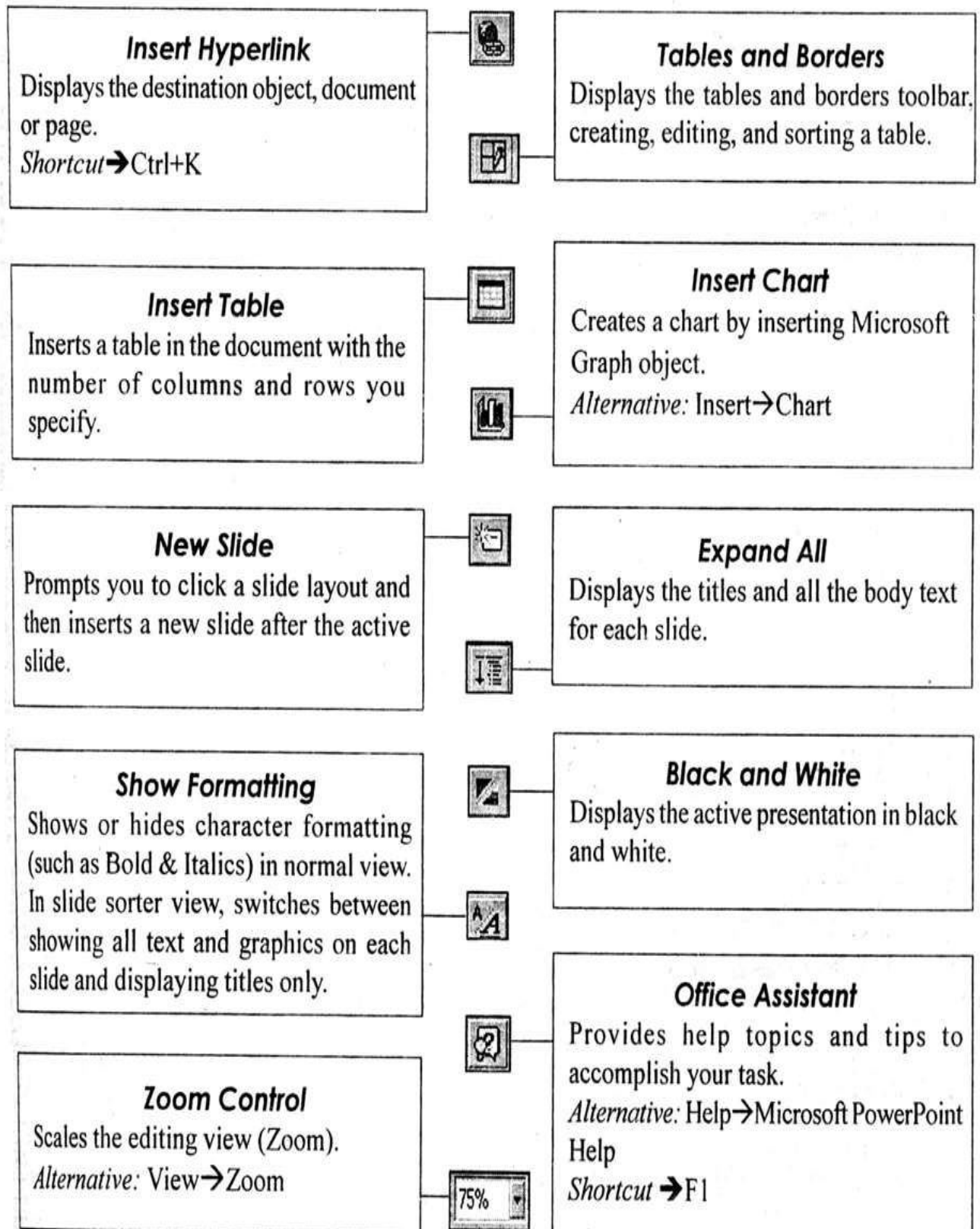


Fig. 14.1 Parts of a PowerPoint Window

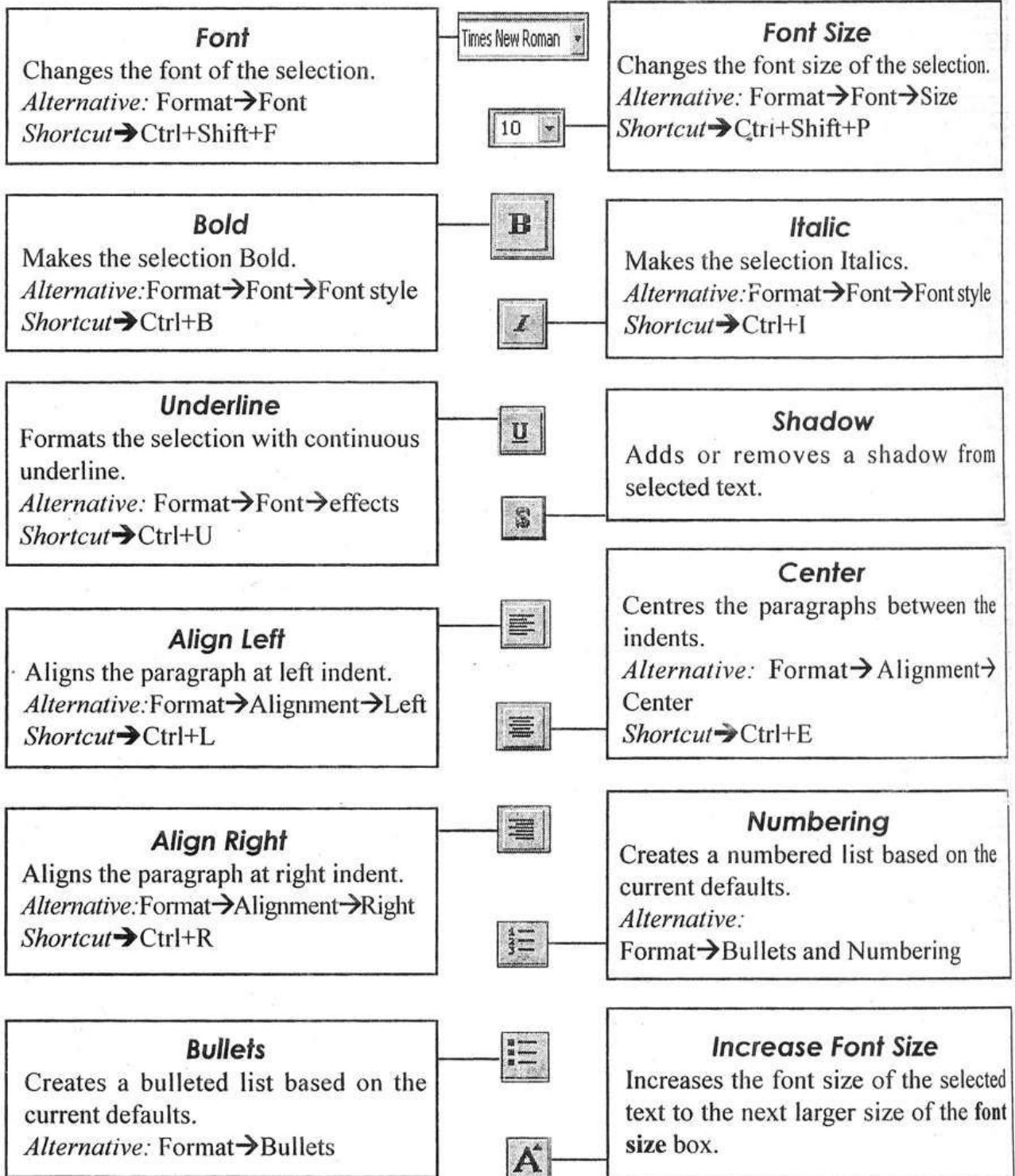
18.2 TOOLBAR, THEIR ICONS & COMMANDS

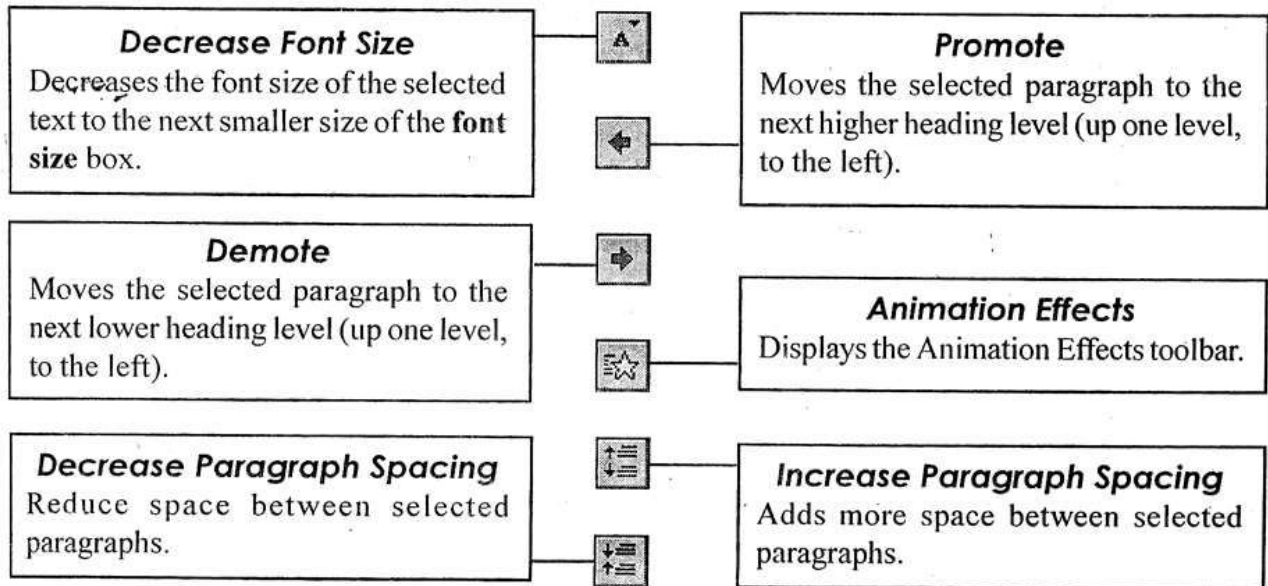
Standard Toolbar



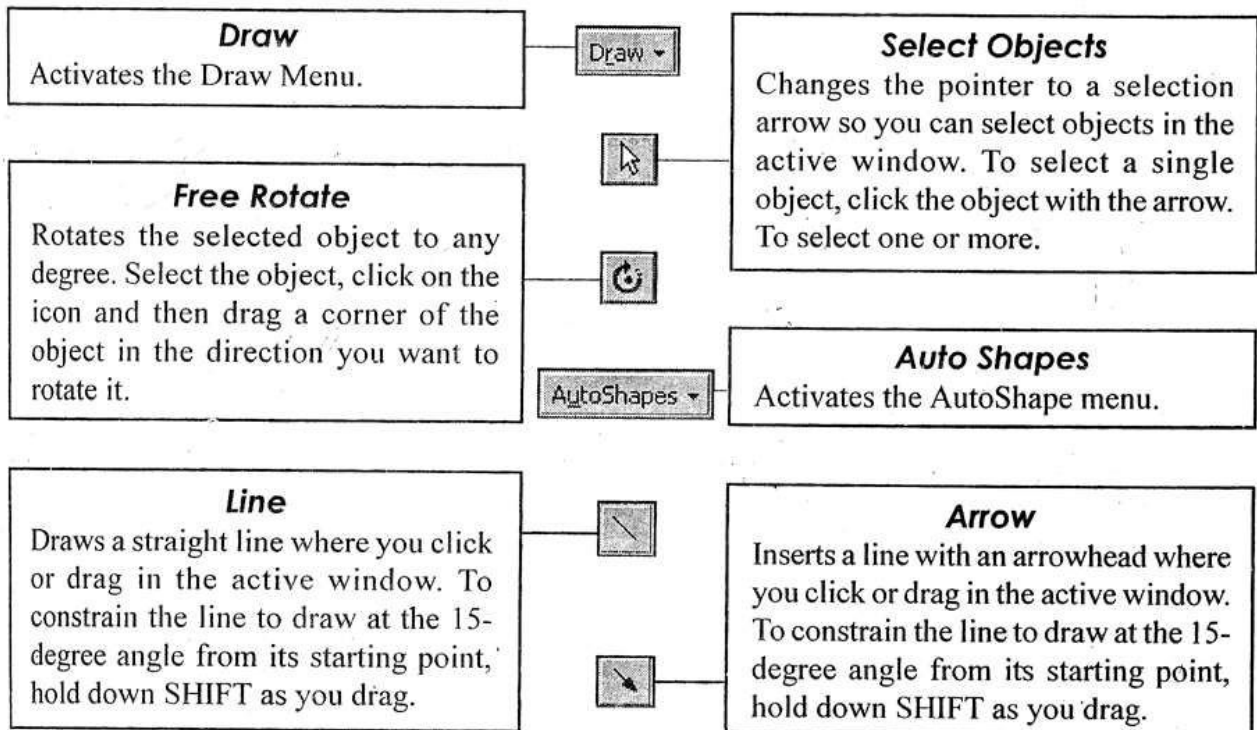


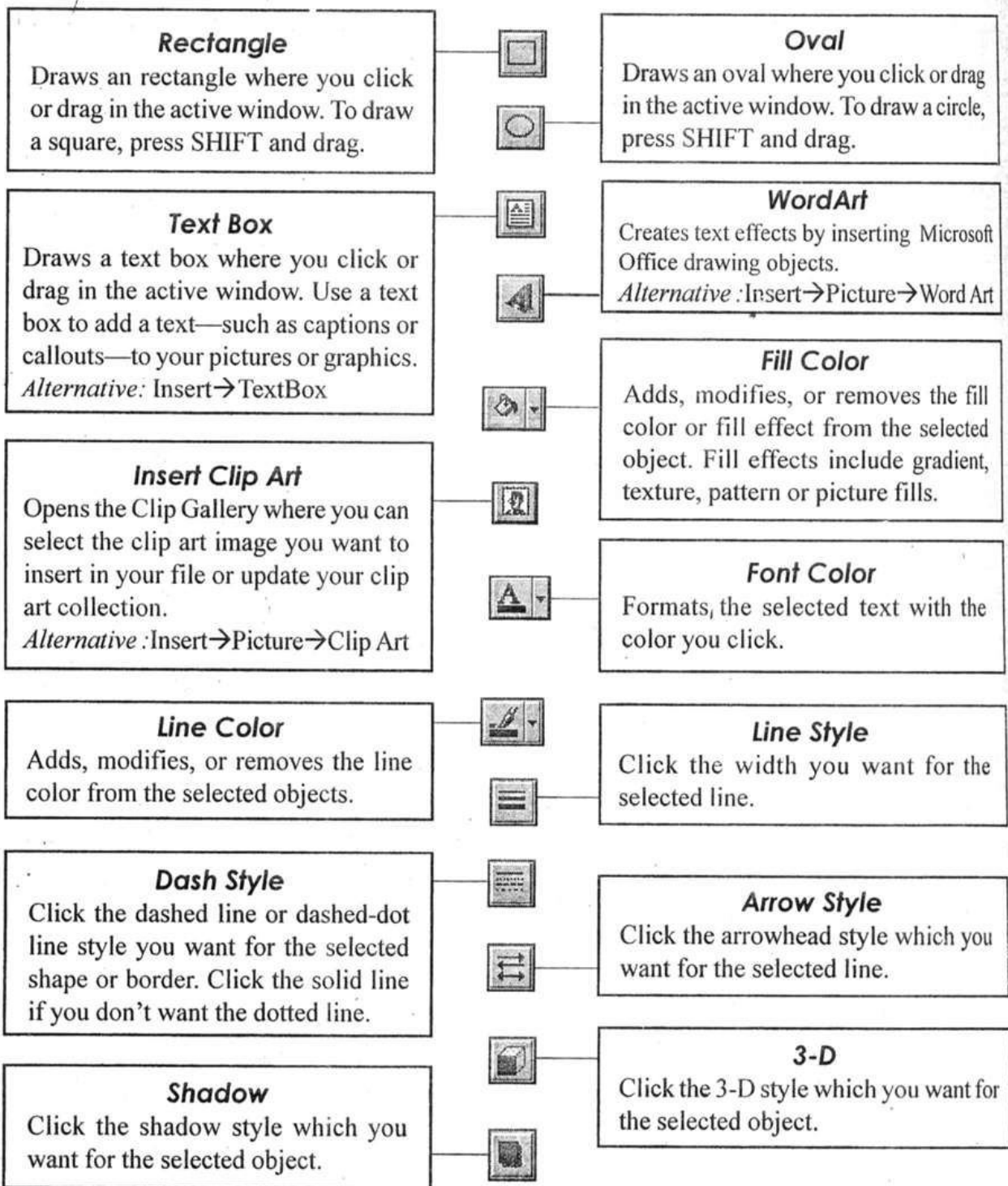
Formatting Toolbar





Drawing Toolbar



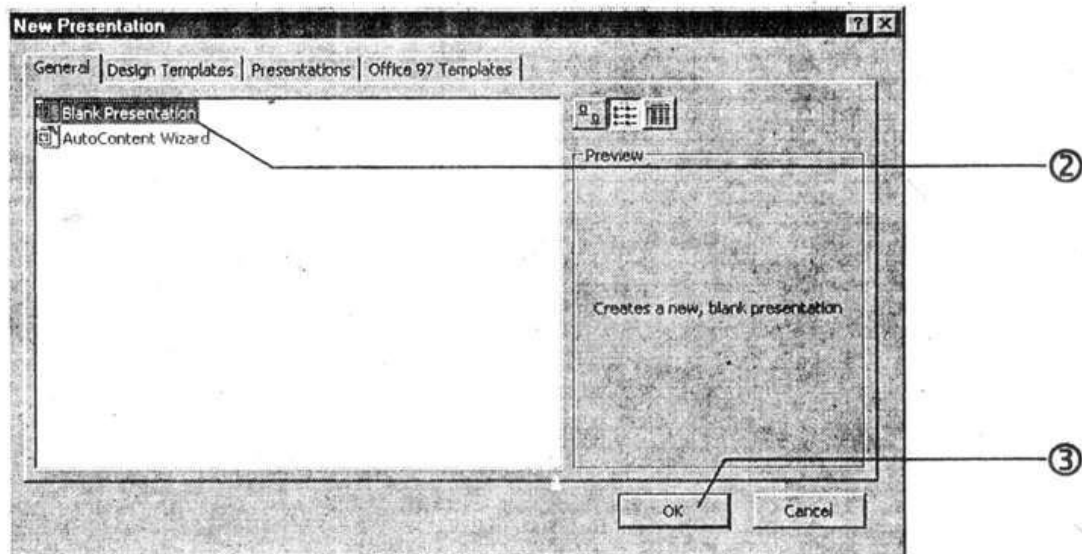
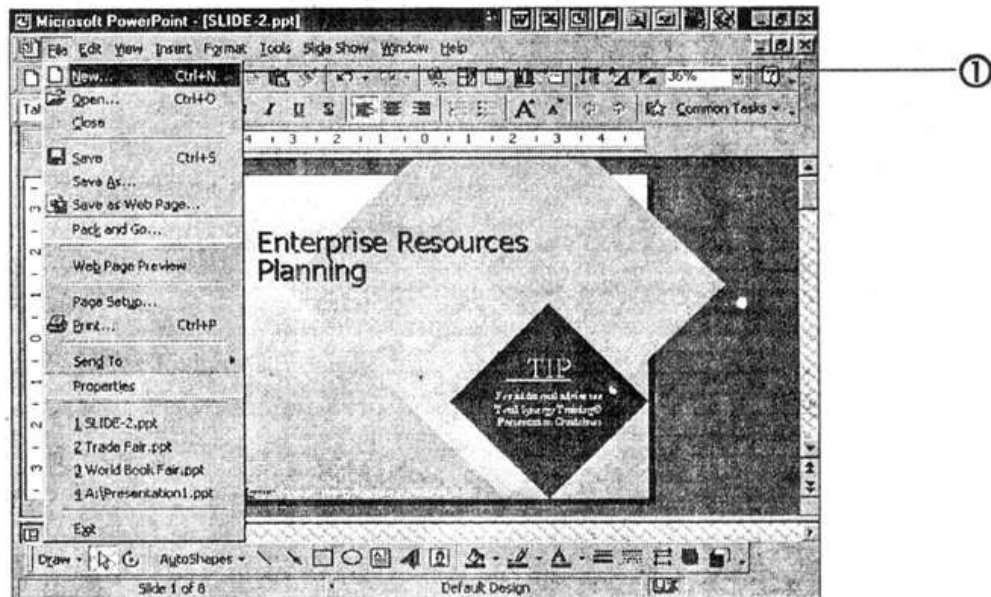


18.3 NAVIGATING IN POWERPOINT

Now that you have a general idea of the various command, & toolbars, let us see how we can carry out simple operation like saving and opening files, creating and printing slides, etc. For this follow the step-by-step instructions given below.

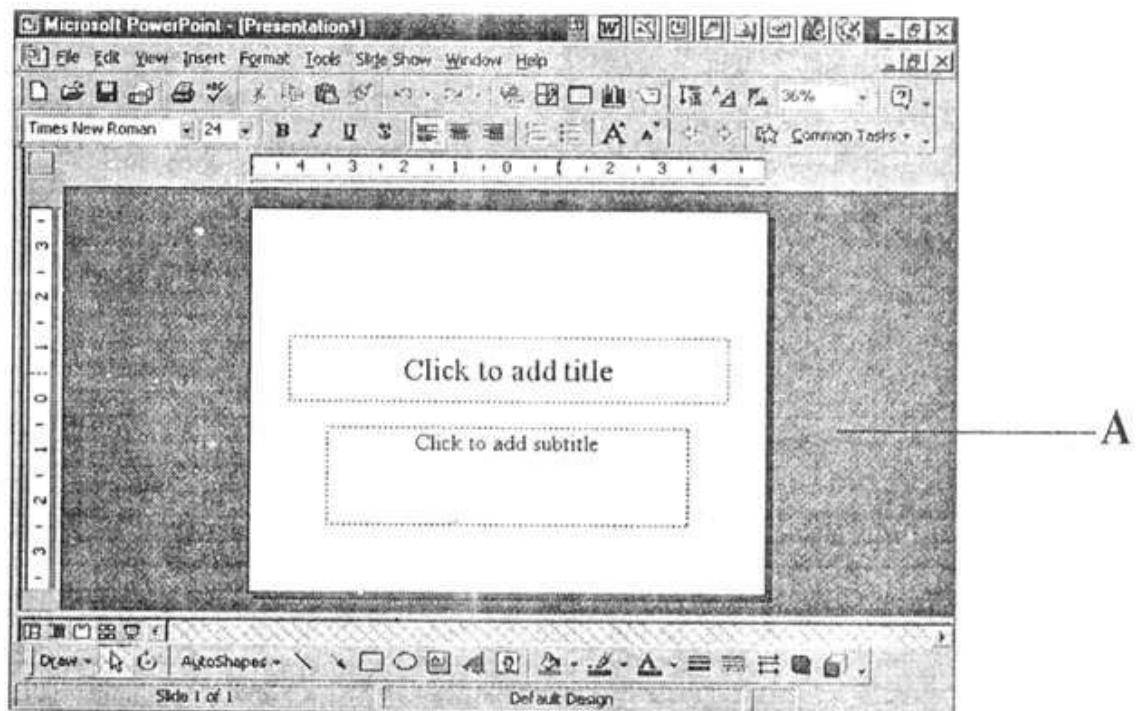
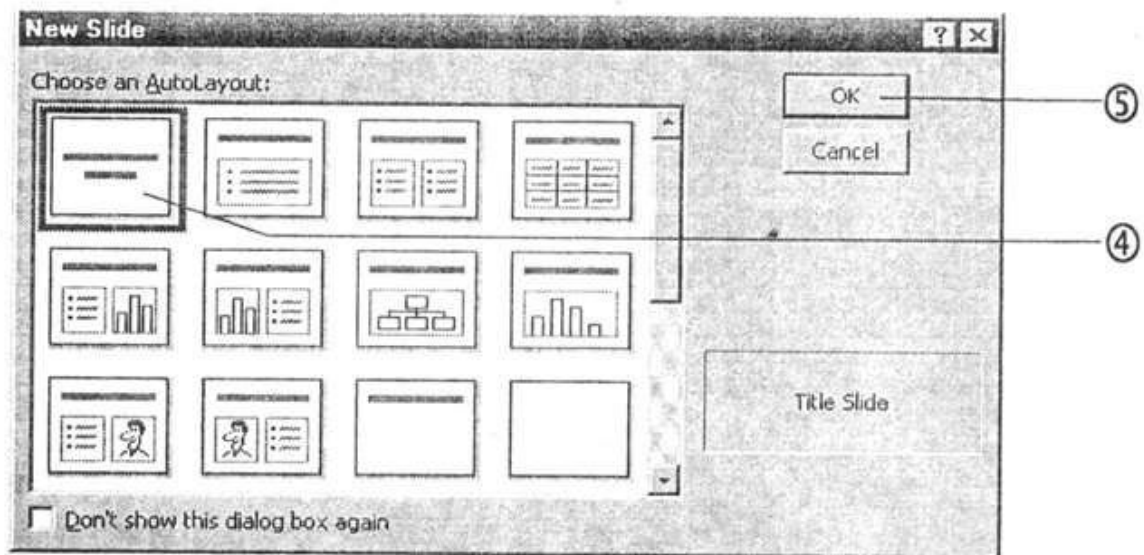
Creating a New Presentation

- 1 Choose New command from File menu.
- 2 The following dialog box would be displayed. Generally this option (Blank Presentation) always comes highlighted, but if not, Click to highlight.
- 3 Click on OK button to continue.



- 4 The following dialog box would appear. Click on the type of slide that you want :
- 5 Click on OK button.

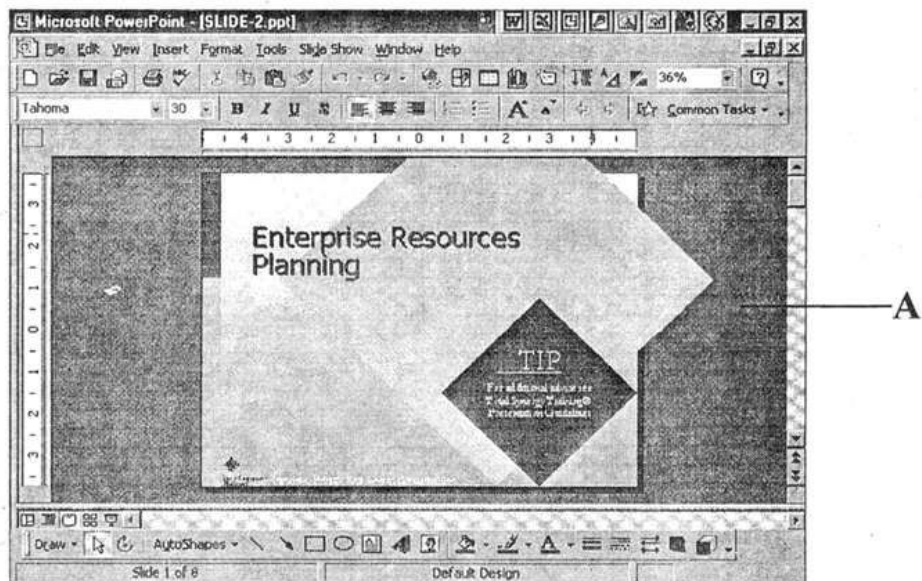
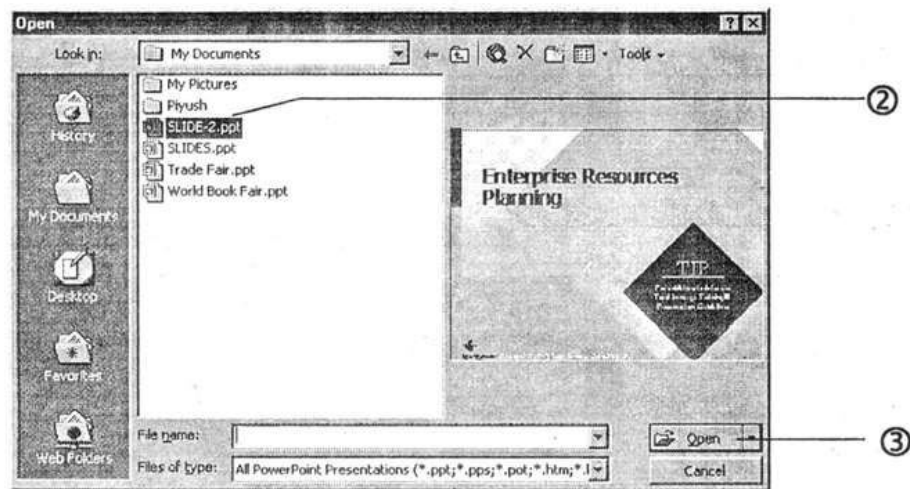
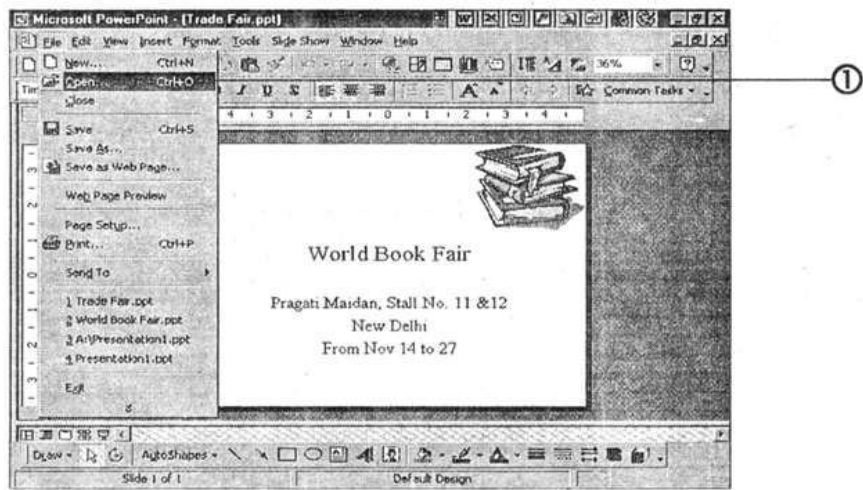
- A The following blank presentation would be displayed.



Opening a Presentation

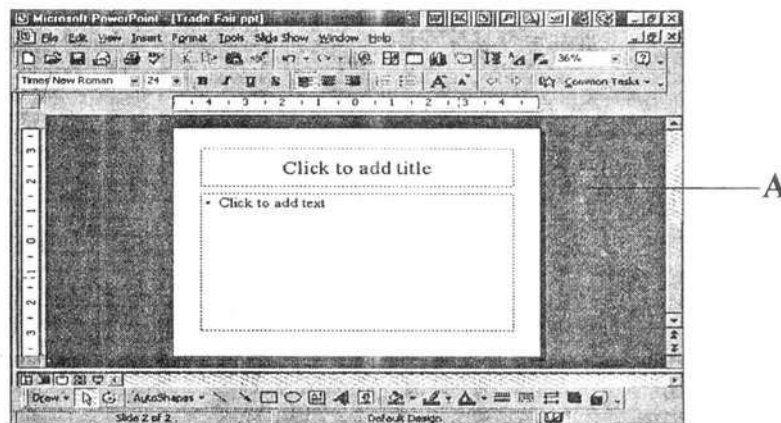
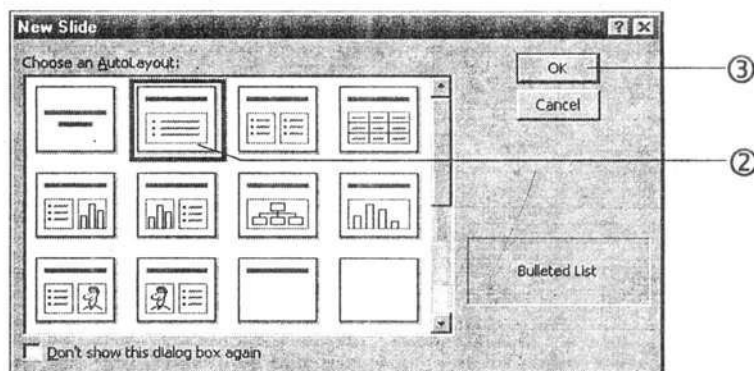
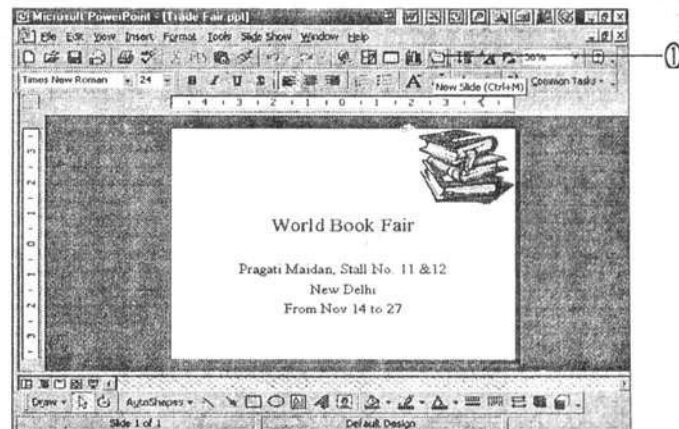
- 1 Choose Open command from File menu.
- 2 Choose the file that you want to open or type the name of the file in the "File Name" window.
- 3 Click on Open button.

A The requested file would open up.



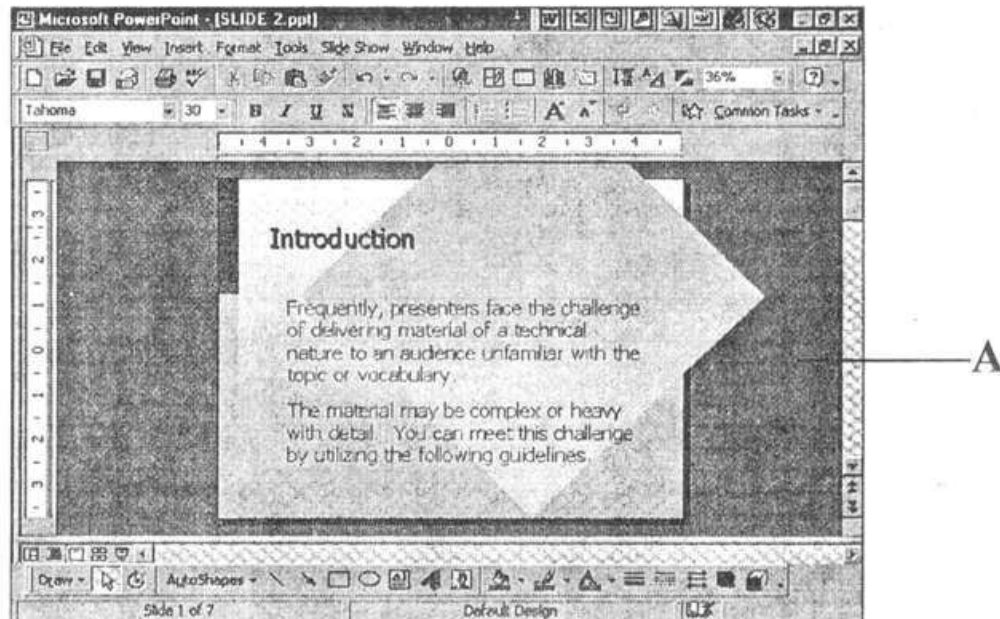
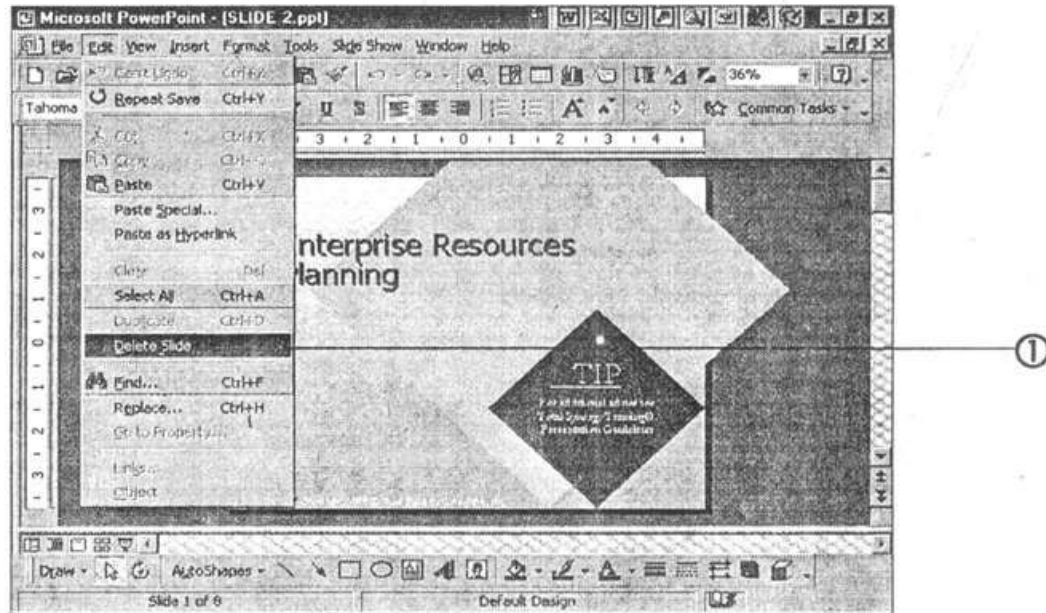
Creating a New Slide

1. Click on the **New Slide** icon from the **standard** toolbar or alternatively, click on the **Common Tasks** roll-down menu and choose **New Slide** command.
 2. Choose the slide type that you want.
 3. Click on **OK** button.
- A A blank new slide would be inserted.



Deleting a Slide

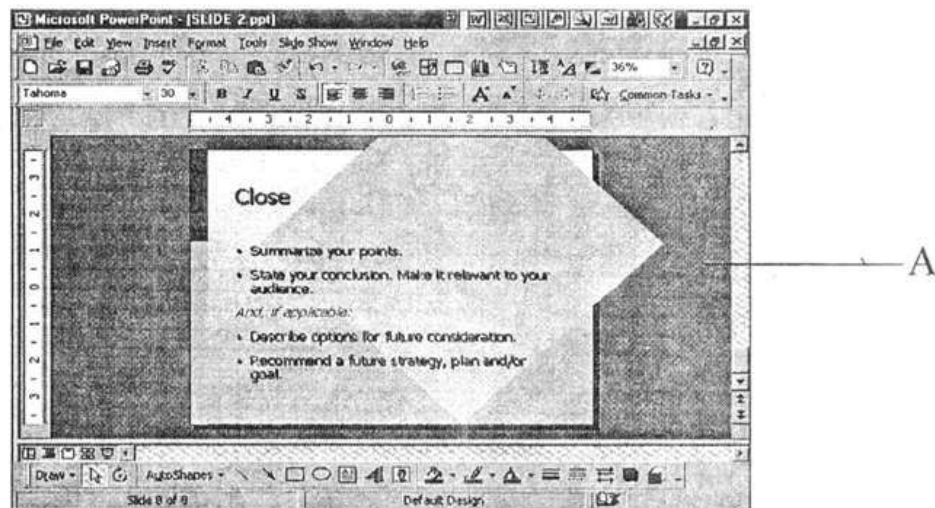
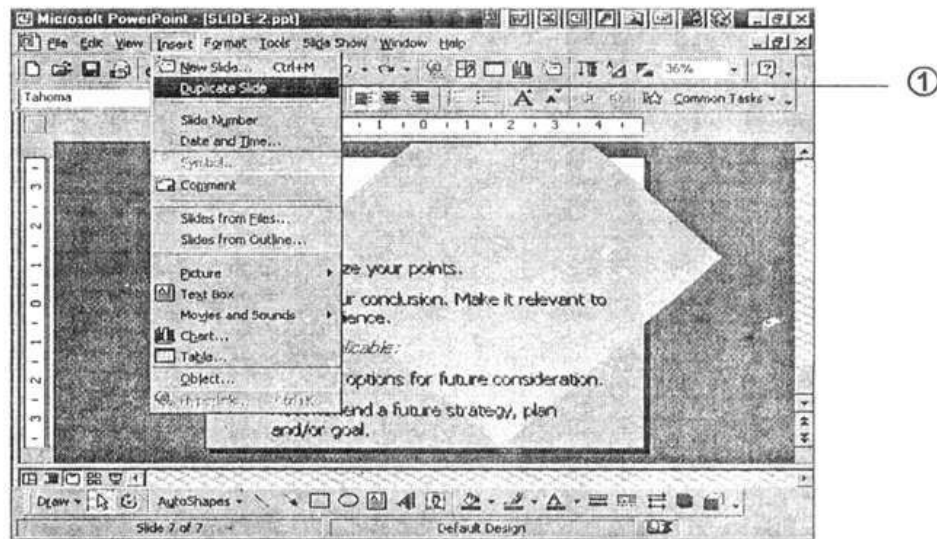
- 1 Choose **Delete Slide** command from **Edit** menu.
- A The slide chosen has been deleted and the number of slides has changed from 8 to 7.



Copying a Slide

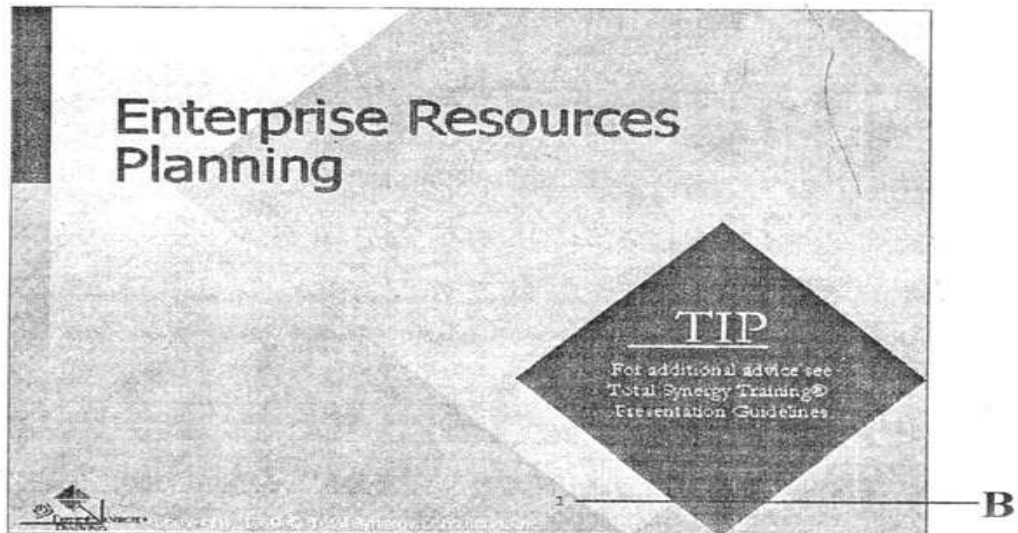
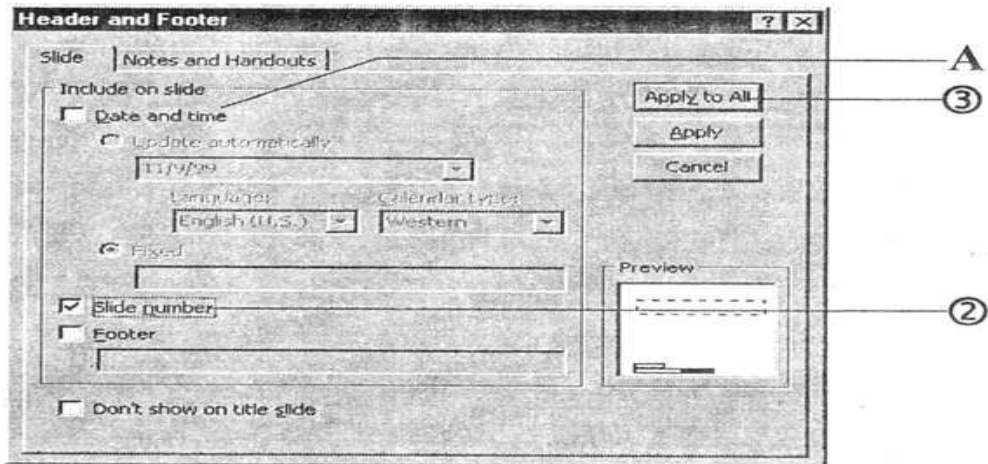
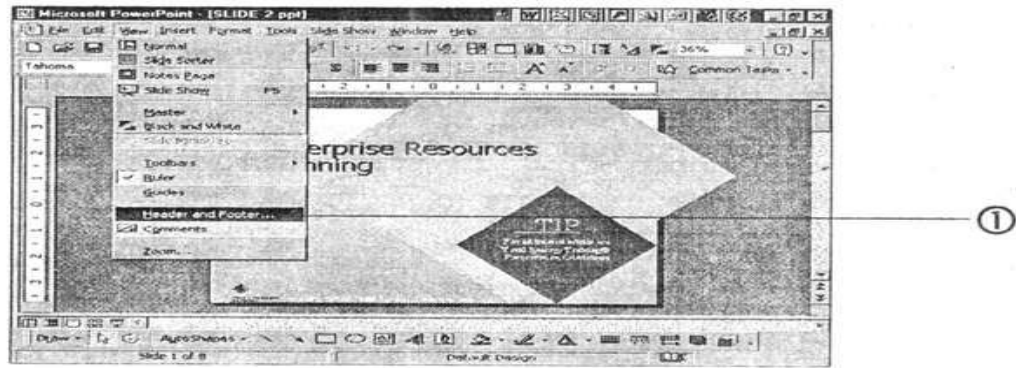
- 1 Choose **Duplicate Slide** command from **Insert** menu.

AA duplicate slide has been inserted.



Slide Numbering

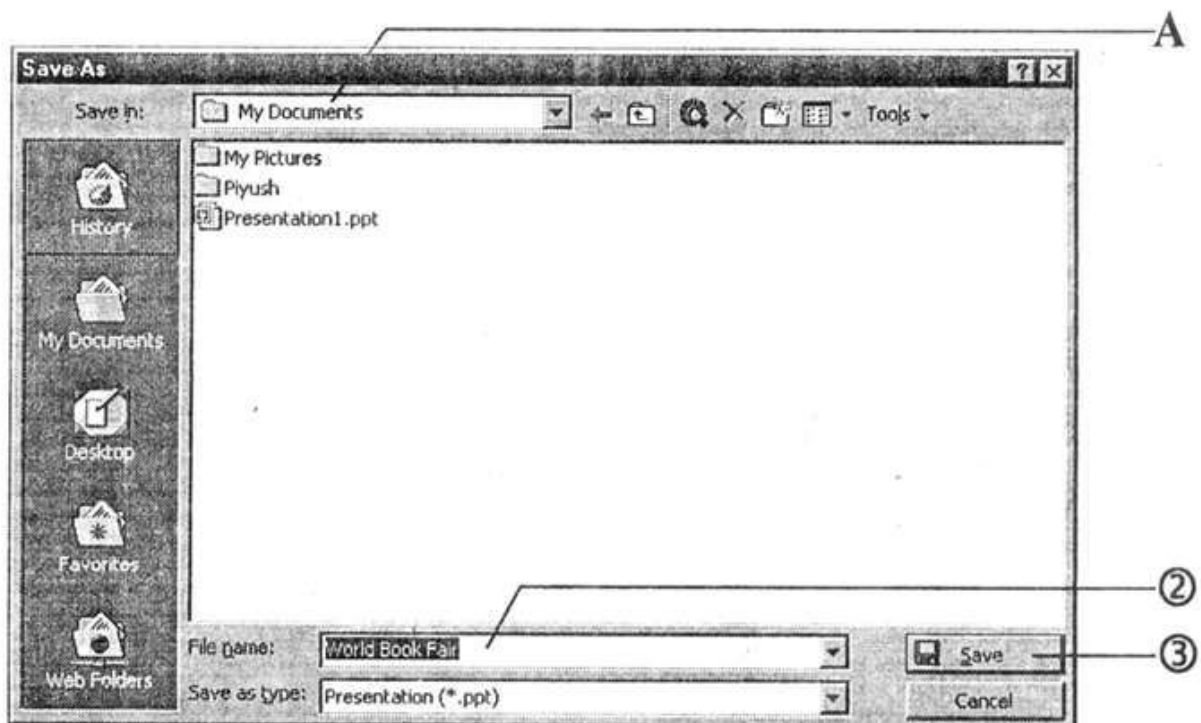
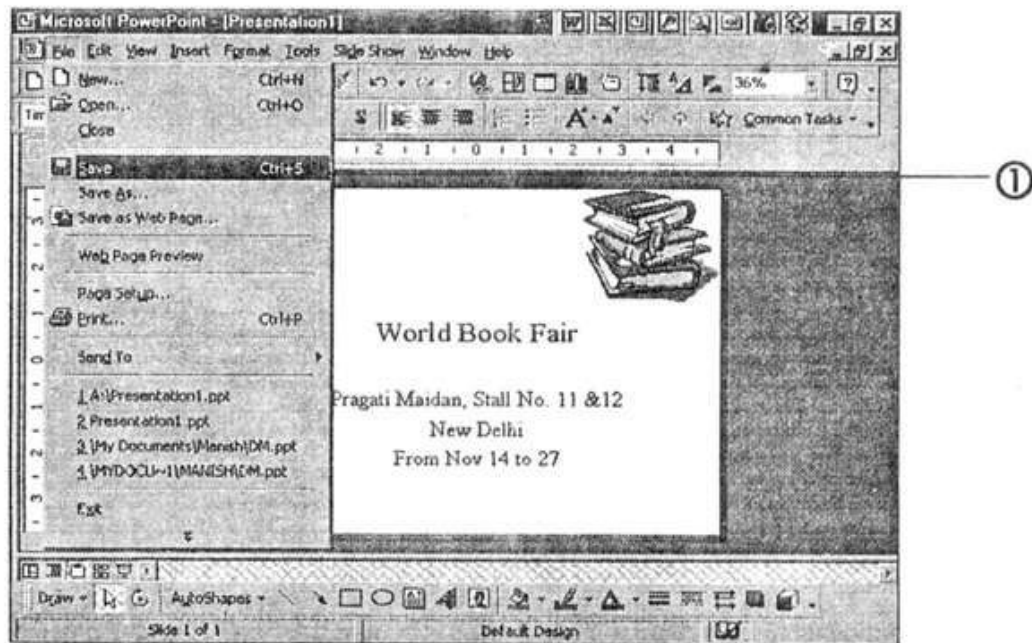
- 1 Choose **Header and Footer** command from **View** menu.
 - 2 The following dialog box would be displayed. Click on the **Slide Number** check box.
 3. Click on **Apply to All** button to apply numbering to all the slides or click on **Apply** button to apply numbering to the current slide only.
- A You can also click on the Date and Time check box to apply the current date and time to the slides.
- B You would notice numbering on all the slides.



Saving a Presentation

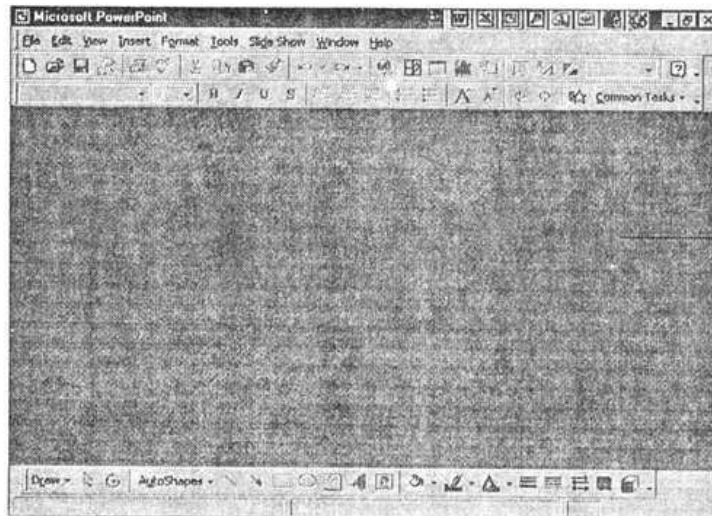
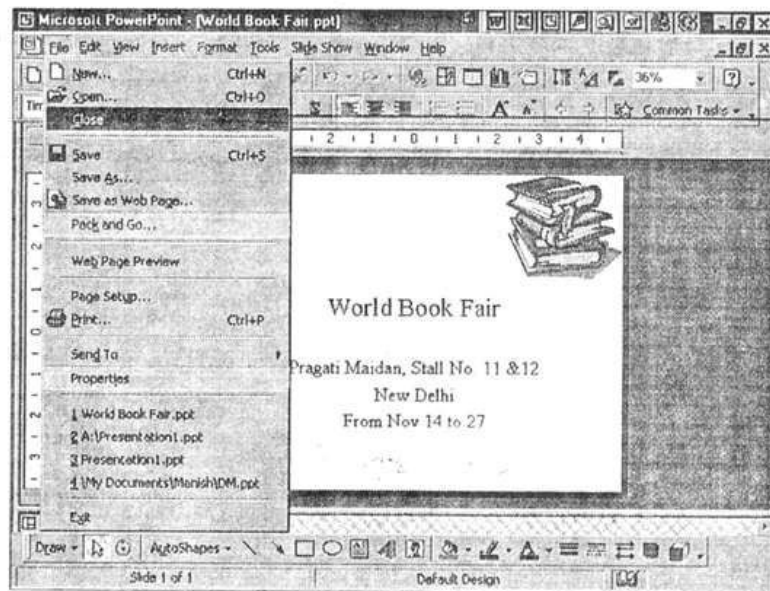
- 1 Choose **Save** command from **File** menu

- 2 The following dialog box would be displayed. Type the name of the presentation.
- 3 Click on **Save** button.
- A In case you wish to save the file into another directory, specify the correct path and directory here.



Closing a Presentation

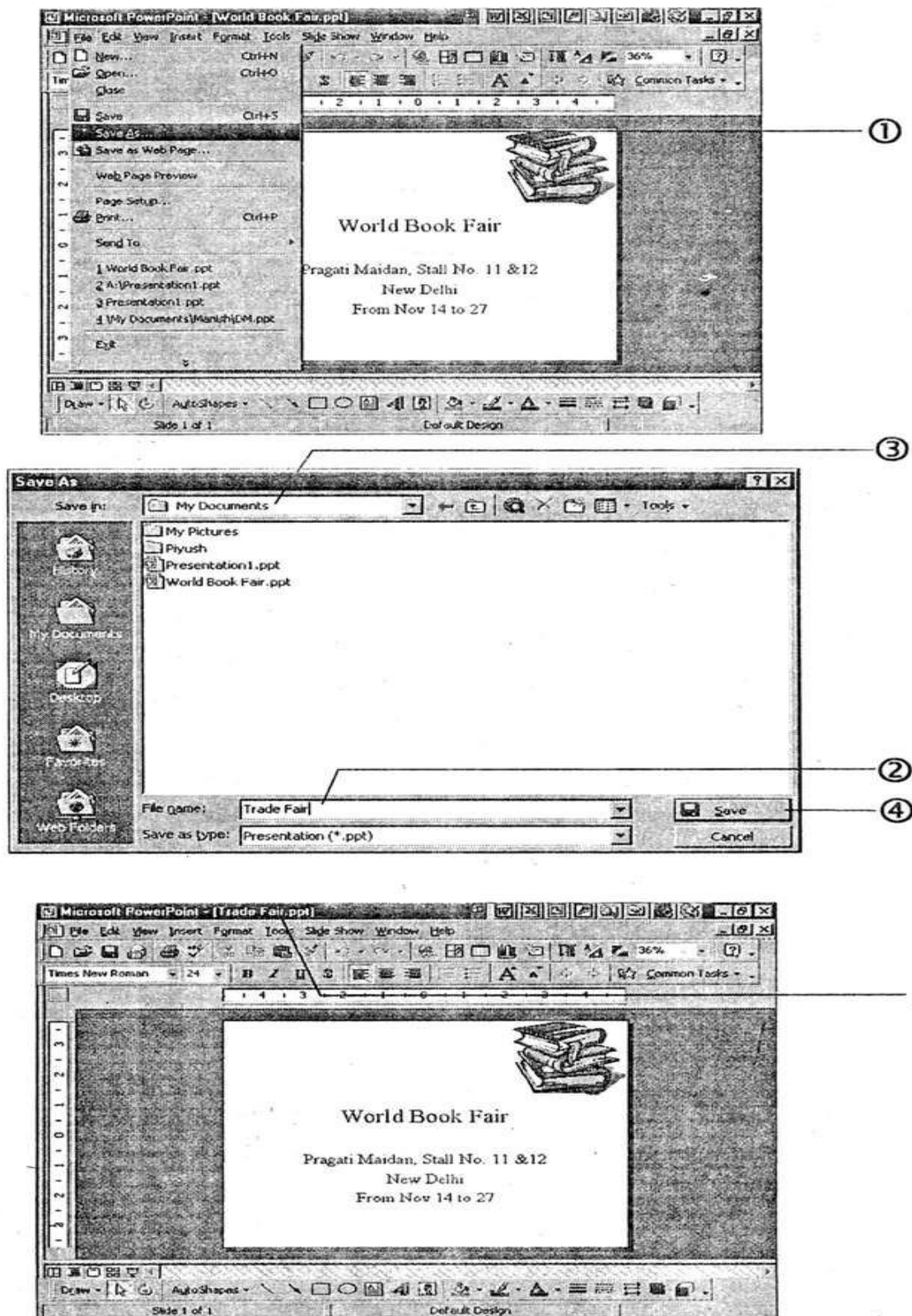
- 1 Choose **Close** command from **File** menu A You would see **PowerPoint's** exit screen.



Saving a Presentation with a Different Name

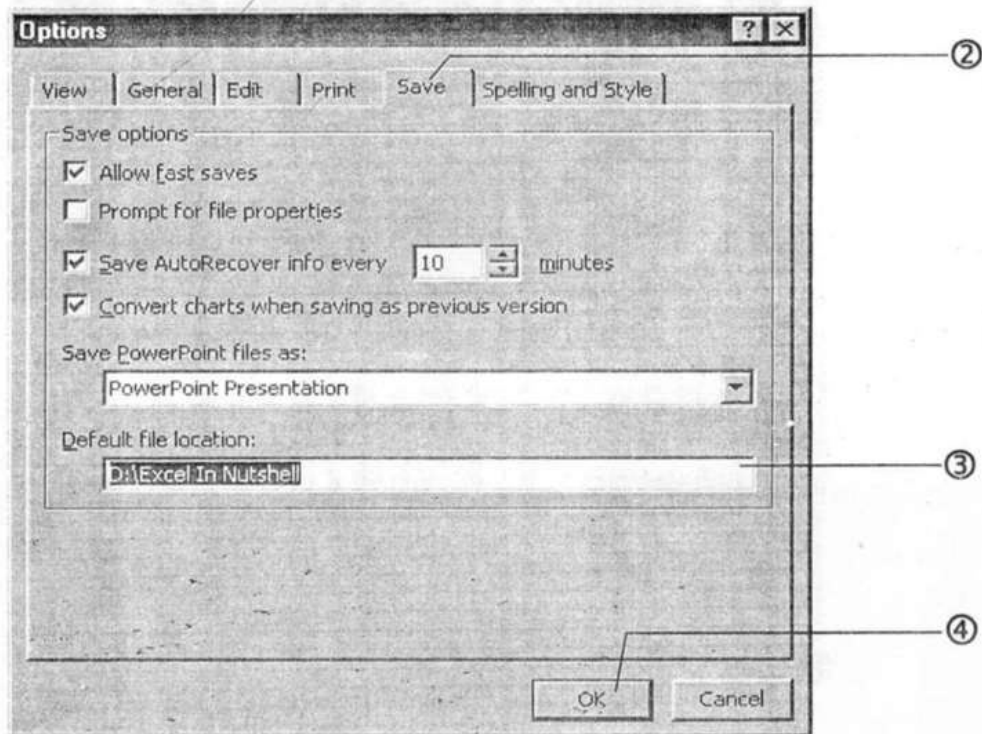
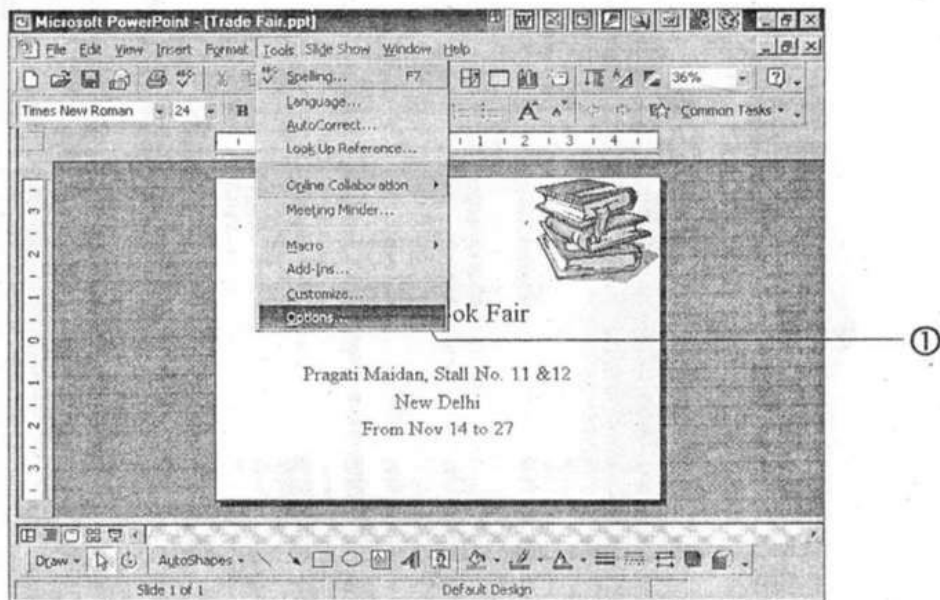
- 1 Choose **Save As** command from **File** menu.
- 2 The following dialog box would be displayed. Type in the name that you want to save the file as.
- 3 Choose the directory in which you want to save the file in.
- 4 Click on **Save** button.

A You would notice that the changed file name appears on the Title bar.



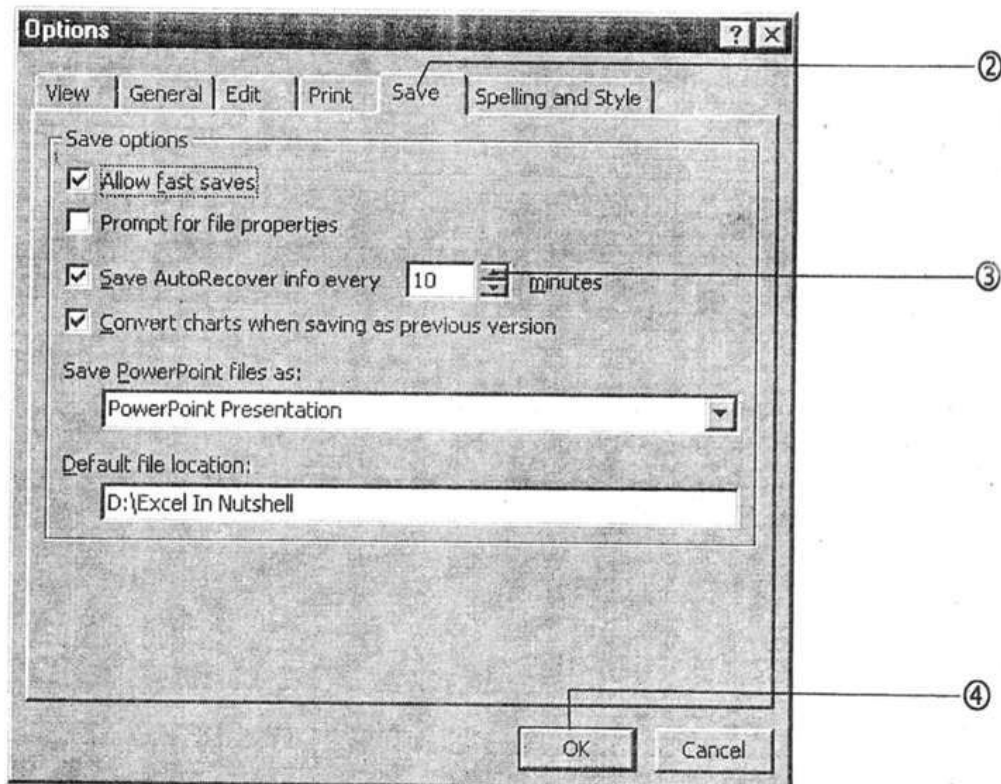
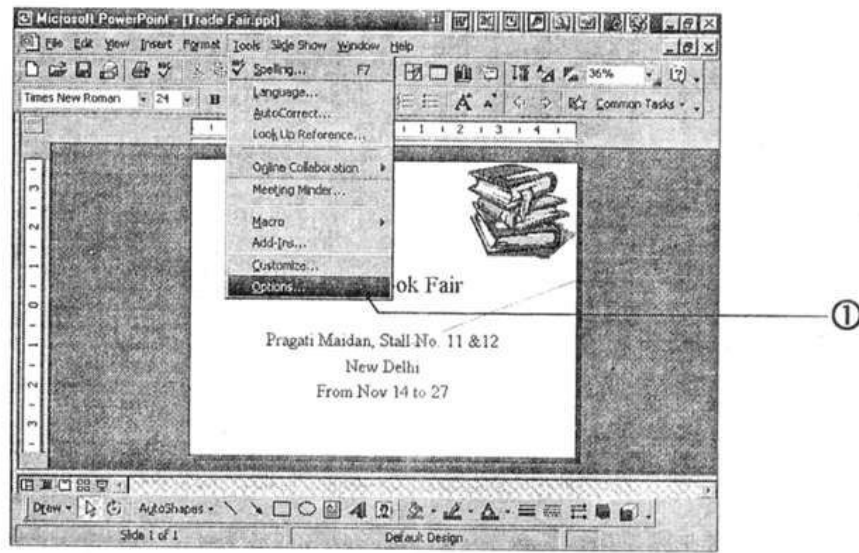
Changing the Default Directory

- 1 Choose **Options** command from **Tools** menu.
- 2 The following dialog box would be displayed. Click once on **Save** folio/tab.
- 3 Type in the name of the Default directory.
- 4 Click on **OK** button.



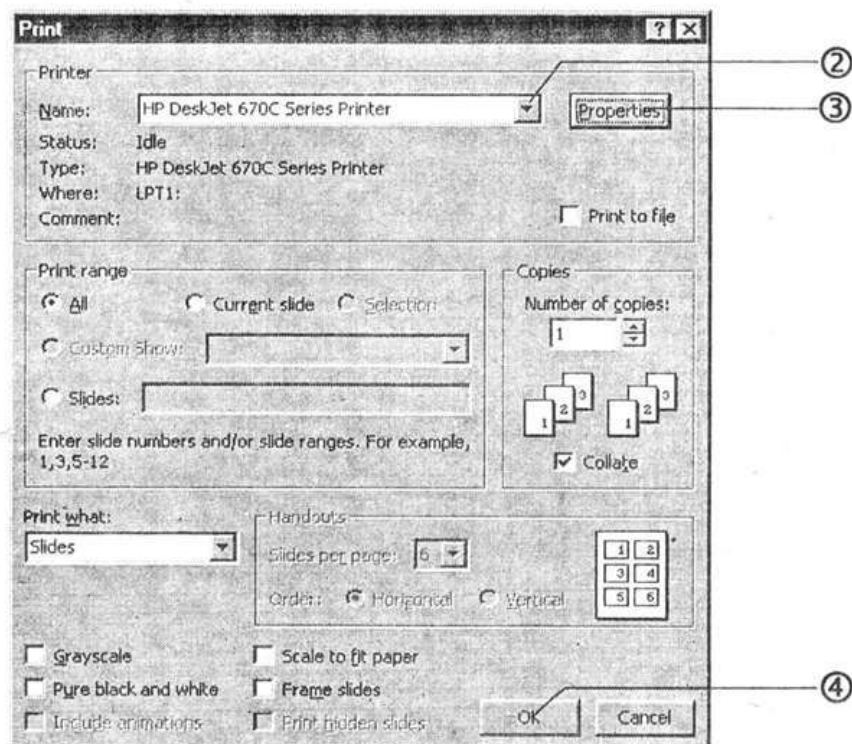
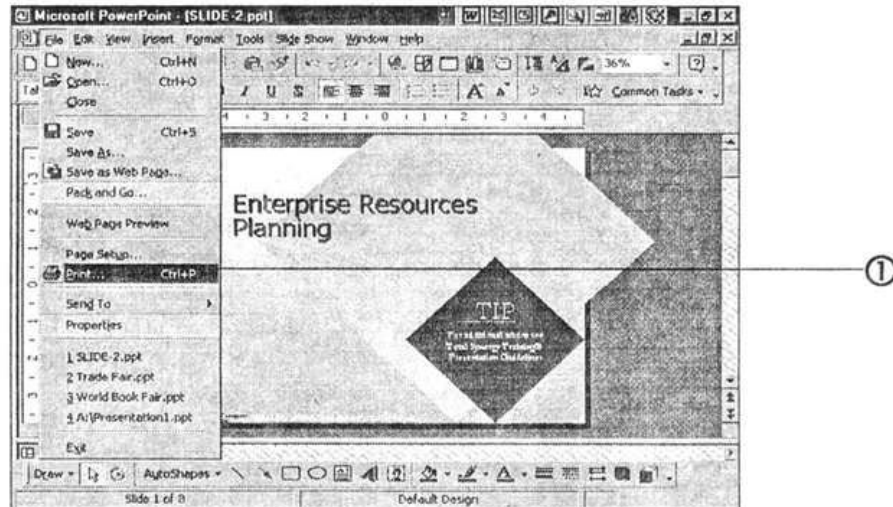
AutoSave

- 1 Choose **Options** command from **Tools** menu.
- 2 The following dialog box would be displayed. Click once on **Save** folio/tab.
- 3 Click on the scroll buttons to increase or decrease the **AutoSave** recover time.
- 4 Click on **OK** button to continue.



Printing a Presentation

- 1 Choose **Print** command from **File** menu.
- 2 Click on the roll-down list and choose your printer.
- 3 Click on the **Properties** button to choose whether you want in **Best**, **Normal** or **Eco fast** quality print.
- 4 Click on **OK** button.



18.4 WORKING WITH POWERPOINT

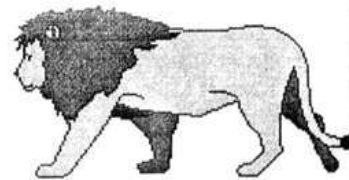
Now that you have got a general idea about PowerPoint let us get down to brass tracks. We will create a presentation in PowerPoint to see how simple it really is. For this purpose, I have deliberately taken a slightly longer route to create this presentation and have even made the user do things, which PowerPoint can very well do itself. The intention behind this has been to demystify the PowerPoint features and show how you can customize templates and wizards provided by PowerPoint and get the exact look that you want.

You have already seen the various parts of the PowerPoint in the previous screen. Also, it is important to familiarize yourself with the PowerPoint standard toolbar, the formatting toolbar, the drawing toolbar, and the auto shapes toolbar. So, if you have skipped that section please go back and take a look there.

Please take a look at the sample exercise pages (ten slides), so that you know what presentation you have to create and then follow the steps mentioned in the following pages.

So, let's begin now!

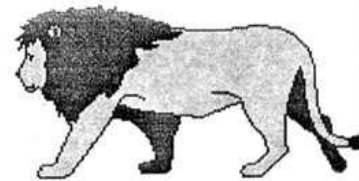
Action Plan
for organising the
Fourth National Games



Mysore City Corporation
31 May – 11 June '99

Slide 1

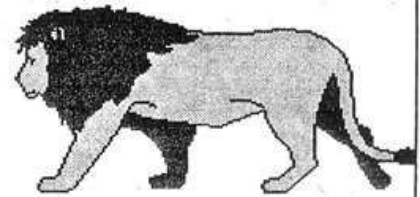
Welcome to the Steering
Committee Meeting for the
National Games !



Mysore City Corporation

Slide 2

Action Plan



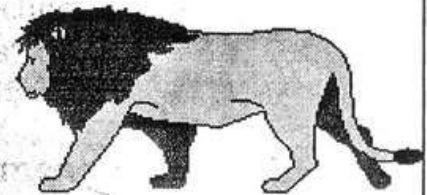
- Mission Statement
- Time Schedule
- Venue Schedule
- Organising Committee
- Responsibility Chart
- Success Indicators

Mysore City Corporation



Slide 3

Mission Statement



3.3.1 Preference Hypothesis: The fundamental

The methodological approach to the fundamental hypothesis (theory) may be summarised as follows:

First, it is hypothesised that the demand for a good is affected by the price of the good itself, the price of other goods, and the income of the consumer.

Second, it is hypothesised that the demand for a good is affected by the price of the good itself, the price of other goods, and the income of the consumer.

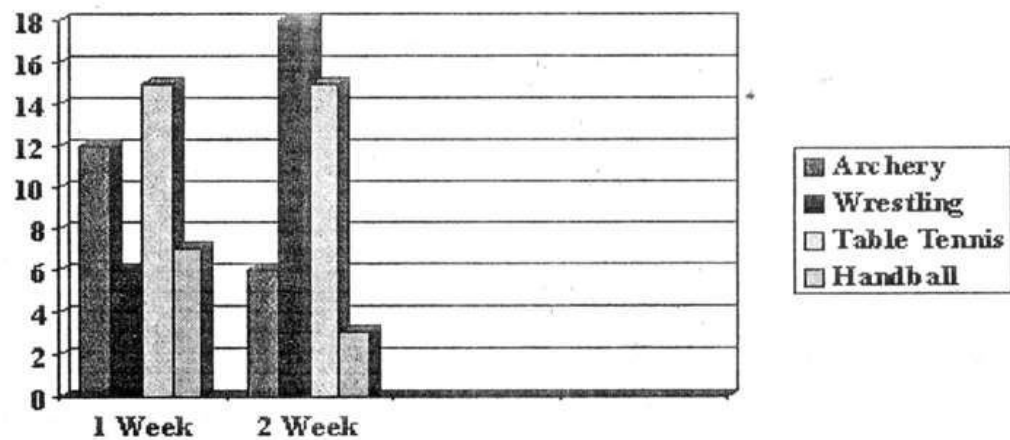
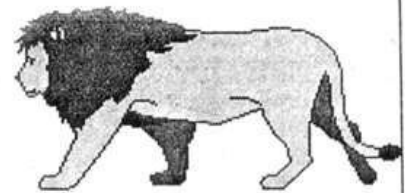
Not only for the people, but also for
organisations officials!

We shall cross all hurdles !!!

Mysore City Corporation

Slide 4

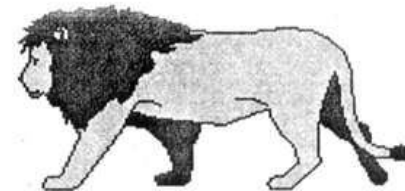
Time Schedule



Mysore City Corporation

Slide 5

Venue Schedule

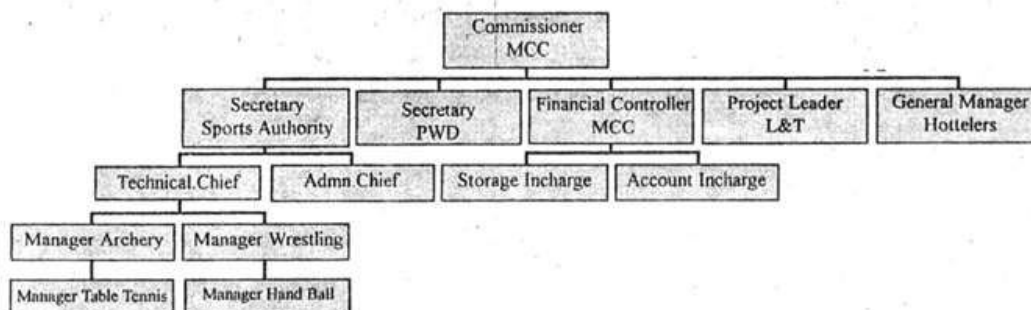
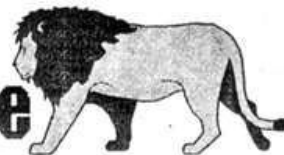


- Archery
- Wrestling
- Table Tennis
- Handball
- Chamundi Stadium
- Channaiah Akhada
- Chamundi Stadium
- Chamundi Stadium

Mysore City Corporation

Slide 6

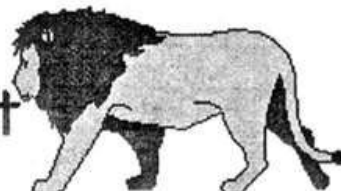
Organizing Committee



Mysore City Corporation

Slide 7

Responsibility Chart



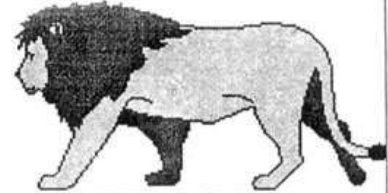
Commissioner	Overall Incharge
Sports Secretary	Incharge for all sports related matters
Technical Chief	All Sports' technical matters
Manager - Archery	Incharge of Archery events
Manager - Wrestling	Incharge of Wrestling events
Manager - Table Tennis	Incharge of Table Tennis events
Manager - Handball	Incharge of Handball events



Mysore City Corporation

Slide 8

Success Indicators

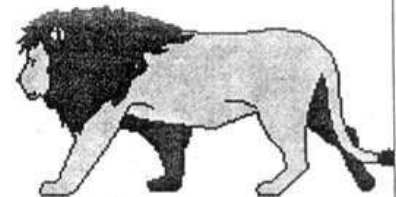


- TIMELY START AND COMPLETION ALL EVENTS
- MAXIMISE SPECTATORS
- PROPER CROWD CONTROL
- PROPER LAW AND ORDER
- PROPER HOSPITALITY TO ALL PARTICIPANTS
- PROPER MEDIA COVERAGE

Mysore City Corporation

Slide 9

BEST OF LUCK
&
GET CRACKING !



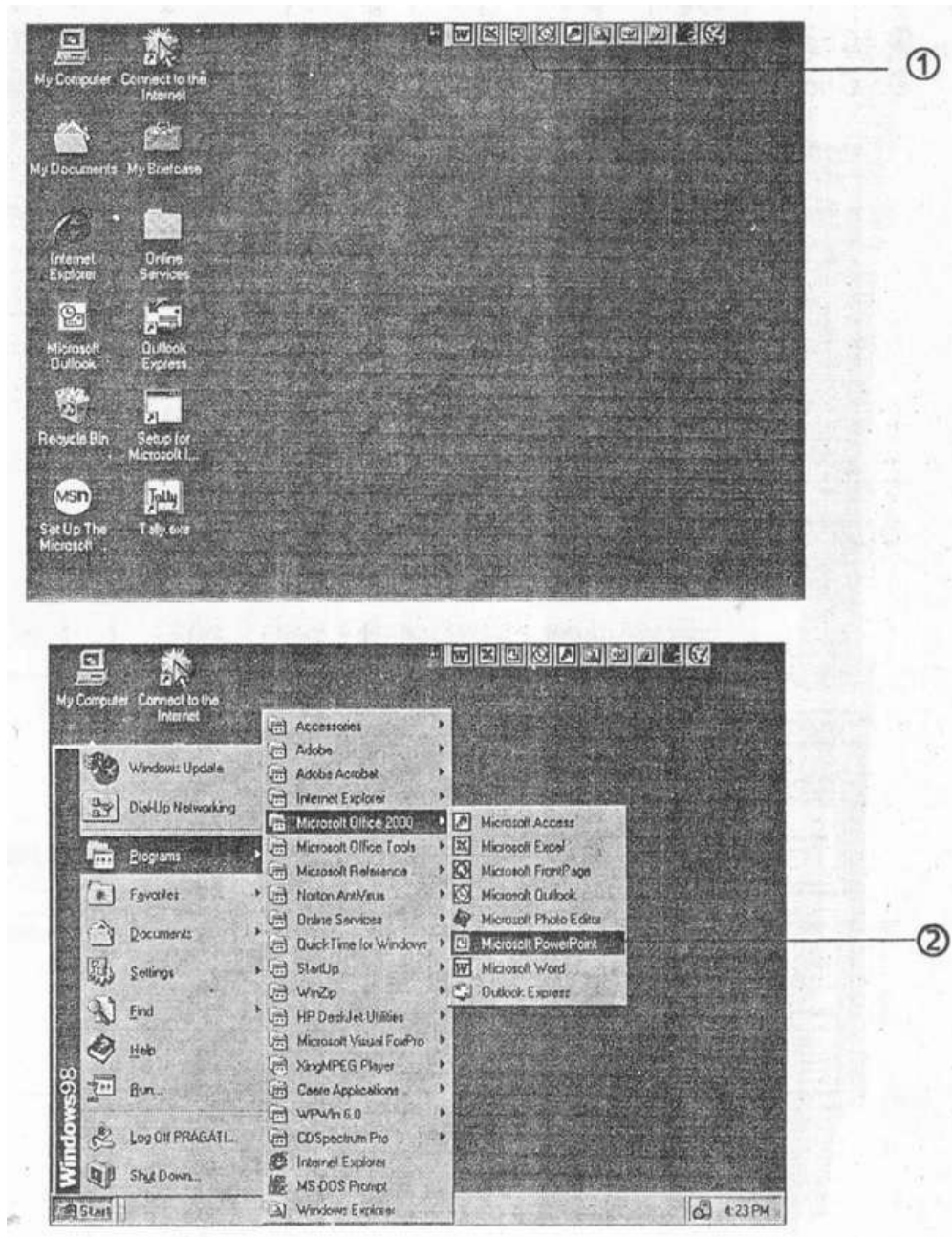
Remember, ALL'S WELL THAT ENDS WELL!

Mysore City Corporation

Slide 10

Starting PowerPoint

- 1 Click here to start **PowerPoint** or
- 2 Choose **Microsoft PowerPoint** from **Microsoft Office 2000** from **Programs** menu.



First Screen

This is the first screen that appears on starting PowerPoint. Let me briefly explain about the option here.

- A This is the quickest way to create a presentation by choosing from predefined subjects and templates.

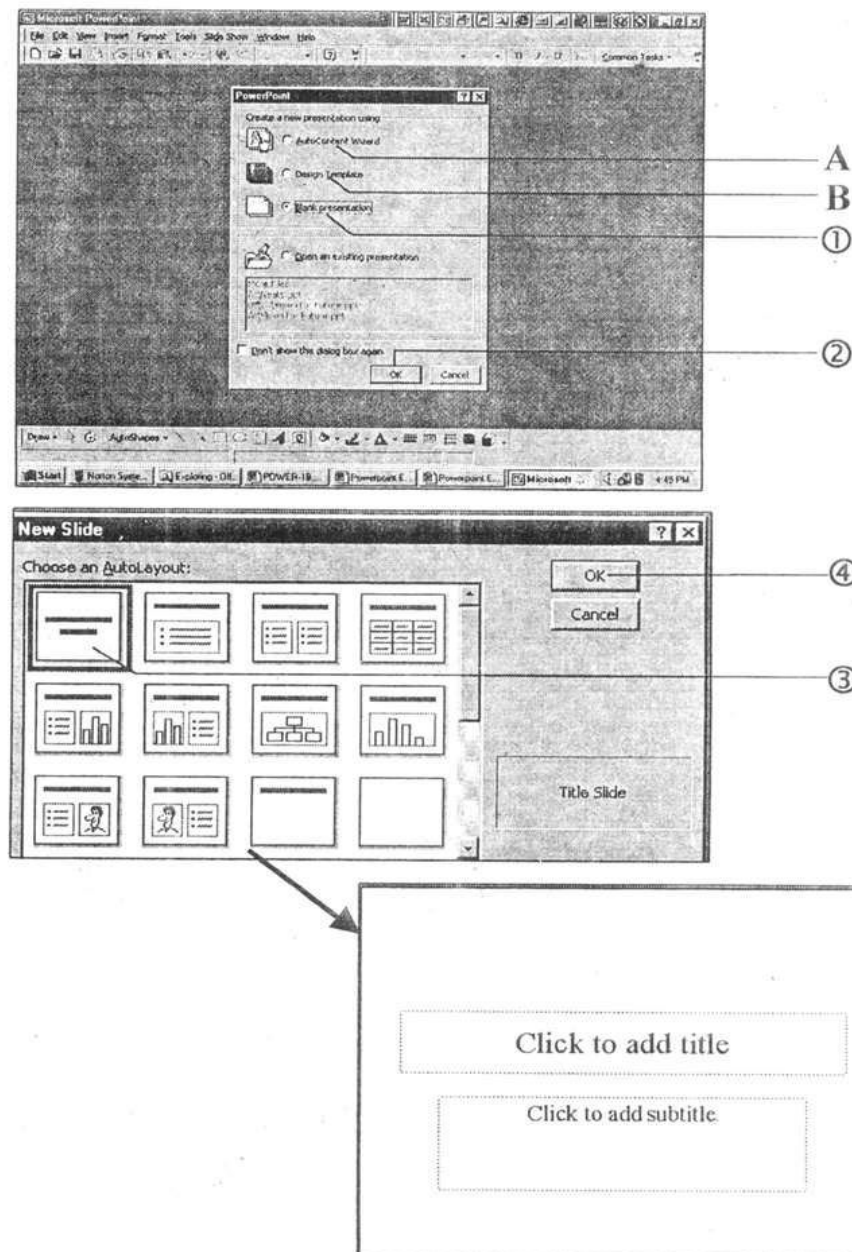
B In case you simply wish to use predefined templates containing colour schemes and background, etc. choose this option.

1 To create a presentation from scratch, click here.

2 Click on **OK** button once to continue.

3 Choose the **Title Slide** from **Auto Layout** box.

4 Click on **OK** button to continue.



CHECK YOUR PROGRESS-I

MULTIPLE CHOICE QUESTIONS:

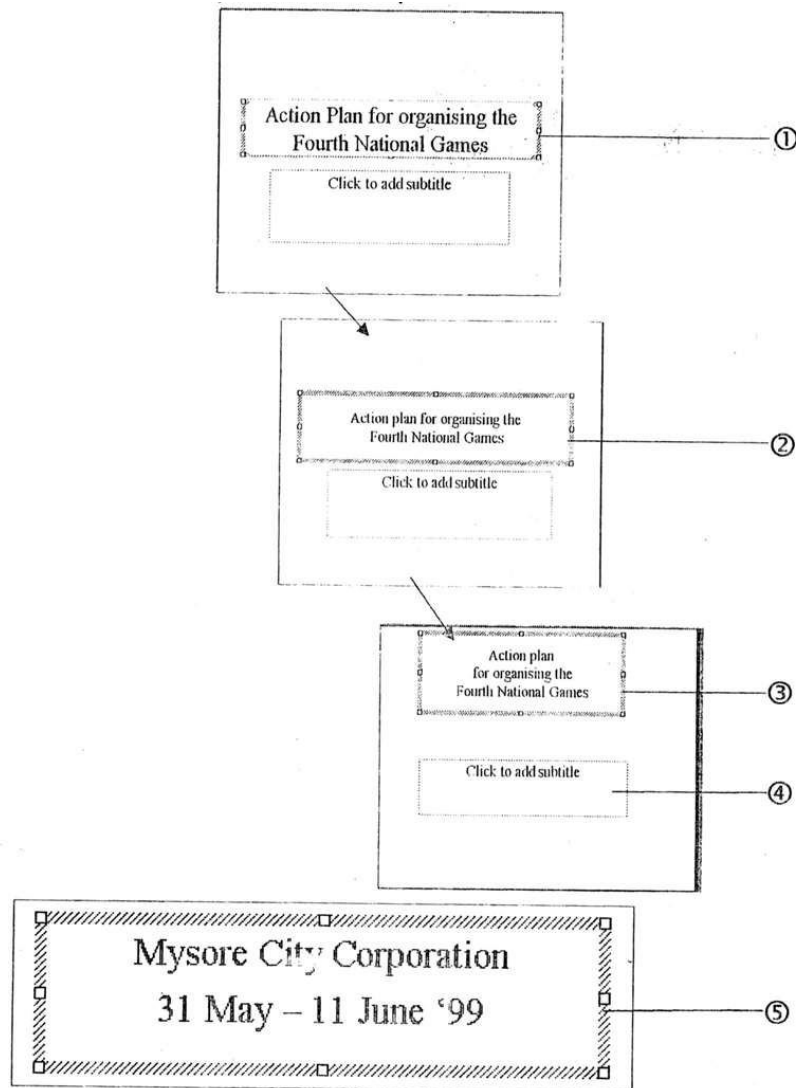
- 1. In MS PowerPoint presentations, the designs regulate the layout and formatting for the slide. These are commonly known as:**
 - a. Blueprints
 - b. Placeholders
 - c. Templates
 - d. Design Plates
- 2. The Handout Master consists of placeholders for all of these except the:**
 - a. title
 - b. slide number
 - c. header
 - d. footer
- 3. Which of these PowerPoint features would allow any user to create a given simple presentation quicker?**
 - a. Animations
 - b. Chart Wizard
 - c. Transition Wizard
 - d. AutoContent Wizard
- 4. The File that consists of readymade styles that we can easily use for our presentation is known as:**
 - a. Pre formatting
 - b. Template
 - c. Wizard
 - d. AutoStyle
- 5. What is the max Zoom percentage in MS PowerPoint?**
 - a. 400%
 - b. 300%
 - c. 200%
- 6. A PowerPoint design template is stored in any file with _____ extensions.**
 - a. .pot

- b. .ptt
 - c. .dtp
 - d. .ppt
- 7. Which of these options in PowerPoint would allow us to carry any set of slides from a given computer to another one?**
- a. Web and Go
 - b. Pack and Go
 - c. Save and Go
 - d. Save As
- 8. Which of these is a type of sound file?**
- a. DAT files
 - b. LOG files
 - c. DRV files
 - d. WAV files
- 9. In the current presentation, if we want to insert a new slide, we can choose which of these?**
- a. Ctrl + F
 - b. Ctrl + O
 - c. Ctrl + M
 - d. Ctrl + N
- 10. Which of these terms do we use when an image in a clip art changes the direction that it was originally facing?**
- a. Mirror
 - b. Rotate
 - c. Flip
 - d. Group
- 11. Which of these refers to the PowerPoint view used for displaying only the text (title and bullets)?**
- a. Slide sorter show
 - b. Slide view
 - c. Outline view
 - d. Notes page view

- 12. Which toolbar provides various options in different master views?**
- a. Standard toolbar
 - b. Formatting toolbar
 - c. Drawing toolbar
 - d. Common tasks toolbar
- 13. We can find any and every command on the:**
- a. Menu bar
 - b. Standard toolbar
 - c. Formatting toolbar
 - d. Drawing toolbar
- 14. Which of these features should one use when we want to type in the text box for notes?**
- a. Insert
 - b. Slide shows
 - c. Zoom
 - d. Slide master
- 15. Which of these is the page setup orientation by default of the slides in PowerPoint?**
- a. Landscape
 - b. Portrait
 - c. Vertical
 - d. None of the above

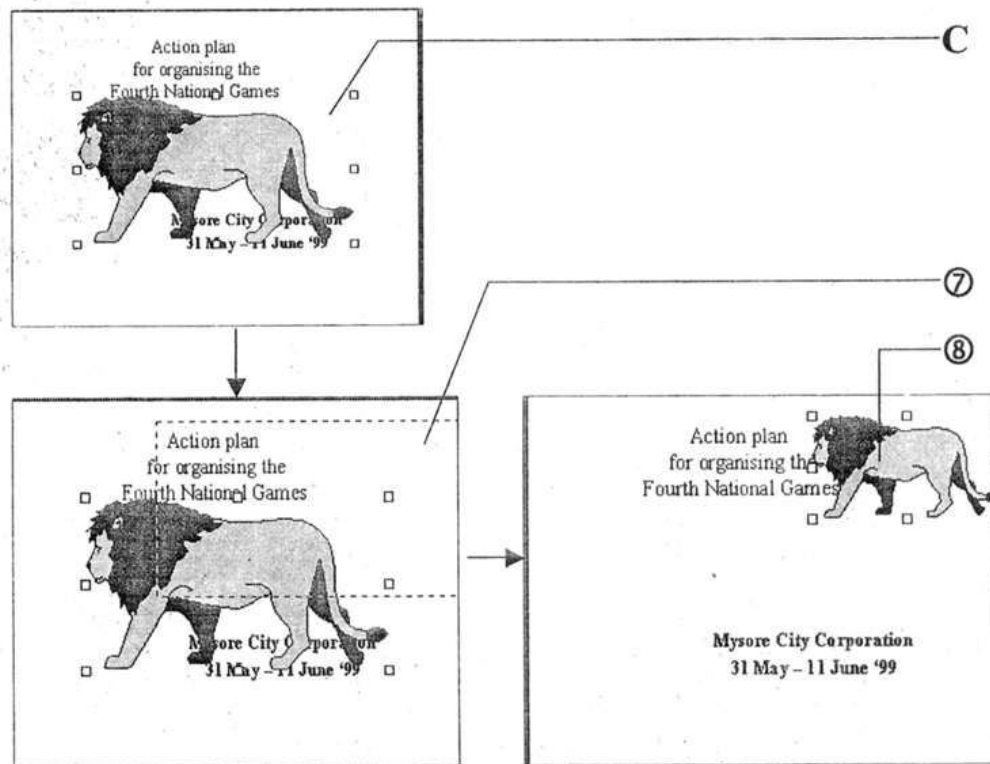
Creating Front, Font Size and Bold

- 1 Type the presentation title.
- 2 Click on the text box boundary to select the entire text box.
- 3 Drag the box to the top of the slide and position at the desired location. You can also resize this box, as shown here.
- 4 Click here to continue.
- 5 Type the desired text.



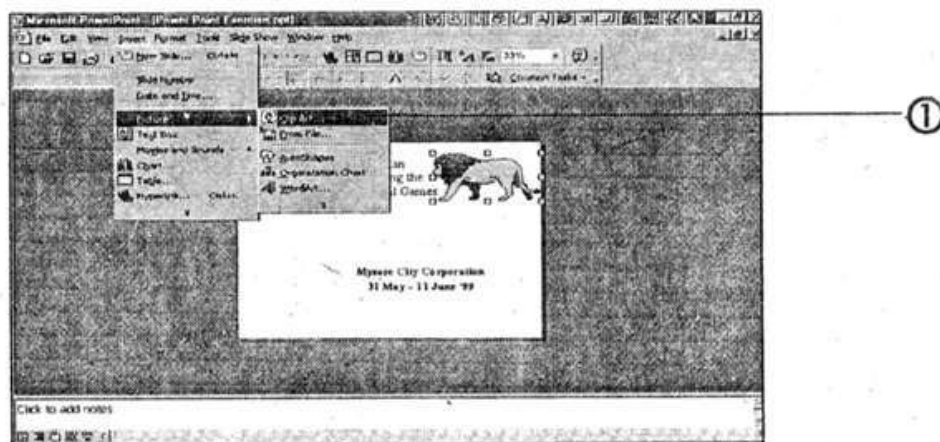
Moving the Frame and Inserting ClipArt

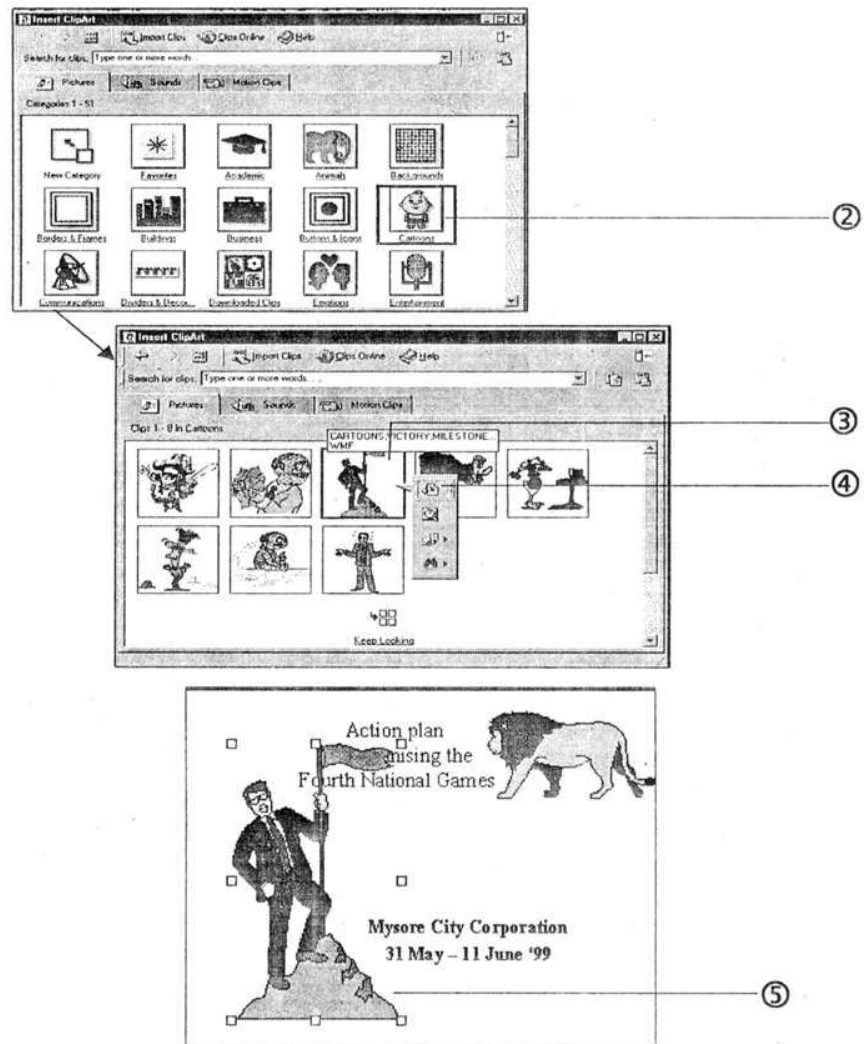
- 1 Click here to increase Font size.
 - 2 Click here to decrease Font size.
 - 3 Click here to apply shadow effect to the text.
 - 4 Choose ClipArt command from Picture option of Insert menu.
- A Highlight the text and format the text to your liking. You can change the Font, Point size, Font style and Alignment. You will notice that the formatting toolbar is very similar to Word or Excel.
- B Once having applied the desired formatting command, drag and move the box towards the right-hand side.
- 5 Click here to choose Animals Category.



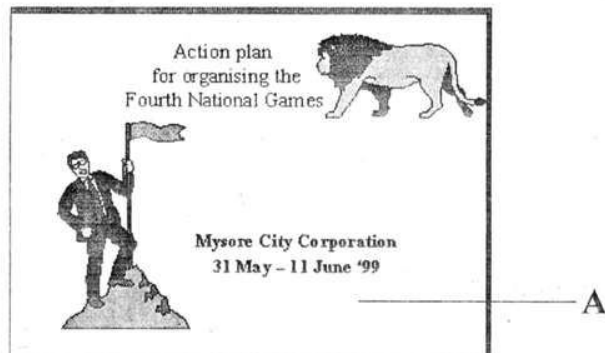
Inserting Picture

- 1 Choose again **Clip Art** command from Picture option of Insert menu.
- 2 Choose 'Cartoons' Category
- 3 Choose the picture.
4. Click here to insert picture.
- 5 Use the technique explained earlier to move the picture to the right position and resize it as per your requirement.





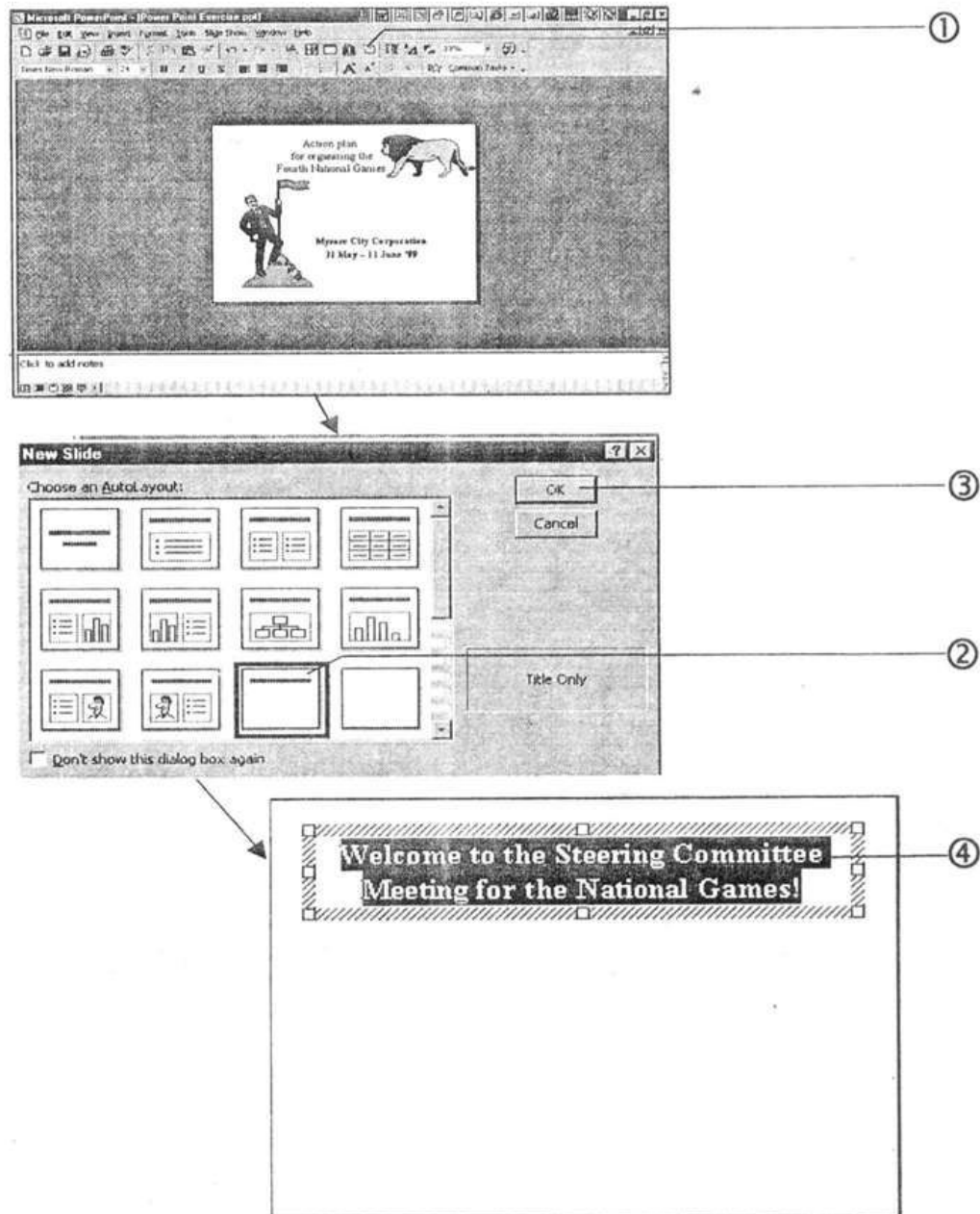
Congrats! You Have Just Finished Your First Slide



A Inserting a New Slide

- 1 Click here to create a new slide.
- 2 Choose **Title Only** from **Auto layout**.

- 3 Click on the **OK** button once to continue.
- 4 Type the text and apply **Bold** attributes.

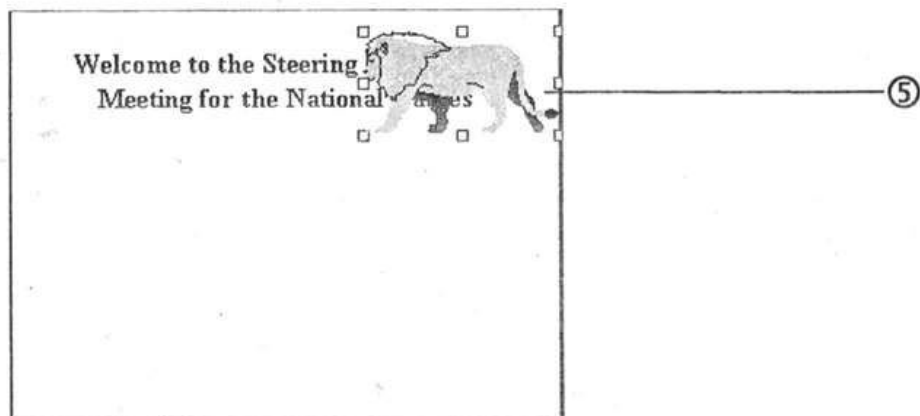
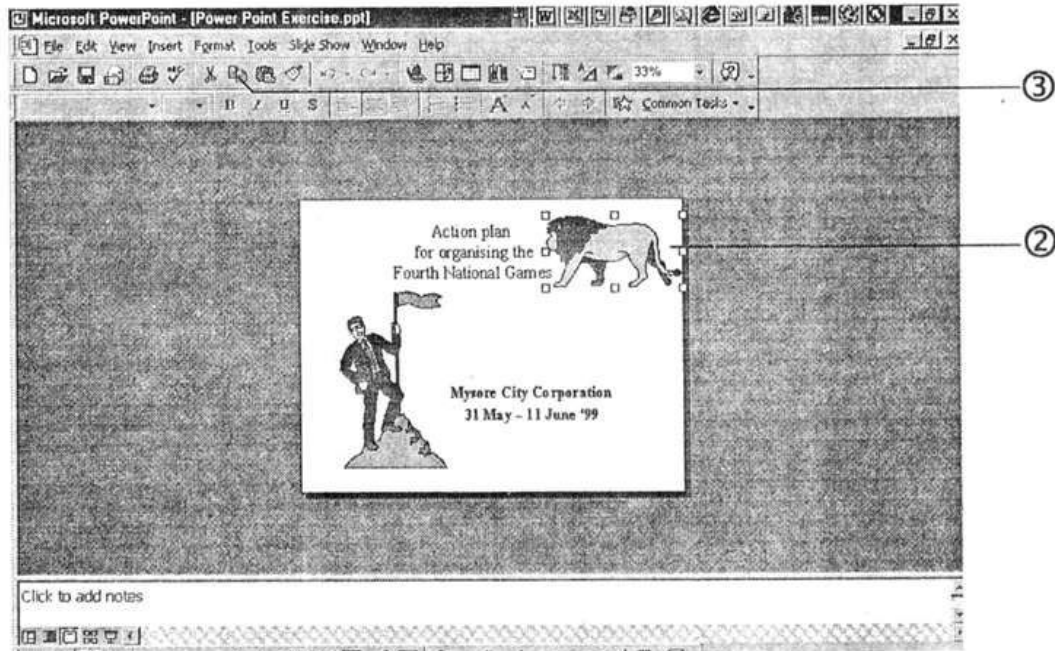


Copying Picture from Previous Slide

Now we wish to insert the Elephant picture (our logo) slide. We have two options. One, we use step for moving and inserting Clip Art, explained earlier to do so. The second and simpler would be to copy it from the previous slide. Doing so would be much faster.

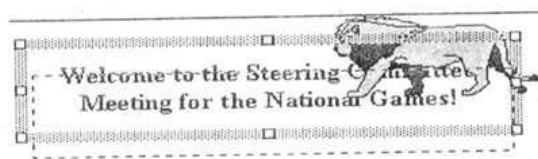
- 1 Press **Page Up** key to go the previous slide.

- 2 Click here to choose this picture box.
- 3 Click here to copy
- 4 Press **Page Down** key to go to the new slide.
- 5 Click on **Paste** icon and get the properly sized and placed picture.



Moving the Text

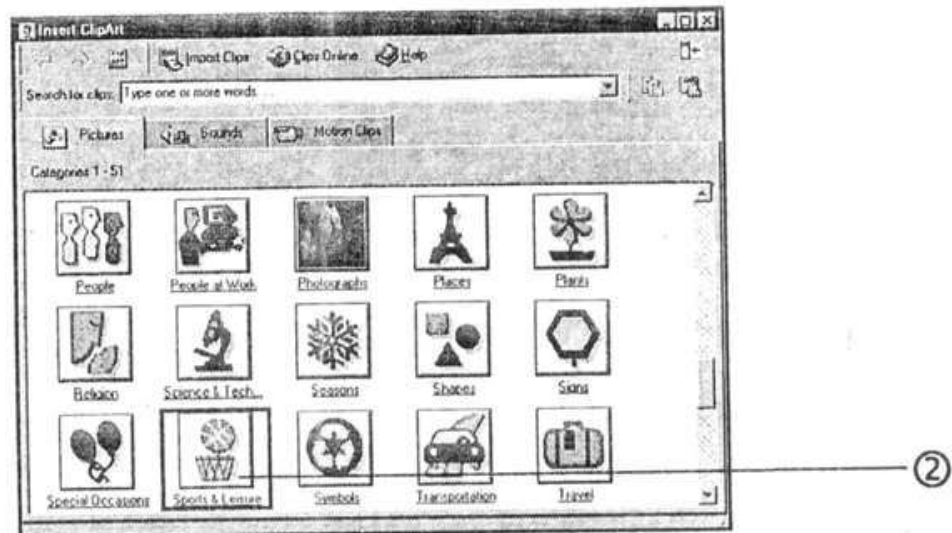
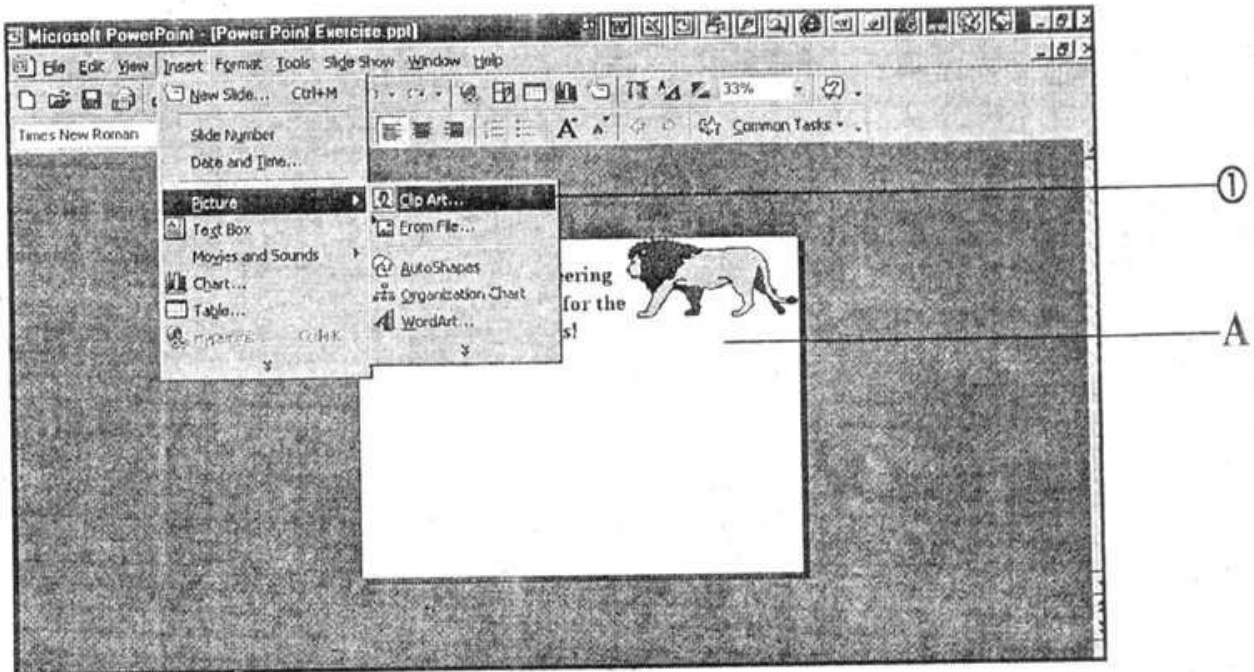
Now since the picture is overlapping on the text box we will have to resize our text box.



Inserting Picture

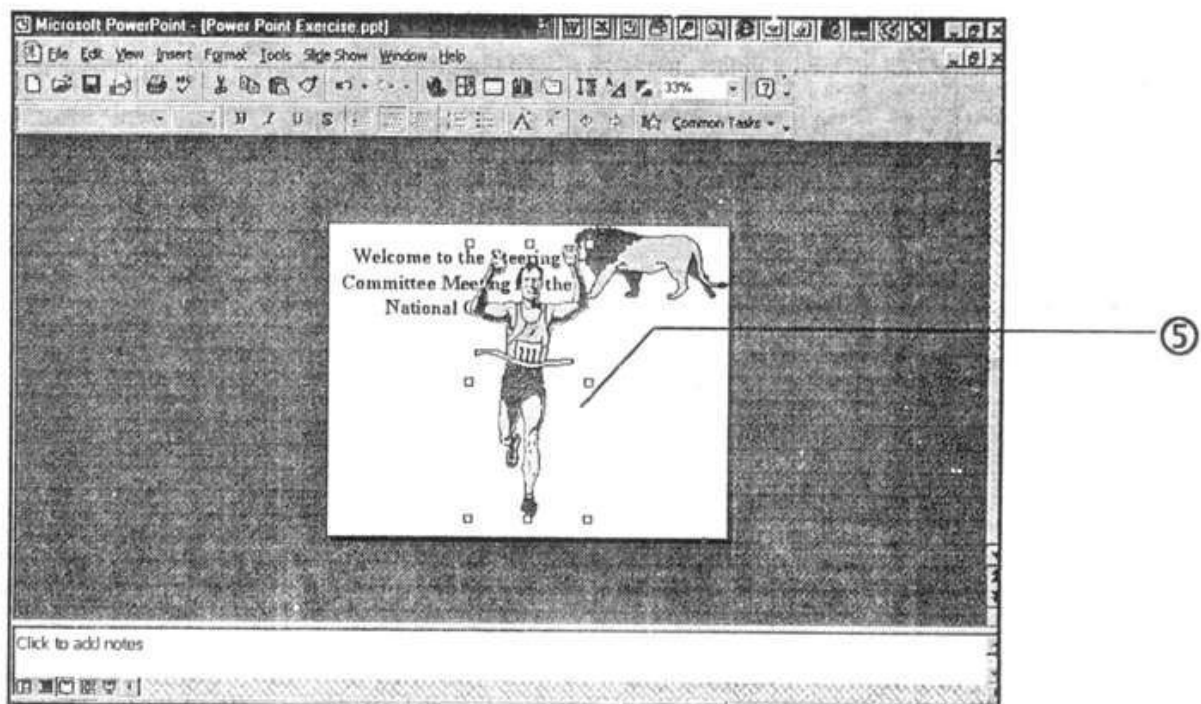
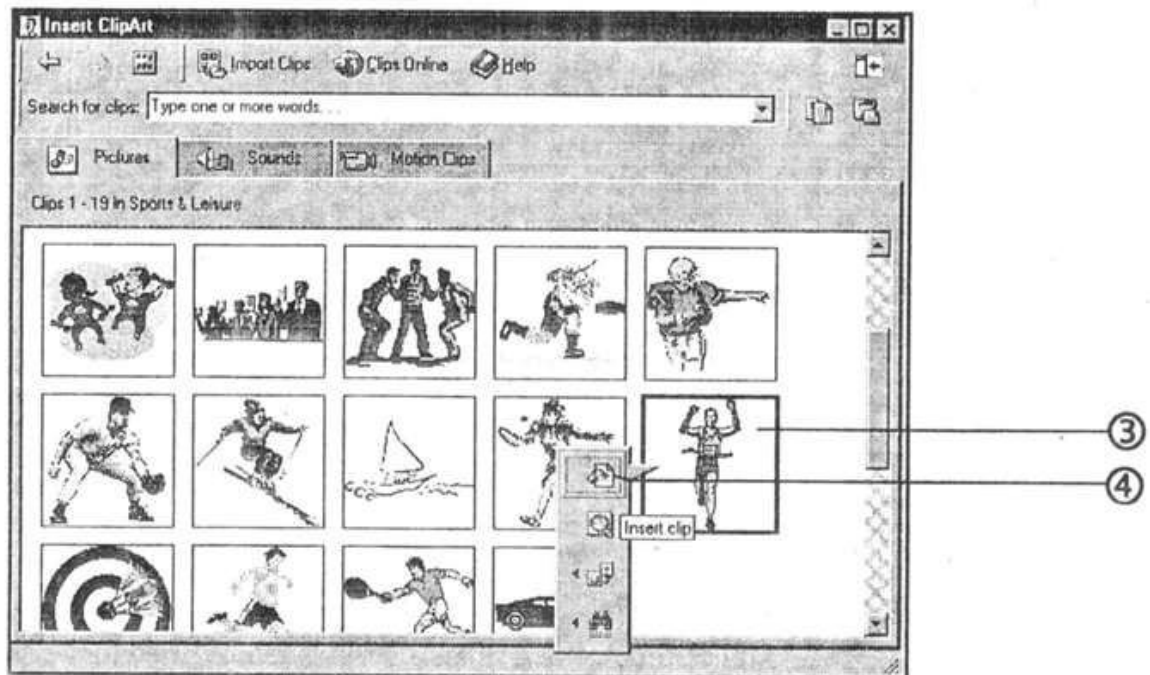
- 1 Choose **Clip Art** command from **Picture** option of **Insert** menu.
- 2 Choose '**Sports and Leisure**'

A Now both the picture and text fit properly.



- 3 Choose the picture.
- 4 Click to insert another picture.

- 5 Resize and drag to place the picture at the desired location
- 6 Now use the same technique to get all the other picture so that the slide looks like this.



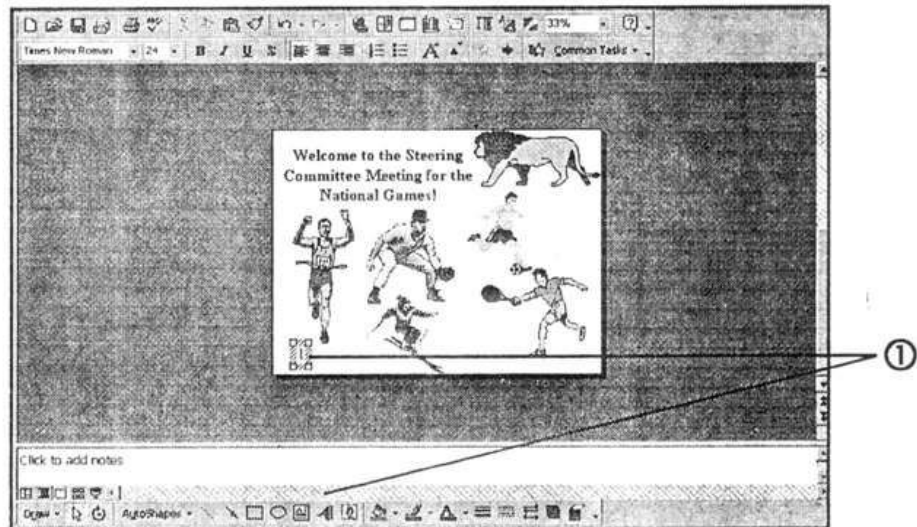


Inserting Text

- 1 Click on this icon. Place the mouse pointer here by clicking it once. A small box with a cursor in it would appear.
- 2 Type 'Mysore City Corporation' and click anywhere outside this box to conclude.

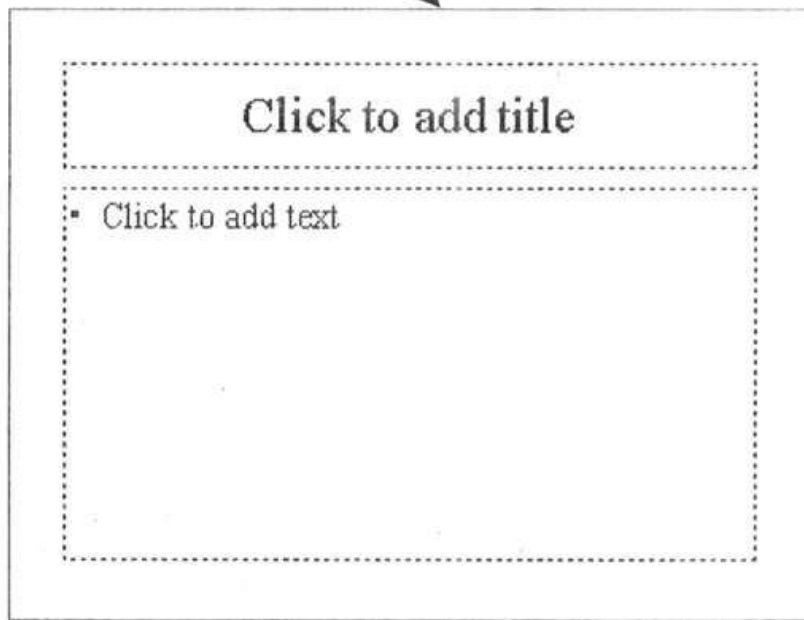
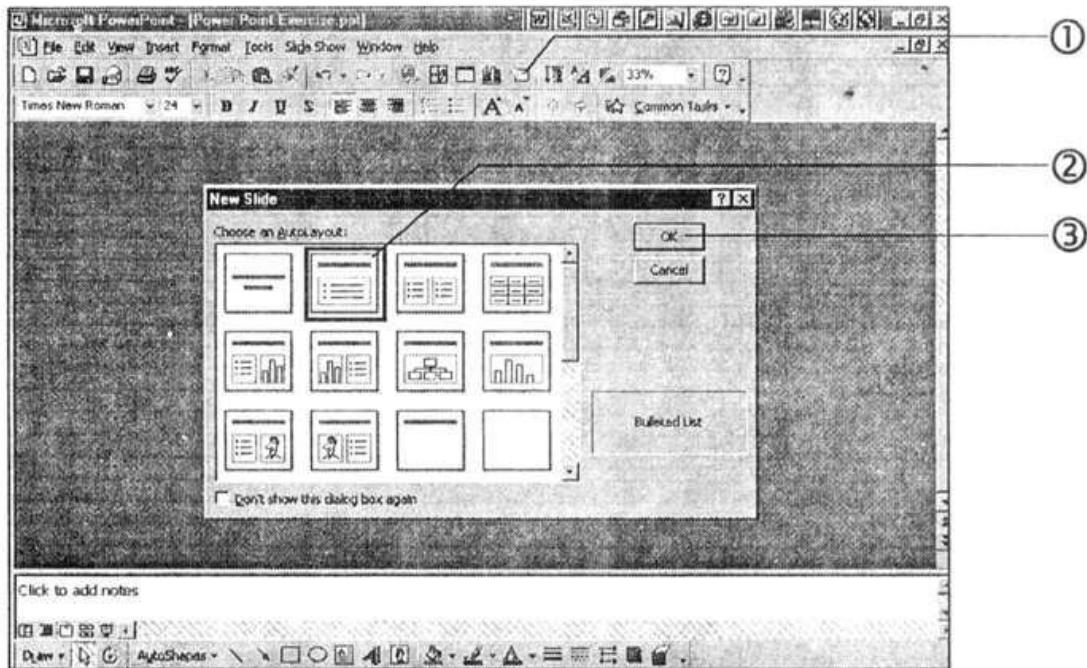


- A You can open the Drawing toolbar by clicking on **Drawing** command from **Toolbars** option from **View** menu.
- B The second slide is complete.



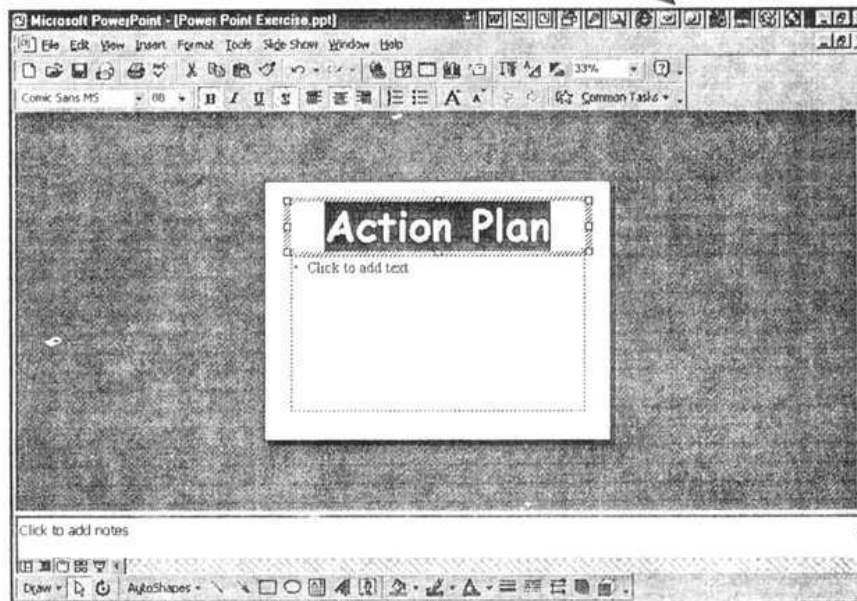
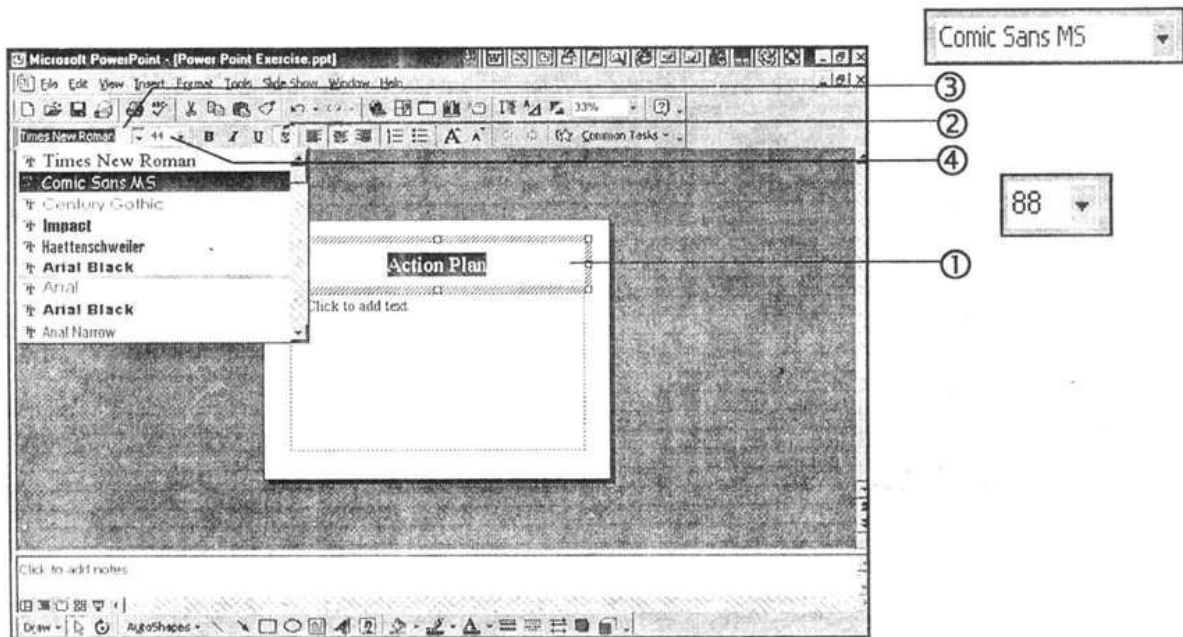
New Slide

- 1 Choose **New Slide** from the Status Bar.
- 2 Choose bulleted list from **Auto layout** box.
- 3 Click **OK to continue**.



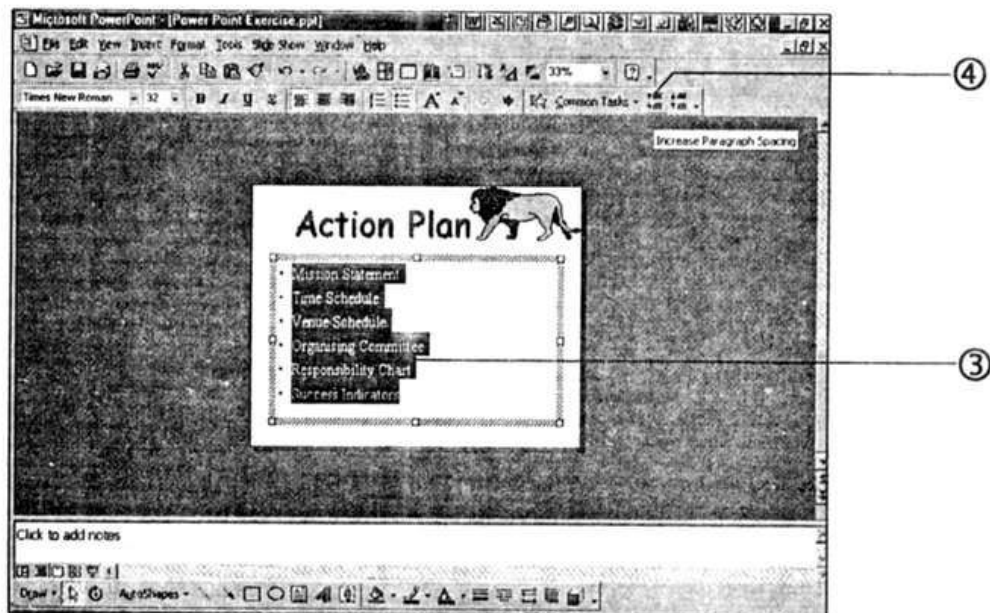
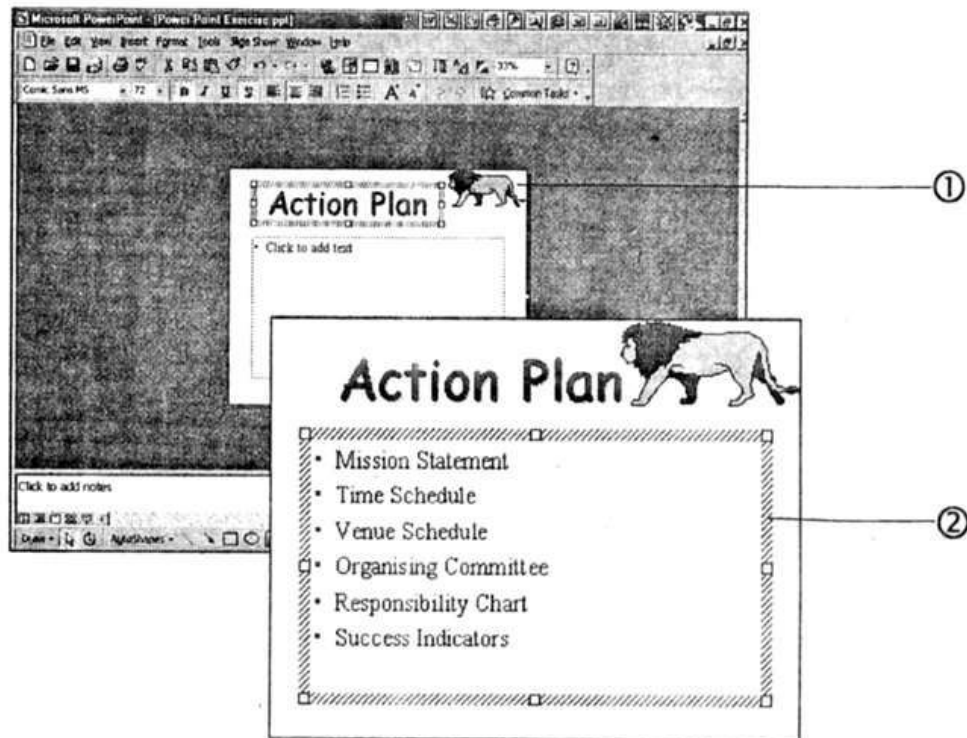
Changing the Font Size

- 1 Type the title and highlight it.
- 2 Click here to give a shadow effect.
- 3 Choose '**Comic Sans MS**' font.
- 4 Increase the size of the font.



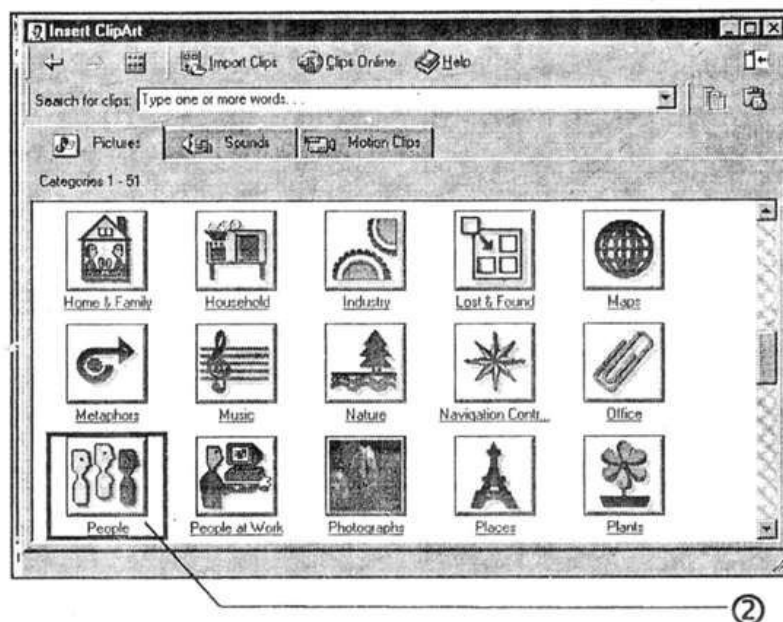
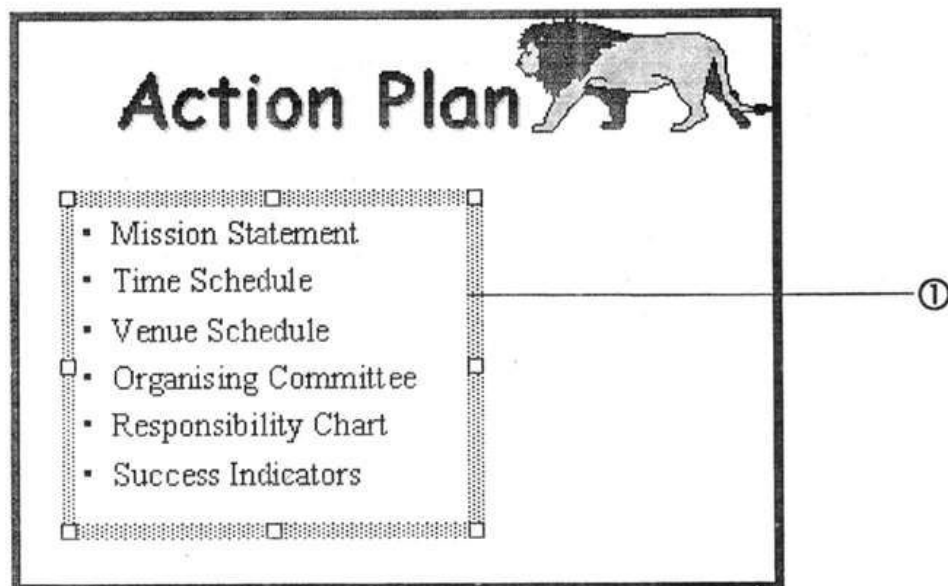
Copying Picture from Previous Slide

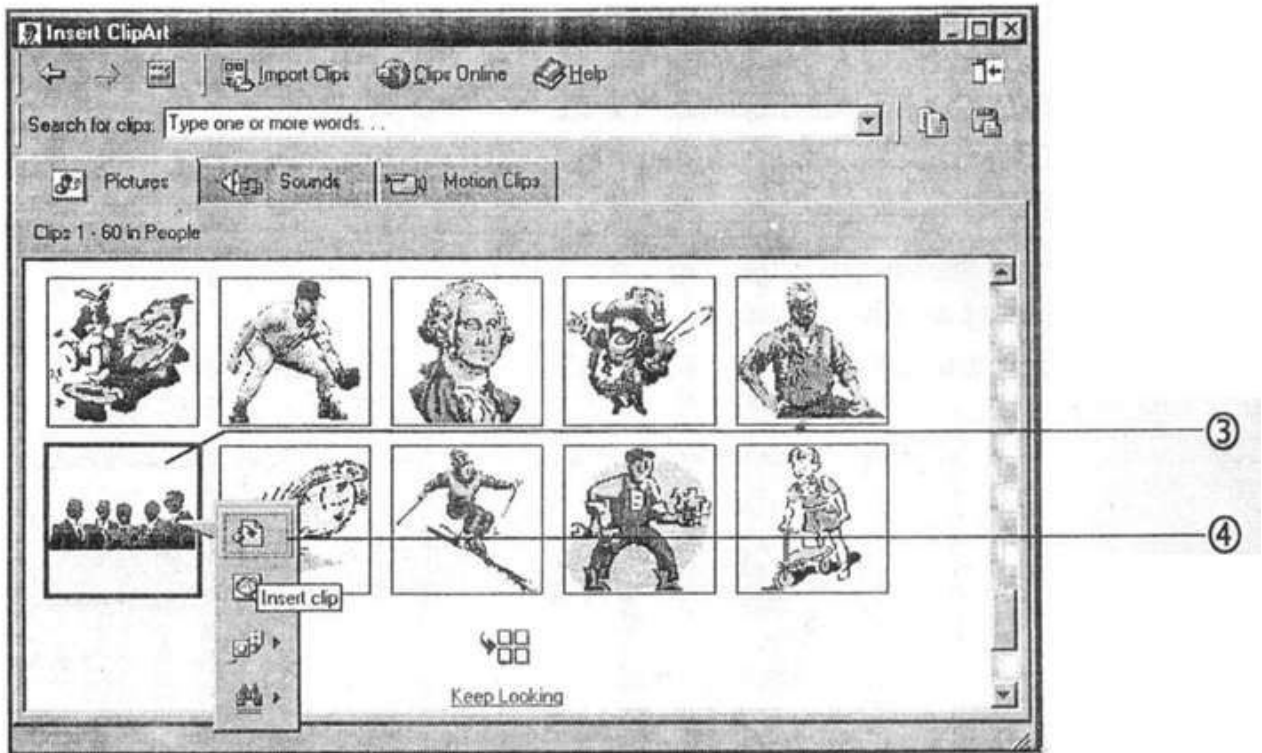
- 1 Copy the picture from previous slide.
- 2 Type the points
- 3 Highlight all the points
- 4 Click here repeatedly to increase the paragraph spacing to desired level.



Sizing Box and Inserting Picture

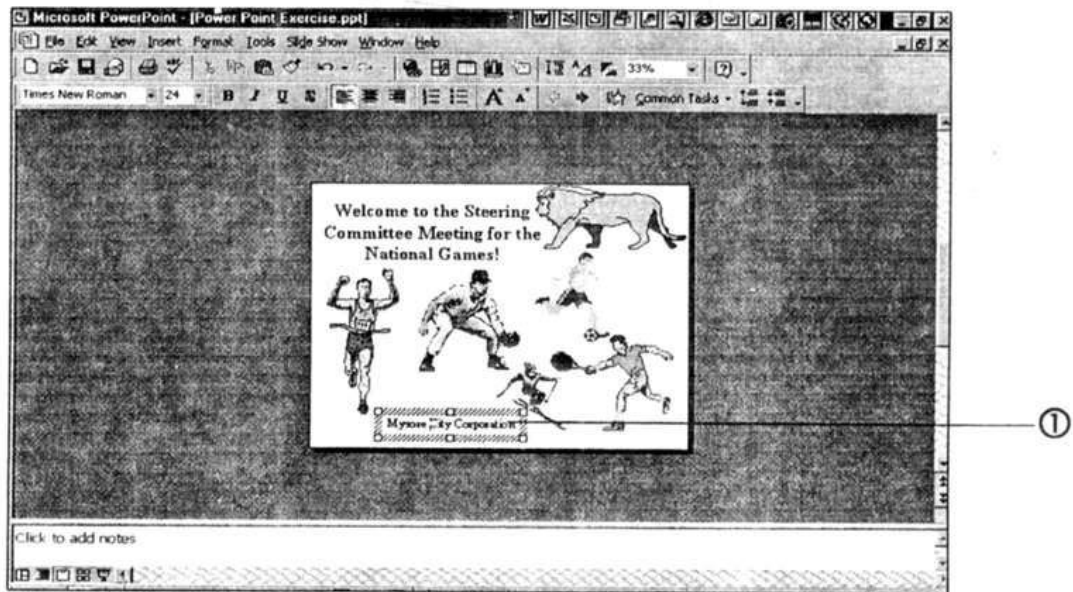
- 1 Size the box to the size of text.
- 2 Choose ClipArt command from Picture options of Insert menu. Choose People from category.
- 3 Choose the picture
- 4 Insert another picture.
- 5 Size the picture box to fit within the space available.



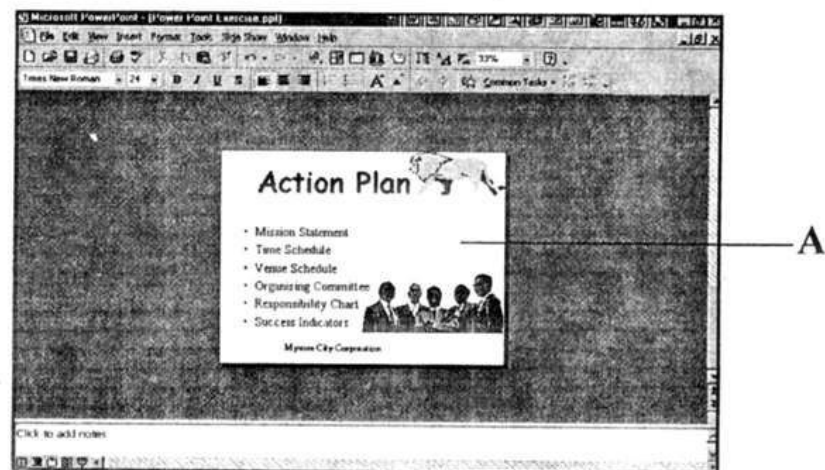


Copying Text from Previous Slide

- 1 Copy this text box from previous slide
- 2 Paste it here.

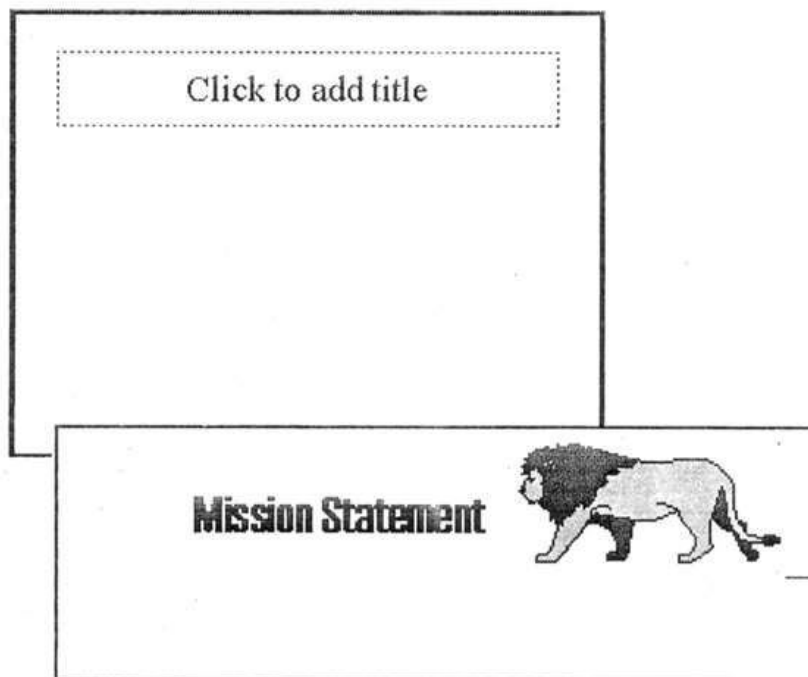
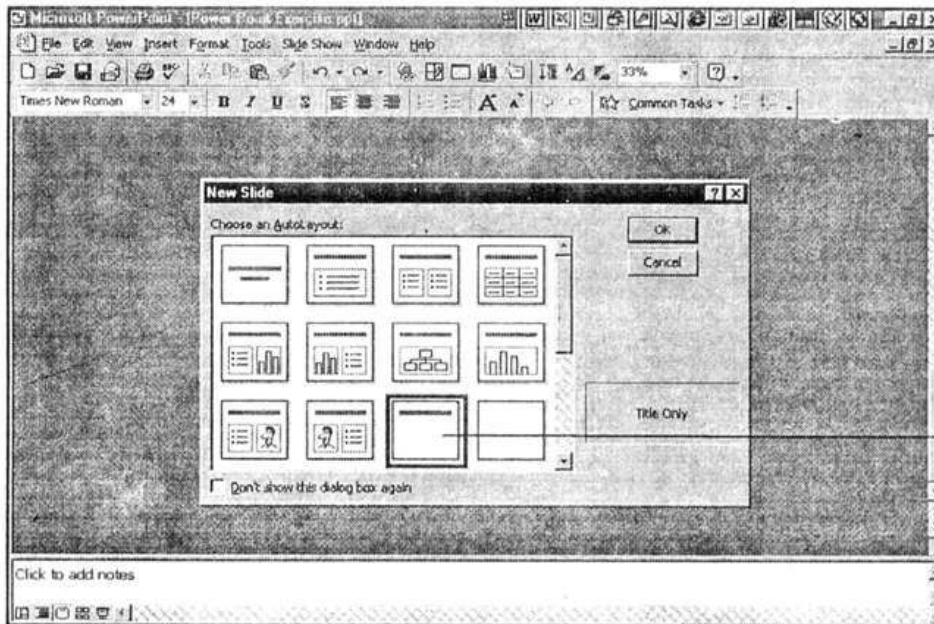


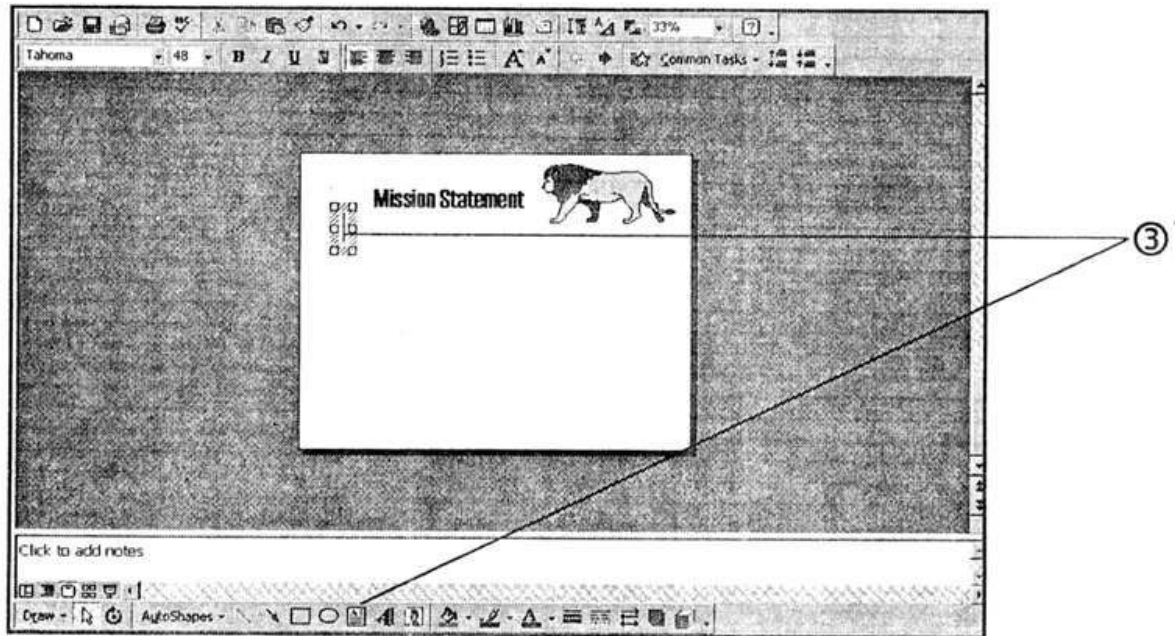
Your third slide is complete!



A New Slide

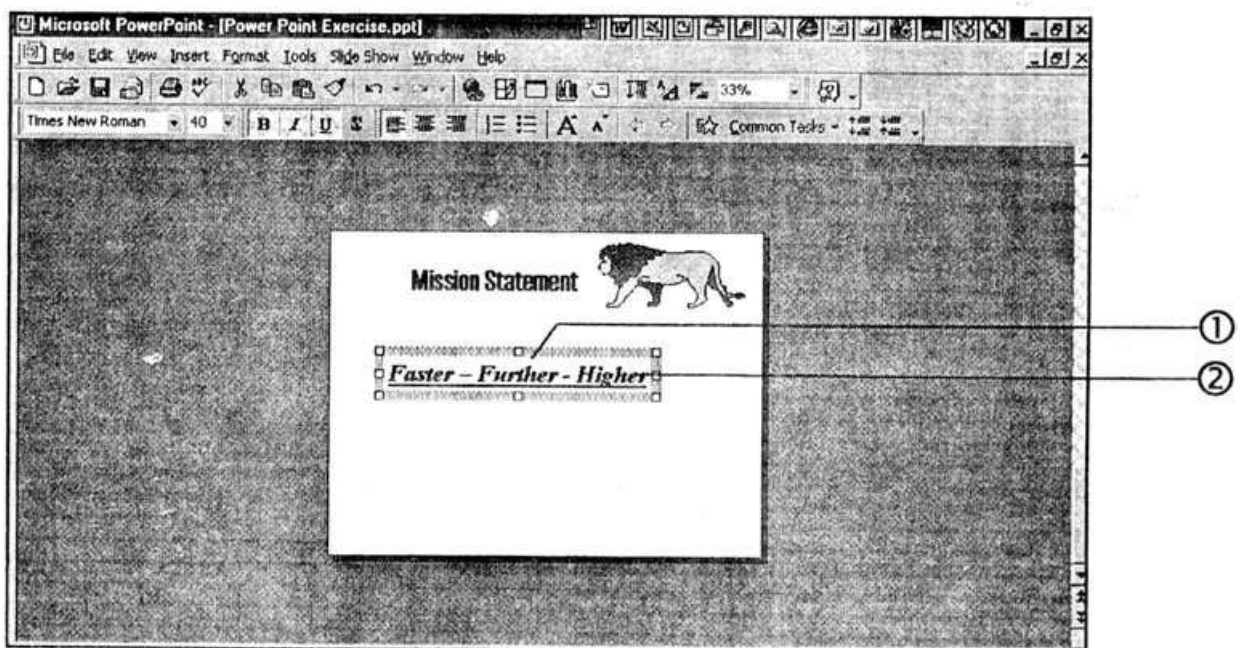
- 1 Add a **'Title only'** slide.
- 2 Type and format the text and copy the logo picture from previous slide.
- 3 Insert a text box.

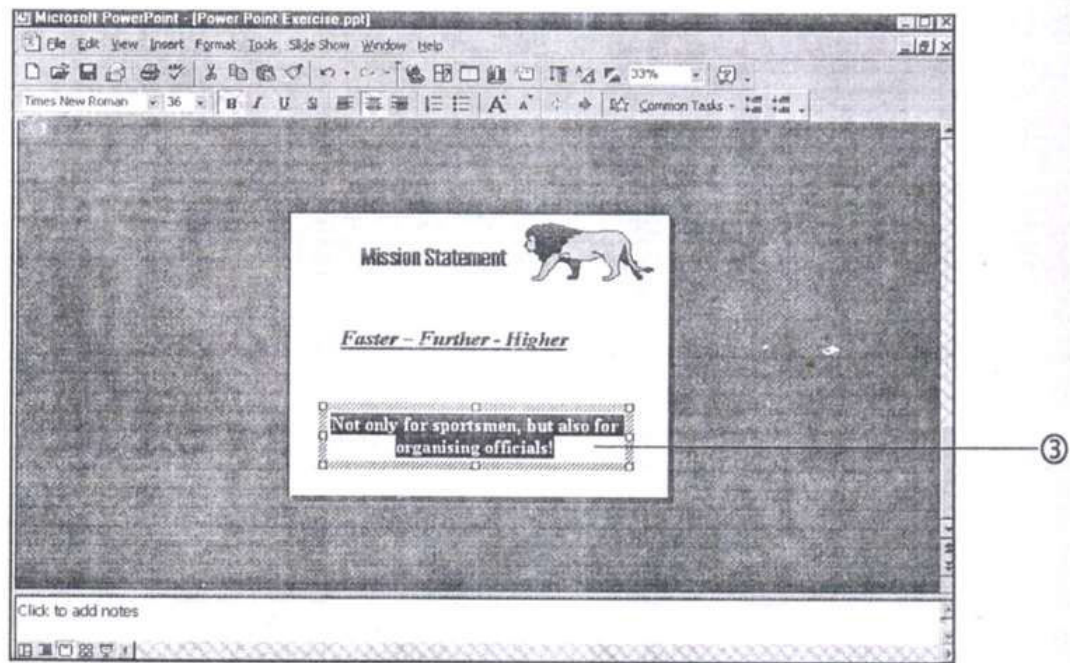




Text Styling

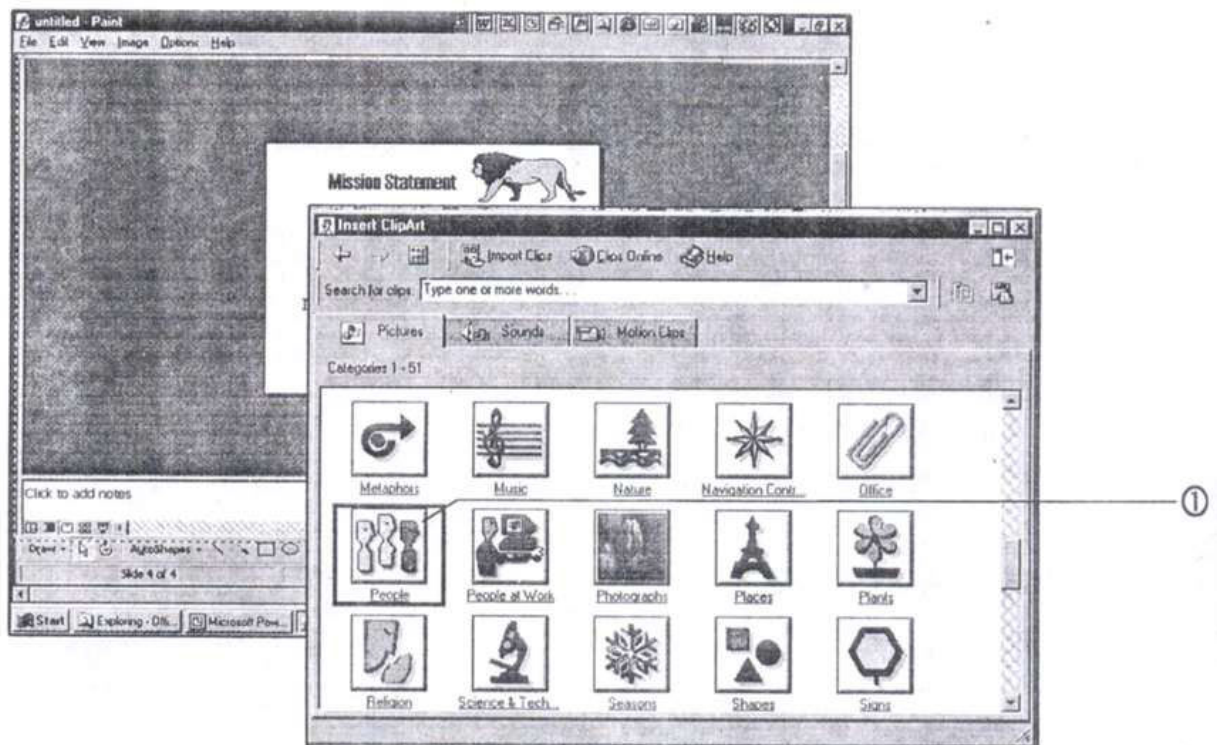
- 1 Type the motto. Increase the font size.
- 2 Apply the **Bold**, **Italics**, **Underline** attributes.
- 3 Insert another text box and type and format the text.





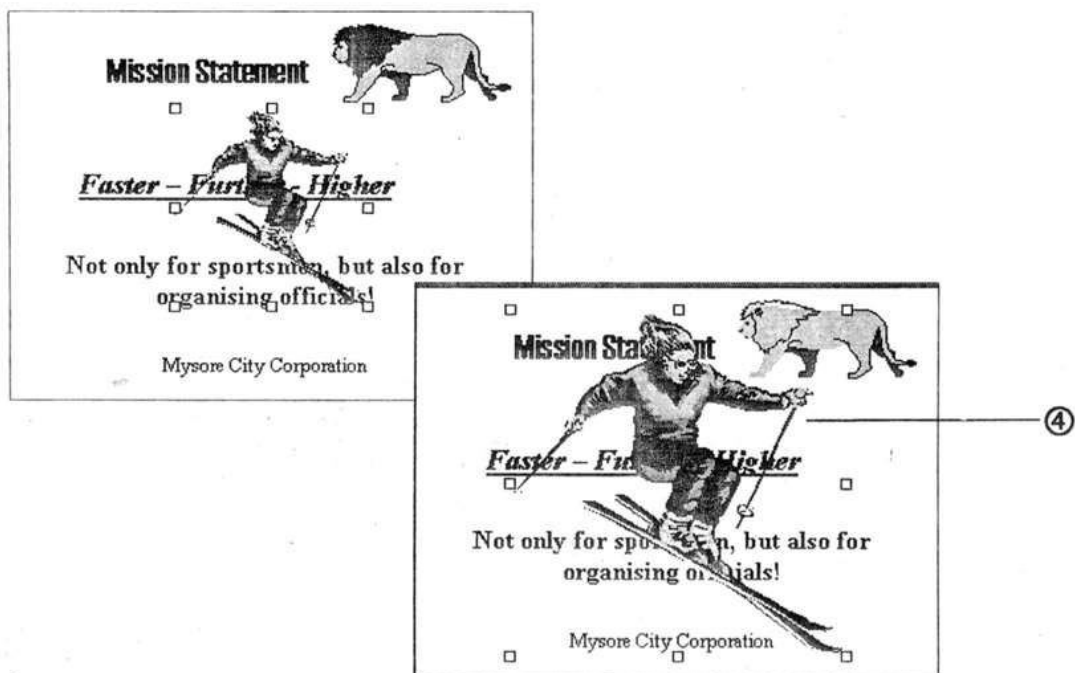
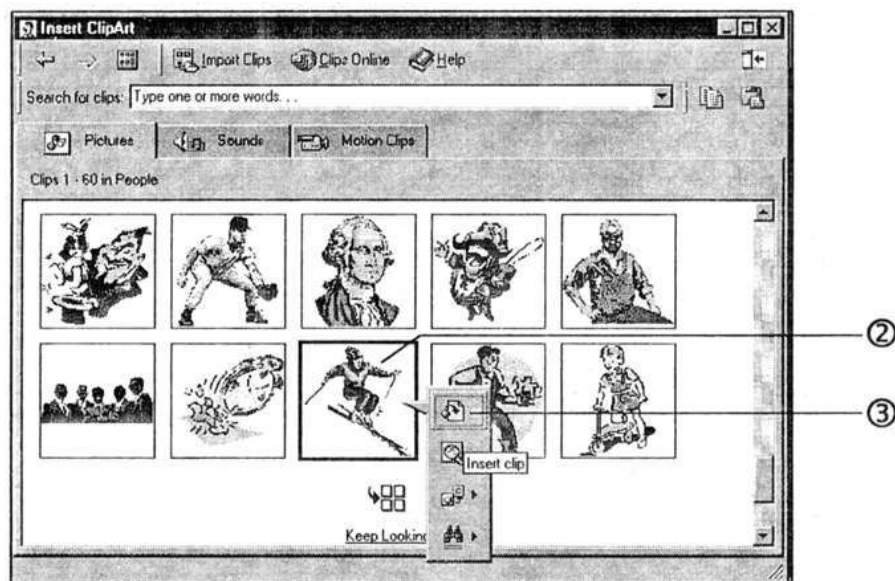
Insert Picture

- 1 Choose **Clip Art** command from **Picture** option of **Insert** menu. Choose **Sports** from Category.



- 2 Choose the picture.

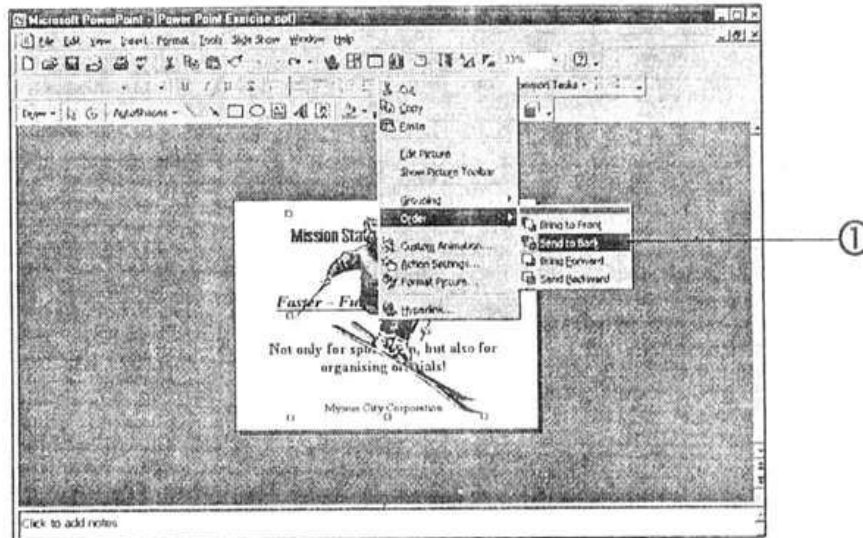
- 3 Insert another picture.
- 4 Size the picture to fit on the full slide.



Send to Back

You might have noticed that the picture is overlapping on your text and therefore, the entire text is not legible. Let us see how we can use a special effect to send this picture in the background, so that instead of picture coming on top, the text comes on the top.

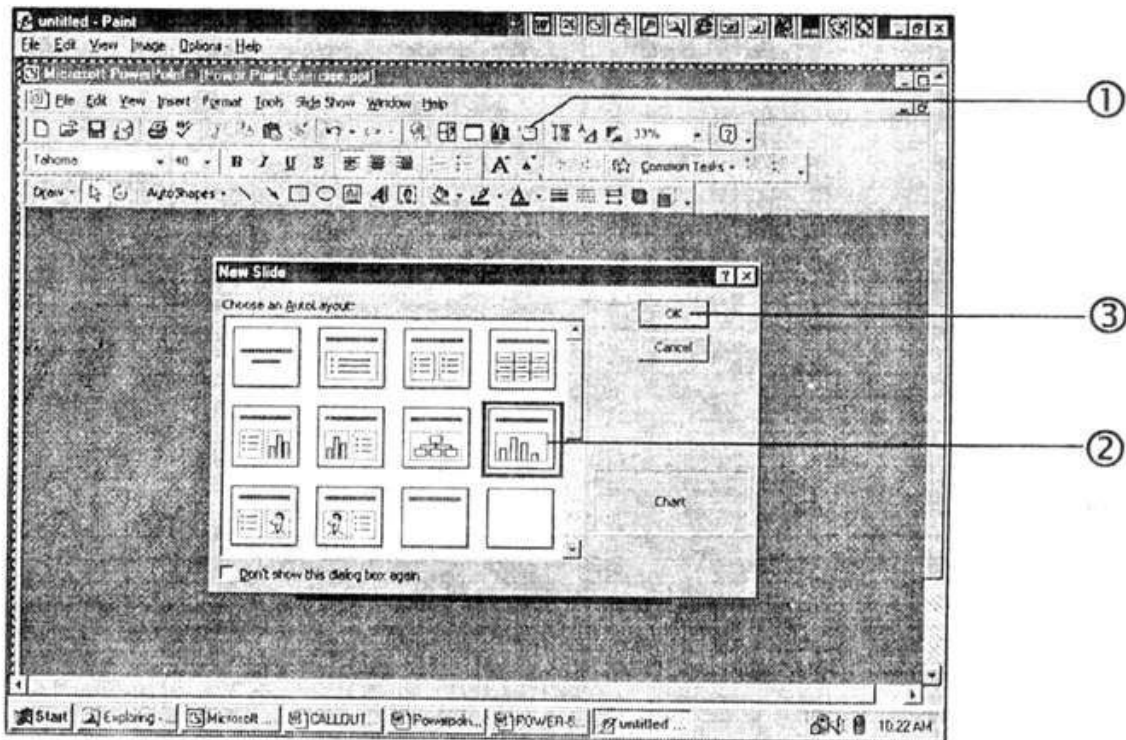
- 1 Select the picture. Right click with the mouse on it. The following menu would appear. Select **Send to Back** command from **Order** menu.
 - 2 Click on the text box once again to add something. Type the additional text and click outside the box to finish.
- A Your fourth slide is complete!



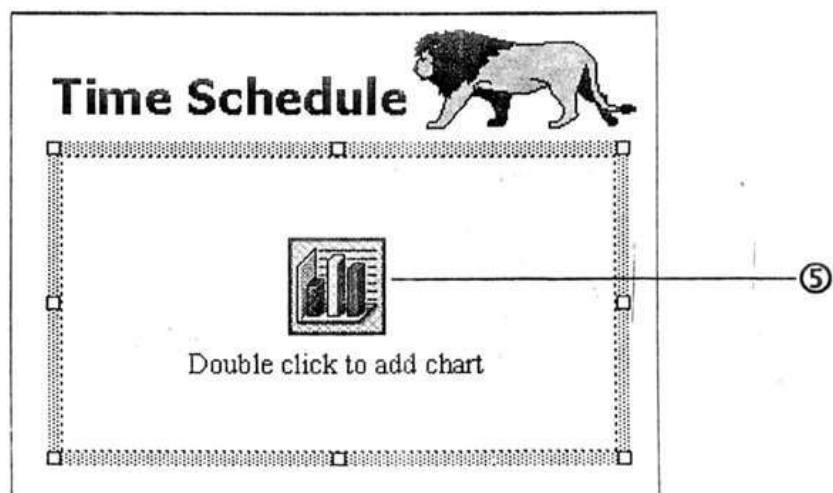
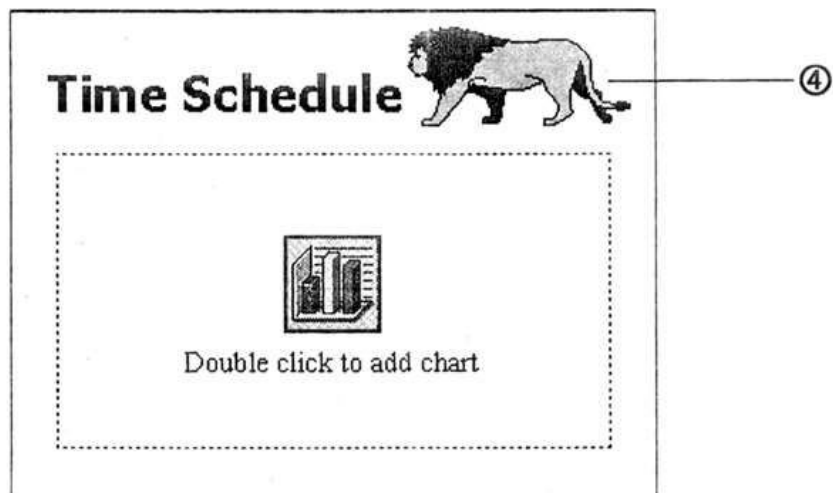
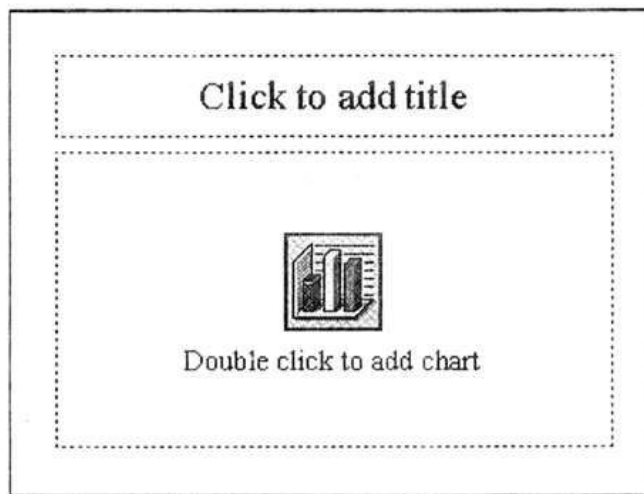


New Slide

- 1 Insert a new slide.
- 2 Choose **Graph** from **Autolayout** box.
- 3 Click on the **OK** button once to continue.



- 4 Add the title and the logo.
- 5 Double click here to start creating a graph.

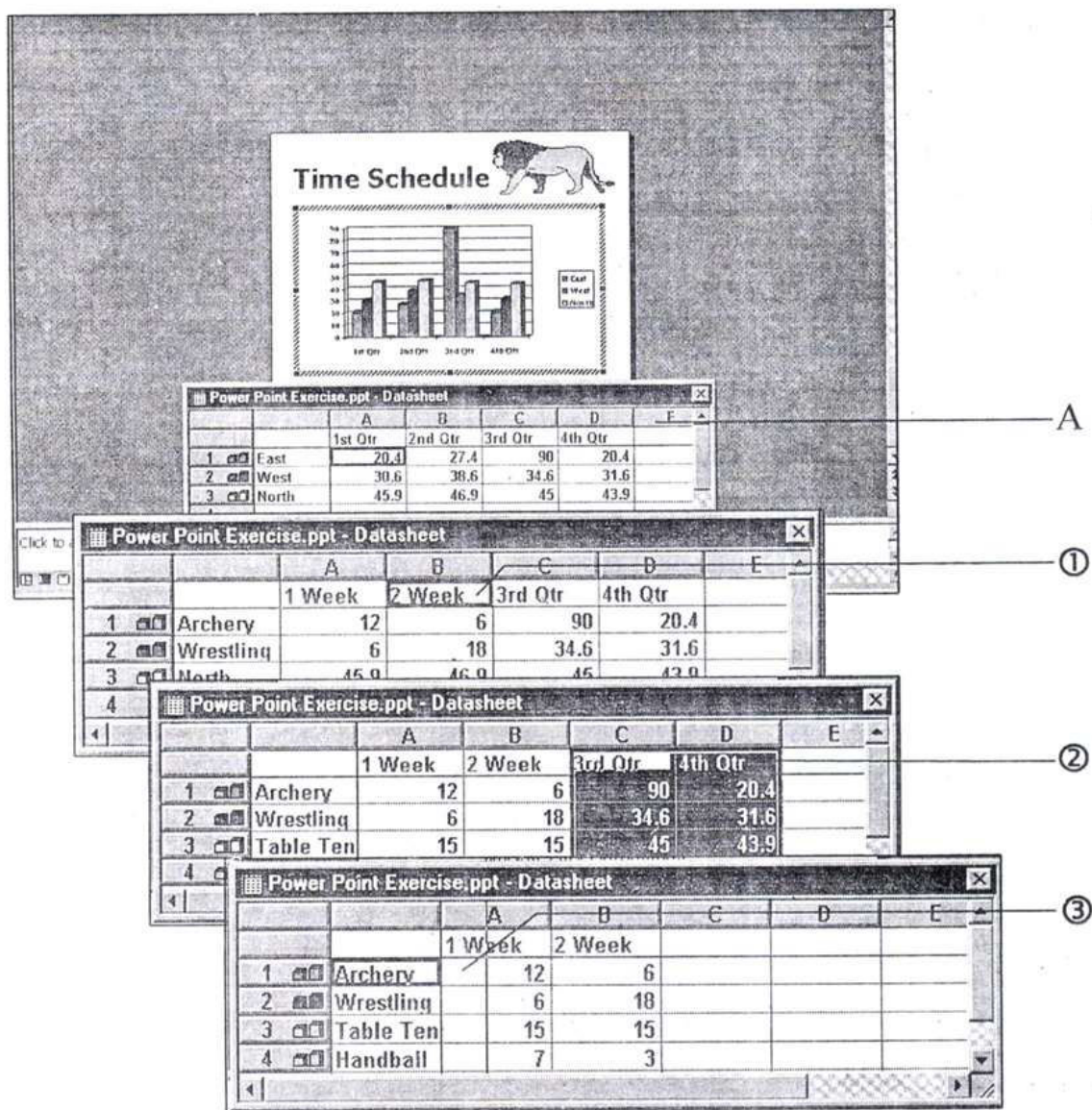


Entering Data to Graph

A This window would appear containing sample data.

- 1 Overwrite the sample data with your data.
- 2 Highlight and press **Del** to delete unwanted data.
- 3 Size this column so that the complete text is visible.
- 4 Click here to close the datasheet.

B You will see graph based upon provided data. You should insert the Footer after this and you fifth slide would be complete.



18.5 LET US SUM UP

Thus, MS PowerPoint is a presentation software program developed by Microsoft that allows users to create and deliver visual slideshows, combining text, images, videos, and other media to present information for various purposes, such as business proposals, educational lectures, or personal projects. It is part of the Microsoft Office suite and provides tools to design and manage slides, add animations and transitions, and share presentations in digital or printed formats.

18.6 GLOSSARY

- **Presentation:** A sequence of slides that conveys information, often used for talks, lectures, or proposals.
- **Slide:** A single page in a presentation, which can contain text, images, videos, and other elements.
- **Slide Layout:** The pre-defined arrangement of content placeholders on a slide, such as title, text, and image placeholders.
- **Slide Master:** Stores the formatting, fonts, color schemes, and layouts for all the slides in a presentation, ensuring consistency.
- **Template:** A pre-designed file that contains formatting, styles, and sometimes sample content to help you quickly create a new presentation.
- **Presentation tool:** Its primary function is to create slideshows, which are collections of individual pages, called “slides,” that convey a message to an audience.
- **Multimedia integration:** Users can enhance their presentations with text, graphics, photos, videos, sounds, and animations to make the content more engaging and understandable.
- **Part of Microsoft Office:** It is bundled with other Microsoft Office applications like Word and Excel, making it a common tool in professional and educational settings.
- **Versatile applications:** PowerPoint presentations are used in diverse fields, from business meetings and investor pitches to school lectures and tutorials.
- **Slide show format:** The software organizes content into slides, which are presented sequentially, often with speaker narration, to guide the audience through the information.
- **Design and editing tools:** It provides a set of tools for inserting content, choosing professional designs, and adding visual effects like transitions and custom animations.
- **Sharing and collaboration:** PowerPoint allows users to save their presentations in various formats, including handouts and webpages, and even share them in real-time for collaborative work.

18.7 SELF-ASSESSMENT QUESTIONS

1. What are the steps involved in creating a new presentation.

2. What is the option available for creating a presentation.

18.8 LESSON END EXERCISE

- Q1. What are the different views in PowerPoint.

- Q2. What is the default file extension for a modern PowerPoint presentation.

- Q3. What is the purpose of a title slide.

Q4. Why are animations and transitions used.

18.9 SUGGESTED READINGS

1. Argyrous, George. 1997. *Statistics for Social Research*. New York: Mc Millan Press Ltd.
2. Goods, W.J. & Hatt, P.K. 1981. *Methods in Social Research*. New York: Mc Graw Hill.
3. Gupta, S.C. 1981. *Fundamentals of Statistics*. Bombay: Himalayan Publishing House.
4. Gupta, S.P. 2004. *Statistical Methods*. New Delhi: Sultan Chand and Sons.

18.10 ANSWERS TO CHECK YOUR PROGRESS

1. **Answer:** (c) Templates
2. **Answer:** (a) title
3. **Answer:** (d) AutoContent Wizard
4. **Answer:** (b) Template
5. **Answer:** (a) 400%
6. **Answer:** (a) .pot
7. **Answer:** (b) Pack and Go
8. **Answer:** (d) WAV files
9. **Answer:** (c) Ctrl + M
10. **Answer:** (b) Rotate
11. **Answer:** (c) Outline view
12. **Answer:** (d) Common tasks toolbar
13. **Answer:** (a) Menu bar
14. **Answer:** (c) Zoom
15. **Answer:** (a) Landscape

INTRODUCTION TO MICROSOFT OFFICE**STRUCTURE**

- 19.0 Objectives
- 19.1 Introduction
- 19.2 Quick Access Toolbar
- 19.3 Tabs
- 19.4 Groups
- 19.5 Font Size
- 19.6 Styles
- 19.7 MS- Excel
- 19.8 MS-Word
- 19.9 Let us sum up
- 19.10 Self-Assessment Questions
- 19.11 Lesson End Exercise
- 19.12 Suggested Readings
- 19.13 Answers to check your progress

19.0 OBJECTIVES

The main objectives of this lesson are:

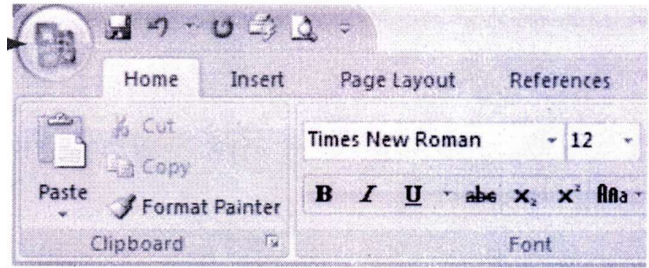
- To understand about the meaning of Microsoft Office.
- To know how to use Microsoft office.
- To have understanding about the tabs, paragraphs, graphical representations and other various tools used in Microsoft office.
- To know about the use of Microsoft office in sociological research.

19.1 INTRODUCTION

The Microsoft Office Button

We'll use **Microsoft Word 2007** for our initial illustrations of Ribbon, Tab and Group examples.

The first thing you'll notice, when you open a 2007 Office application is that there is no longer a File choice in the Menu Bar. The arrow above points to the **Microsoft Office Button** - which replaces File.

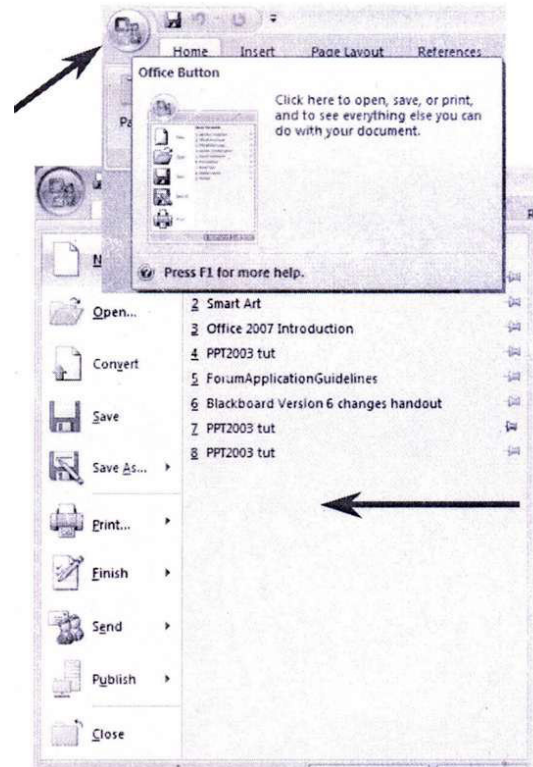


As you **move** your **cursor over** the **Microsoft Office Button** a **preview image** (image on right) **will appear**.

Click the Microsoft Office button.

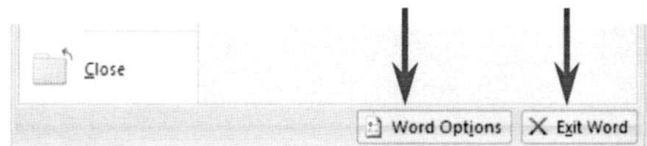
When you **click** the **Microsoft Office button**, it will turn orange and a **“File like” menu will appear** (similar to the image on the right).

You'll notice that you now have little images for choices and that some of them have little arrows pointing to the right.



If you look at the **bottom** of the **Microsoft Office Button menu screen** you will see two buttons. Since we're using

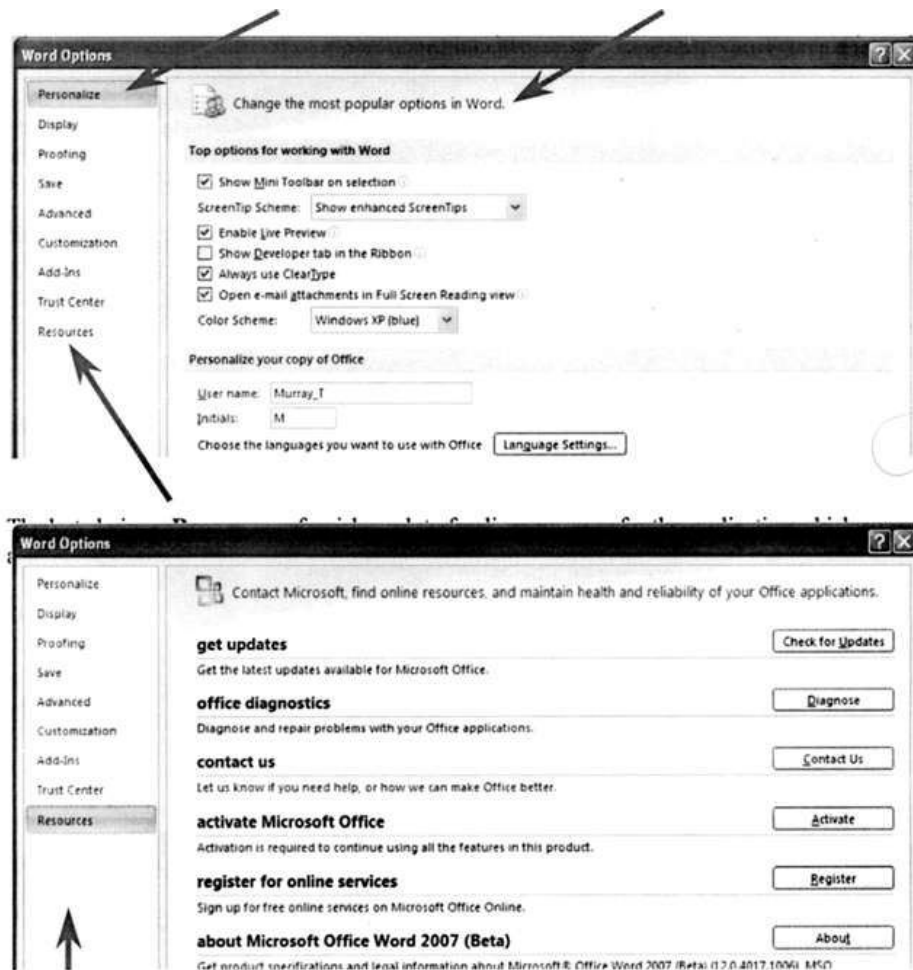
Word, the buttons indicate **Word Options** and **Exit Word**.



The buttons change with each application (e.g. PowerPoint will indicate PowerPoint Options).

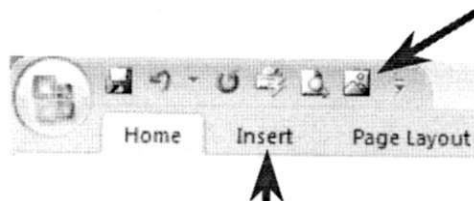
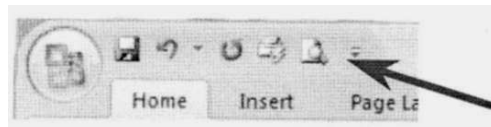
When you **click** the **Word Options button** the **image below** will **appear**. **Notice**, on the **left side** of the **menu screen** there are a number of choices (e.g. Personalize, Display, Proofing, etc.). When you click a choice on the left side of the screen, the options for that choice appear on the right. Take a few minutes and **move through these choices to familiarize yourself with this menu screen**. You will see

that Microsoft has placed a lot of resources that were under File Tools-Options, in previous versions of Office, in this menu.



Notice all of the useful online resources available to you.

19.2 QUICKACCESS TOOLBAR

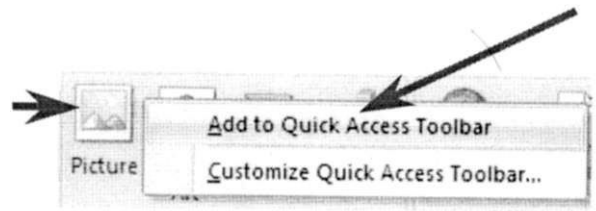


the **upper left corner** - to the right of the Microsoft Office Button - you will see an area called the **Quick Access Toolbar** (image on left). This area is quite handy as it currently

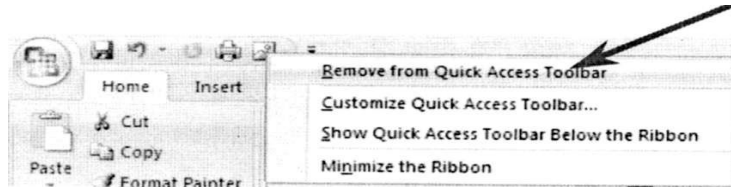
contains several of the most used buttons in Office applications - Save, Undo, Redo, Print and Print Preview. You can customize this toolbar by adding and removing as many Quick Access button choices as you desire.

In the Quick Access Toolbar (on the left) you can see we added the Insert Picture button - since we are using it a lot for this tutorial.

To **add** this button to the toolbar we **first clicked** the **Insert Tab** and then **RIGHT** clicked the **Insert Picture button**. One of the choices was **Add to Quick Access Toolbar**.



When we **clicked this choice** the Insert Picture button was added. You can add any button you choose by doing this.



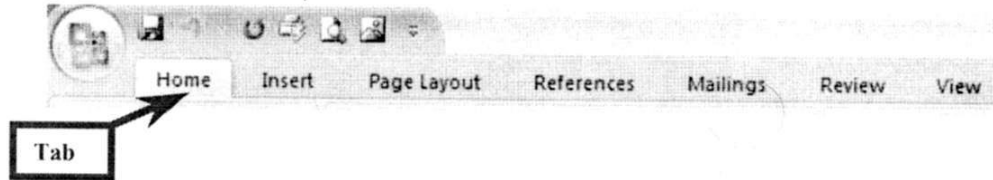
To remove buttons from the Quick Access Toolbar just **RIGHT** click on the button you desire to remove and **choose Remove from**

Quick Access Toolbar.

Ribbons

This is the new term you hear a lot about in 2007 Office. Ribbons stretch across the top of your application screen with features to assist you as you click the Ribbon Tabs. To us, Tabs and Ribbons are the same. It like unreeling holiday ribbon from a spool and seeing new images on the ribbon - very cool! So, we'll cover Tabs/Ribbons in great detail.

19.3 TABS



Below the Microsoft Office Button and Quick Access Toolbar we see a series of

Tabs/Ribbons.

Tabs are similar to the Drop-Down Menu choices in previous versions of Office. The Tabs are, logically, a bit different for each 2007 Office application to assist you with the most common features of that application. All the 2007 Office applications begin with the Home tab.

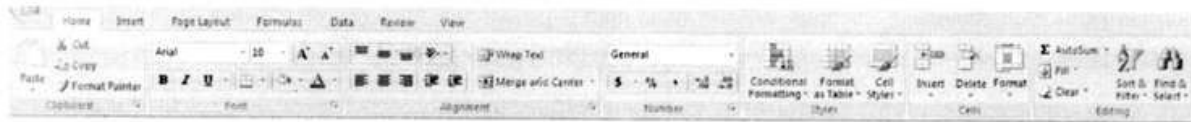
The **Home** Tab/Ribbon for **Word 2007** looks like the image below.



The **Home** Tab/Ribbon for **PowerPoint 2007** looks like the image below.



The **Home** Tab/Ribbon for **Excel 2007** looks like the Image below.



The **Home** Tab/Ribbon for **Access 2007** looks like the Image below.



You'll quickly notice that the **Home Tab/Ribbon** for each application shows the **Clipboard** as the left "**Group**" (except in Access) In Word and Excel, the Font Tab/Ribbon is to the right, but in PowerPoint, because working with slides is paramount, the Slides Tab/Ribbon comes next. If you have 2007 Office installed on your computer, **open these four applications and take a few minutes looking at each application's Home Tab/Ribbon.**

Notice, the **Tabs** to the **right** of the **Home Tab/Ribbon** are **tailored** to each **application**. We'll work a bit with this in a little while.

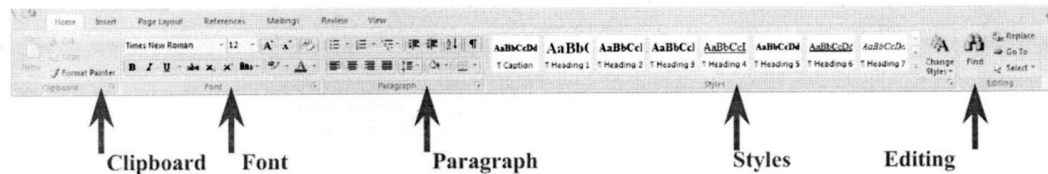
CHECK YOUR PROGRESS-I

Fill in the blanks:

- Microsoft _____ comes with the Microsoft package.
- First screen of MS-Word called MS-Office Word _____.
- The Quick Access tool bar is a _____ tool bar.
- Formatting refers to change the _____ of document.
- Saving a document refers to keep the document safely for _____.
- _____ collects items after using 'Copy' or 'Cut' command.
- Page _____ are the blank spaces around the edge of the page.
- _____ determines the distance of the paragraph from either left or right margin.
- The first line of the paragraph is not indented, but subsequent lines are indented, this is called _____.
- Ribbon is called _____ user interface.

19.4 GROUPS

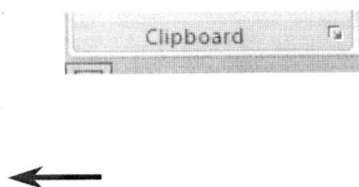
In the image below, the arrows point to a new topic - **Groups**.



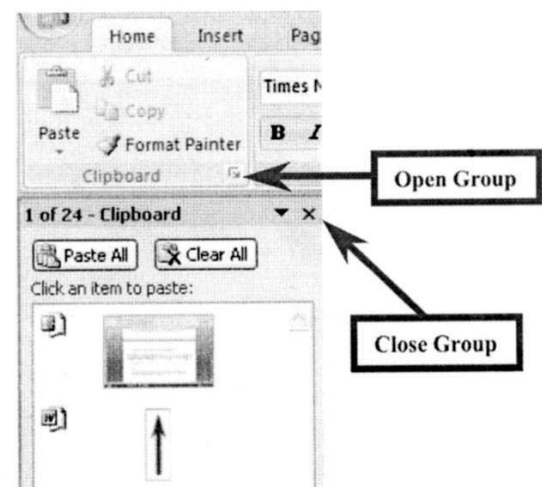
Clipboard Group

The Tab/Ribbon bar images (in this tutorial) are hard to read, so we've placed **arrows** (in the **image above**) for the **Groups** in the **Word Home Tab/Ribbon**. Again, the Tabs/ Ribbons, and Groups, will vary depending on the application you're using. Let's look a bit at the **Groups in Word**.

The first **Group** on the **Word Home Tab** is **Clipboard**. To **open a Group**, you **move your cursor over the little down pointing arrow in the lower right corner of a group**.



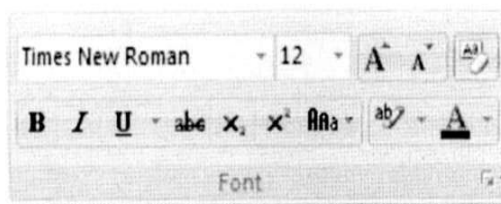
This arrow is enlarged in the image below.



When you **click this arrow** the **image on the right** appears. Notice that the Clipboard appears on the left side of your screen and shows any text or images you've copied. To close this group, click the "X" in the upper right corner of the Group.

19.5 FONT SIZE

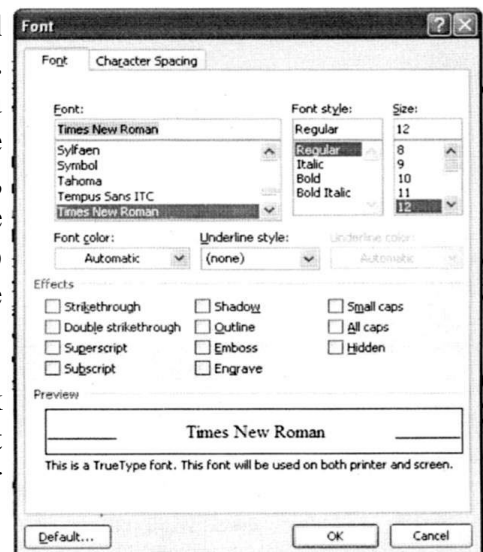
Notice, in the **Font Group** area (above), you have the **most used**



Font features. However, if you **desire all of the font features**, just click the **Open Group** arrow to the

right of Font.

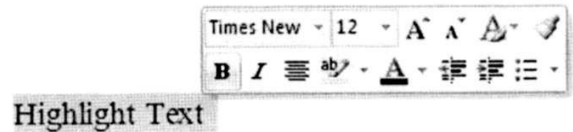
An old friend - the **Font menu screen** appears (when you click the Open Group arrow). You'll see this a lot as you learn more about 2007 Office. Many of the "tried and true" menu screens will appear in logical places.



Select Text Mini Toolbar

When you're working with text and fonts a really ingenious **“new thing”** occurs as you highlight text - a Select Text Mini Toolbar appears!

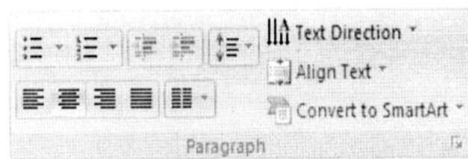
In the **image** on the **right** we **highlighted** - Highlight Text. When we **paused the cursor over the highlight**, a **“shadow like” toolbar**



appeared. When we **move our cursor over the toolbar**, it is ready for us to use it to modify our text.

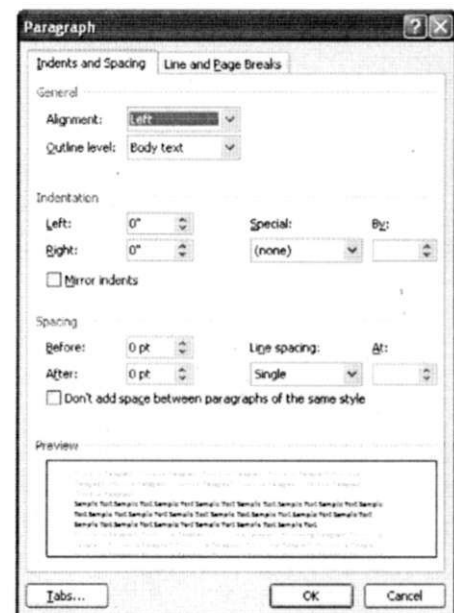
This is really handy as many of text formatting features are in the Mini Toolbar. The first time you try this, be patient, it sometimes takes a few tries.

Paragraph



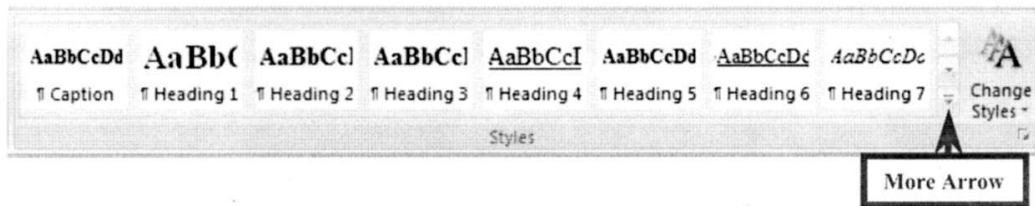
Notice in the **Paragraph Group** area (left) you again have the most used Paragraph features. However, if you **desire all of the paragraph features**, just **click the Open Group arrow** to the right of Paragraph.

The **Paragraph menu screen appears** when you **click the Open Group arrow** to the right of the Paragraph Group. You should now have a “feel” for how the Tabs/Ribbons and Groups work together to assist you.

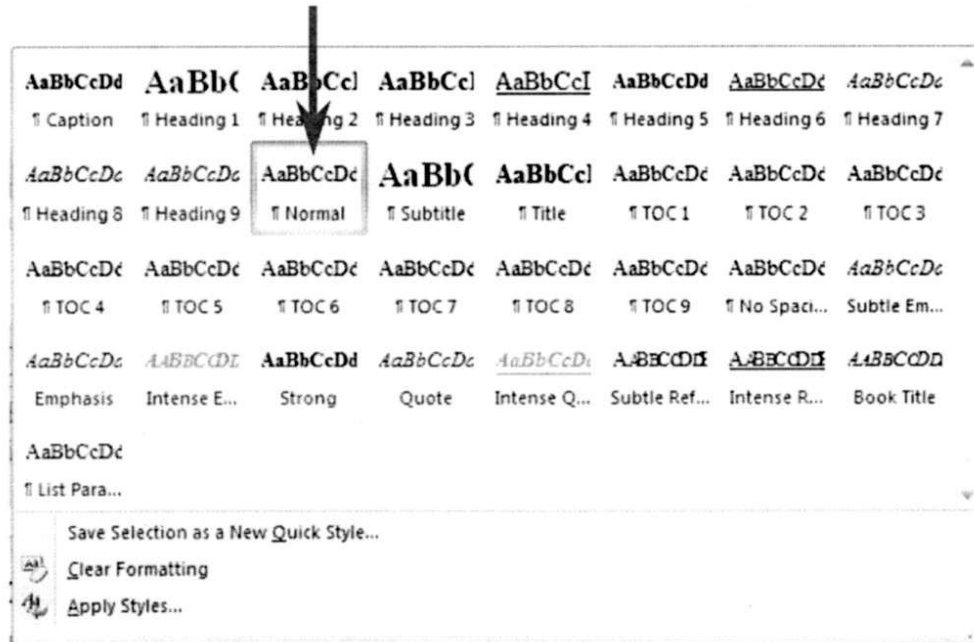


19.6 STYLES

Hang on! The next Group on the Word Home Tab/Ribbon is Styles. If you go back to Page 6 and glance at the Word, PowerPoint and Excel Home Tabs, you'll see that the right portion of a Tab is where the application selections change to fit the application. In **Word** you **can now select a style** from the **Styles Group** (image below). If you **click the More arrow** in the lower right corner of the **Styles group**, you will see additional choices.

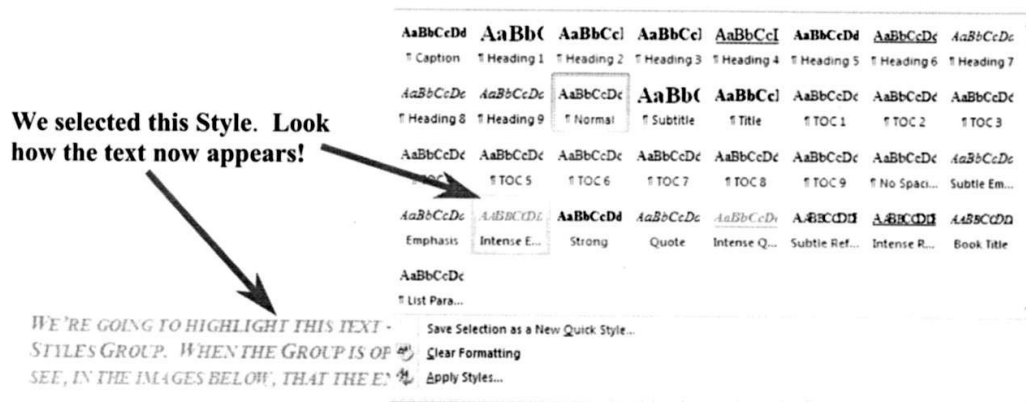


When you **click the More arrow**, you will see an image similar to the one below. Notice that we are in **Times New Roman - Normal**. On the **next page** we'll show you one of the really, really neat new features in 2007 Office.

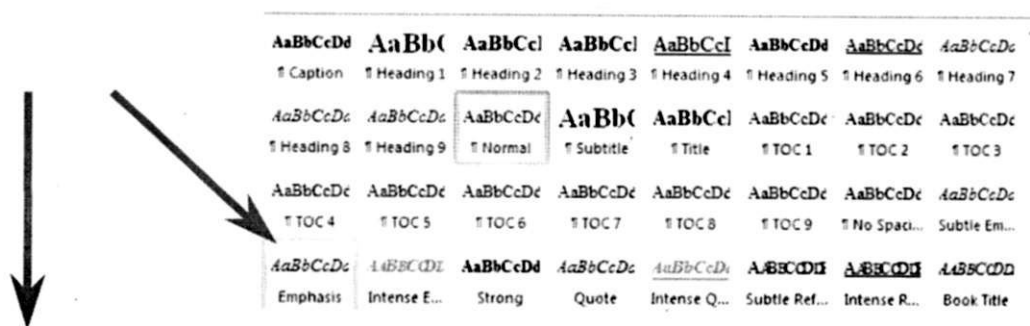


Fasten your seatbelts!

We're going to **highlight this paragraph** (when we have finished typing it). Then we're going to **open the Styles Group**. When the Group is open we'll **move our cursor over the choices**, and as we do, you'll see, in the images below, that **the entire paragraph changes to that Style!**



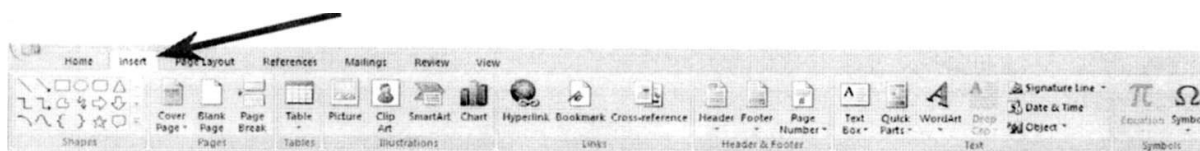
And another.....



We're going to highlight this text - when we have finished typing it. Then we're going to open the Style Group. When the Group is open, we'll mover our cursor over the choices, and as we do, you'll see, in the images below, that the entire paragraph changes to that Style!

Other Tabs/Ribbons -

When you move to the other Tabs/Ribbons, you'll notice that they contain their own Groups associated with that Tab. The **Insert Tab/Ribbon** (below) has logical "things" that you would insert into a document - Shapes, Pages, Tables, Illustrations, Links, Headers/ Footers, Text and Symbols. Again, depending on your choices, many selections allow you to "preview" what you've highlighted - similar to the two illustrations above.

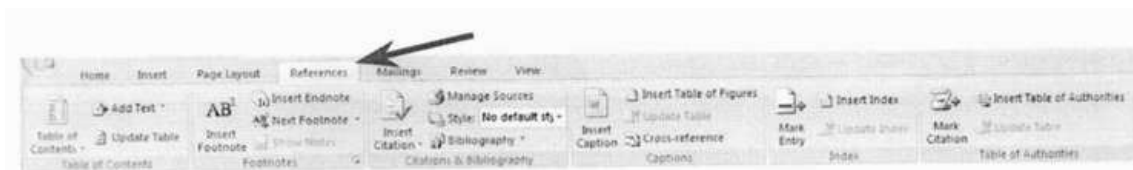


It is **suggested** that you **click the Tabs/Ribbons** in **each application** you'll be using to get a "feel" for them.

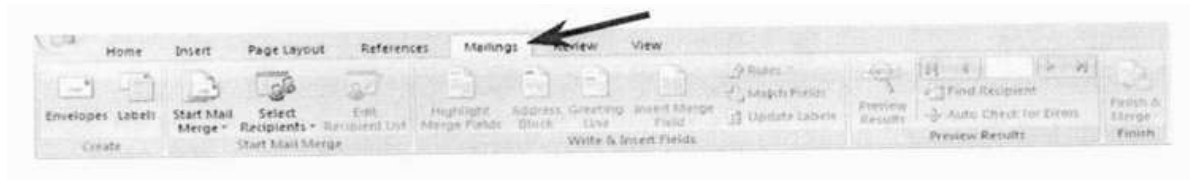
The **Page Layout Tab/Ribbon** also has logical selections - Themes, Page Setup, Page Background, Paragraph and Arrange.



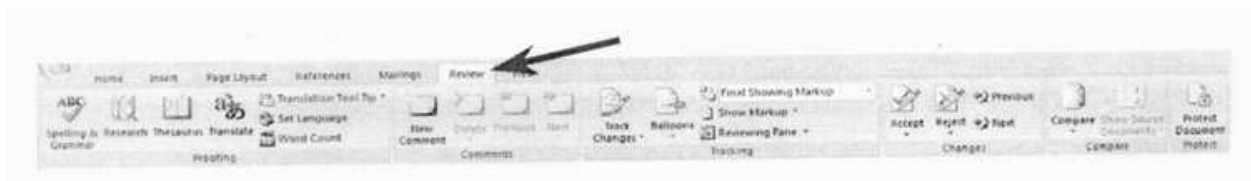
The **References Tab/Ribbon** will really come in **handy** for those publishing **long documents, articles or books** - Table of Contents, Footnotes, Citations & Bibliography, Captions, Index, and Table of Authorities.



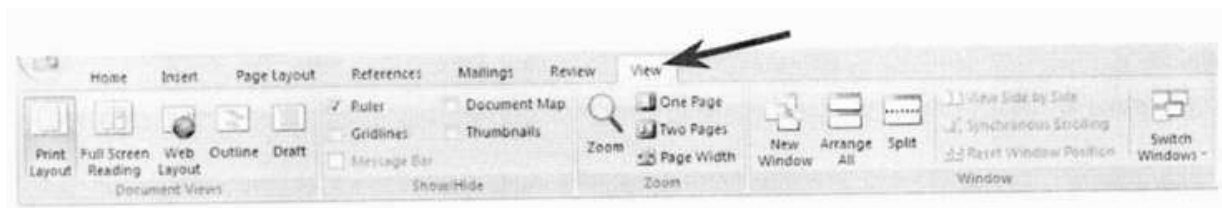
The **Mailings Tab/Ribbon** lets you work with Envelopes, Labels, Mail Merge, Fields and Preview. It includes Create, Start Mail Merge, Write and Insert Fields, Preview Results and Finish.



The **Review Tab/Ribbon** has the Proofing Tools, Comments, Tracking, Changes, Compare and Protect features.



The **View Tab/Ribbon** allows you to change the document Views, do Show/Hide, Zoom and arrange your Windows.

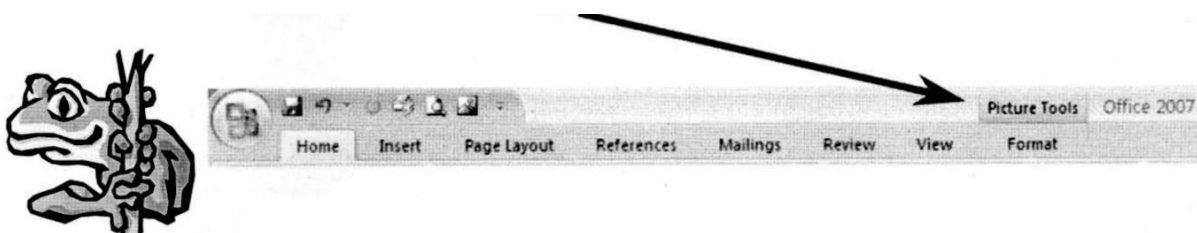


This gives you a “feel” for how the Tabs/Ribbons work in **Word 2007**. Again, **it would be prudent to look at the other 2007 Office applications you will be using - to get a similar sense for these new features.**

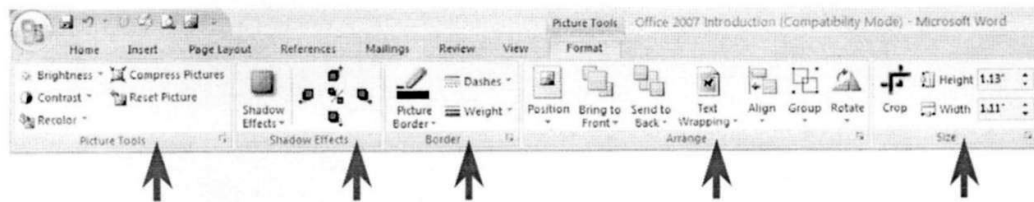
Now we’ll look at several other neat features of 2007 Office. Picture Tools

Currently, when you **click** an image in **Word 2007**, **PowerPoint 2007** or **Excel 2007**, a **Picture Tools Tab/Ribbon** will be **available** to you. We placed a Microsoft Clip Art frog on the left.

When we **click** the frog a **Picture Tools Tab** appears above of the **other Tabs/ Ribbons**.

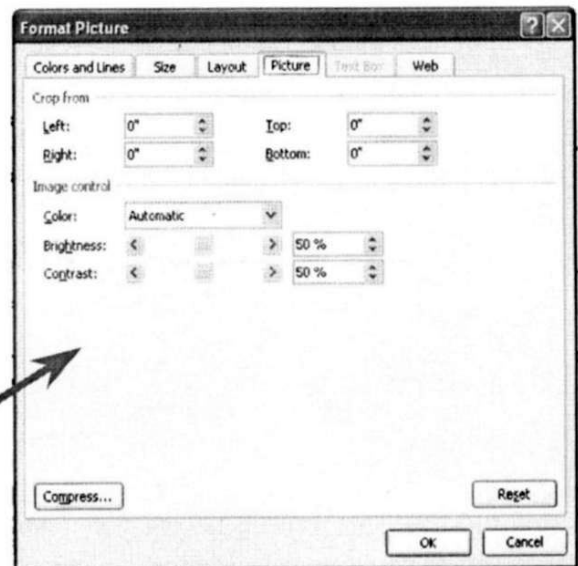
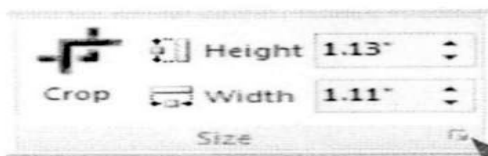


When you **click the Picture Tools Tab** (we're still in Word) the **Picture Tools Ribbon** below appears.



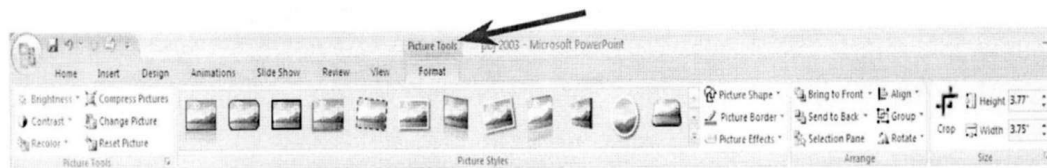
Notice, like the other Ribbons, that Picture Tools also has its own Groups - Picture Tools, Shadow Effects, Border, Arrange, and Size.

You can **click the Open Group arrow** at the lower right of some groups to see more of the Group.



We clicked the **Open Group arrow** on the **Size Group** and the **Format Picture Menu Screen** appeared.

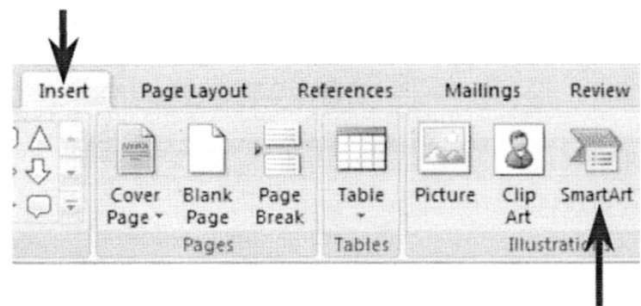
If we are in **PowerPoint** - and click an image - **Picture Tools** becomes available. The image below shows that there are different selections since we are now using PowerPoint.



SmartArt

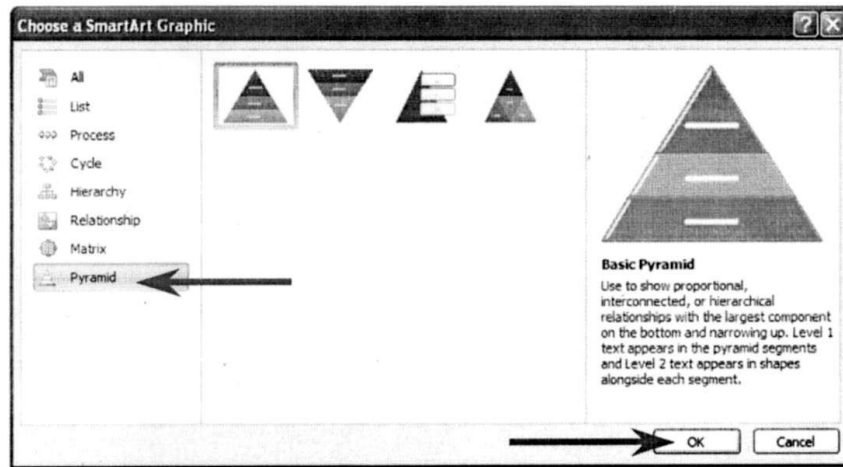
In the **Insert Ribbon/Tab** at the bottom of page 10 there is a new selection that improves on the “old” Drawing Toolbar - especially **SmartArt**. SmartArt is a part of **Word, Excel and PowerPoint**.

We've **enlarged the Word Insert Ribbon/Tab** (right) to show the SmartArt selection. When you **click SmartArt** a **Choose a SmartArt Graphic** menu (image below) will appear.



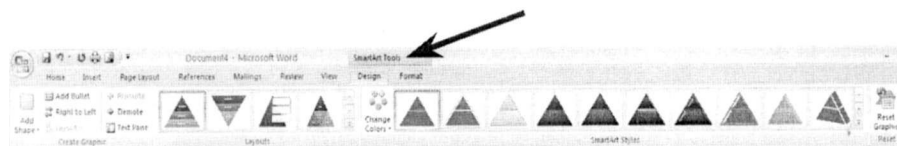
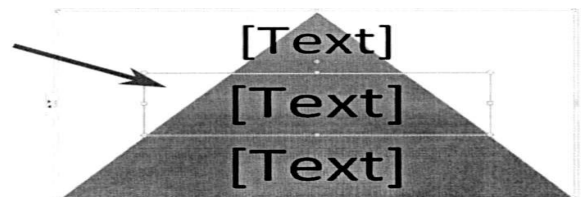
If you have used SmartArt in the past, you'll quickly see that it has been **greatly enhanced**.

We'll **click** on the **Pyramid** and then click the **OK** button.



A Pyramid Diagram, similar to the one on the right, will appear. Now it gets **exciting!**

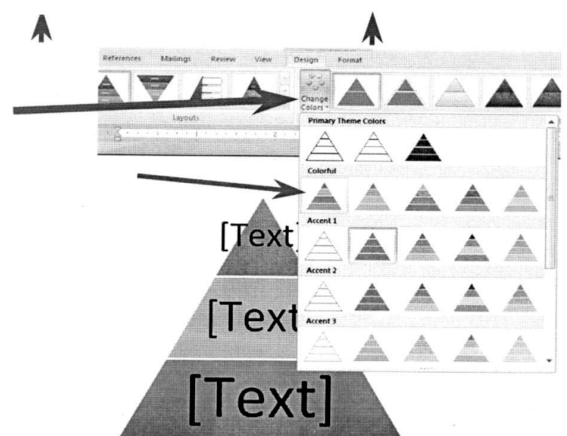
When you **click** the **Pyramid** you'll notice a **new SmartArt Tools Ribbon/Tab** appears (top of next page).



Similar to Picture Tools, you'll **notice** several **Layout** and **SmartArt Styles Groups** designed for enhancing the Pyramid on which you're working.

If you click the **Change Colors** button in the **SmartArt Styles Group** an **image** like the one the **right** will appear. As you **move your cursor arrow** over the **Primary Theme Colors**, you'll see that the **Pyramid** changes to that **color**. We choose the one you see marked by the **arrow** on the **right**.

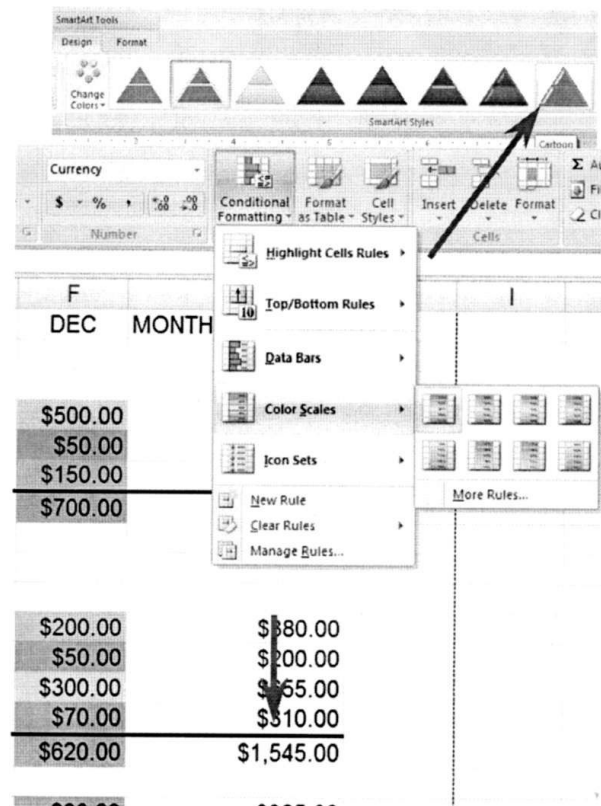
Our Pyramid now has this shading!



If you now **move your cursor arrow over one of the images in SmartArt Styles** you'll see an image similar to the one on the right.

Comparable to the Text Styles on Page 10, you can see how 2007 Office is enhanced to assist you with these great previews.

We'll work with these in the individual 2007 Office tutorials.



Other Ribbons/Tabs/Tools

19.7 MS-EXCEL

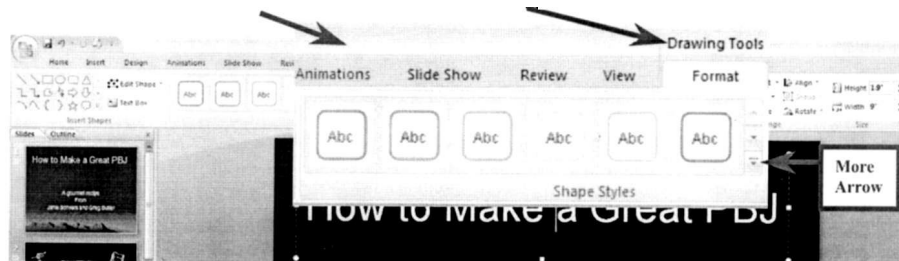
If you are in Excel, **you can now highlight a row, column or entire spreadsheet with really eye-opening effects.**

In the **image on the right**, we opened the spreadsheet developed with the Excel 200 tutorial. We **highlighted the December column** and then **clicked Conditional Formatting**. The **drop-down menu** you see on the right **appeared**. We then **clicked Color Scales**, when the **area to the right of Color Scales appeared**, we **moved our cursor over the secretions**. As with other 2007 applications, when you move your cursor over the choices **you will get a temporary preview of how your selection will appear.**

Notice, in Conditional Formatting, there are also **Data Bars** and **Icon Sets** selections. If you were to choose these you would see small bar charts or little flags, smiley faces, etc. appear in the area you highlighted. And the list goes on and on. Really awesome!

PowerPoint

You saw on Page 12 that **Picture Tools** is a significant part of Power Point 2007. **Text and titles are also very important.** If you **click a Text Box** an image (similar to the one below) will appear. Notice that a **Drawing Tools Tab/Ribbon is available.**



We clicked the **Drawing Tools Tab** then **clicked** the **More arrow** to the **lower right** of the **Shape Styles Group**. An image similar to the one below appeared.

As with other Tools Tabs/Ribbons, when you **move your cursor arrow over a selection** in the Format area - **a preview of how your text will look with that selection appears**. We chose the one marked by the **arrow below** and our **title looks like** the one on the right of the image.



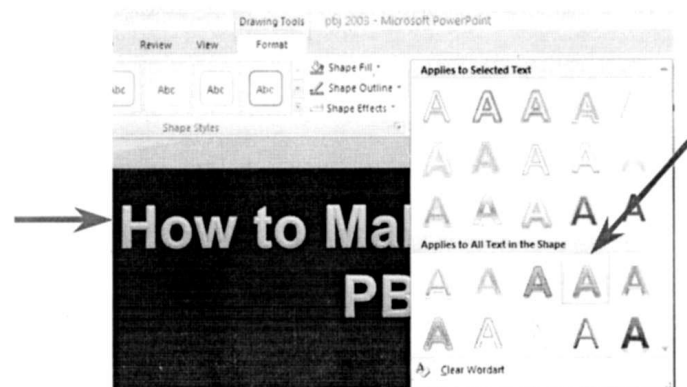
Also, in the Drawing Tools Tab/Ribbon, is the Group **WordArt Styles**.



We **clicked** the **More arrow** to the right of Word Art Styles and the image on the right appeared. Once again, as we **moved our cursor over the choices**, a **preview** of our

title **appeared** in that WordArt Style.

We'll work with these Styles, and more, in the PowerPoint tutorial.

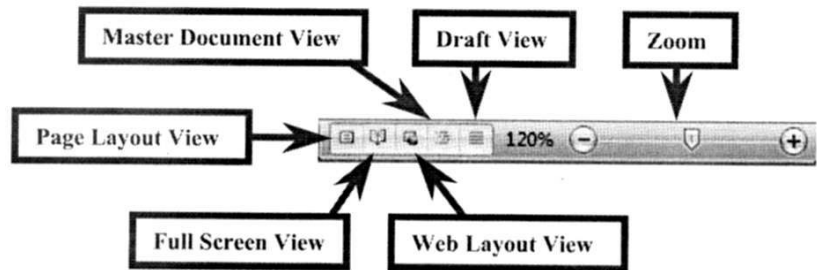


The Lower Right Corner

Another efficient feature of 2007 Office is in the lower right corner of Word, Excel, PowerPoint and Access. When you open these applications, you will see that the “zoom” feature is now available, as well as other logical “view” features for each application.

19.8 MS-WORD

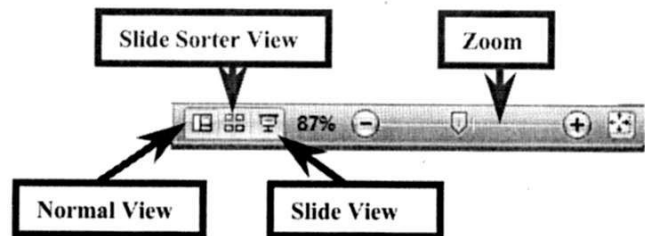
The image on the right is the **Word View Toolbar** (located on the bottom Right of the Word screen).



You'll notice that normal Word document views and zoom features are available.

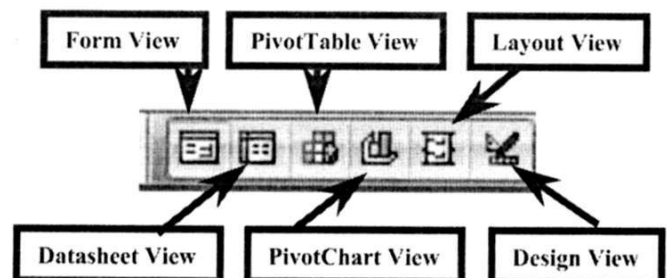
PowerPoint

The **PowerPoint View Toolbar** looks similar to the image on the right.



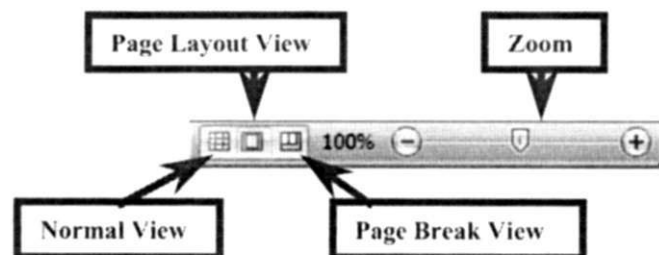
Access

The **Access View Toolbar** looks similar to the image on the right.



Excel

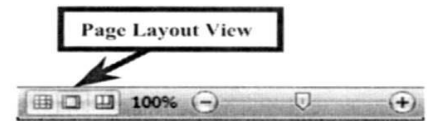
The **Excel View Toolbar** looks similar to the image on the right.



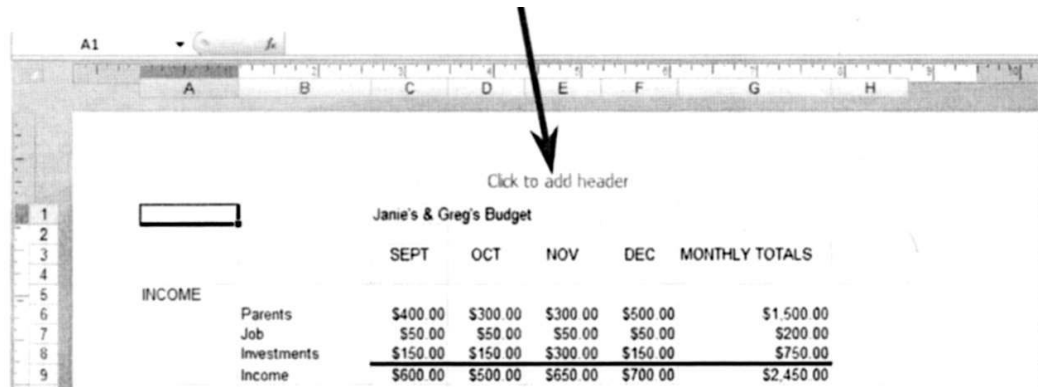
We have found these View toolbars to be very handy as we've worked in these applications.

More Excel

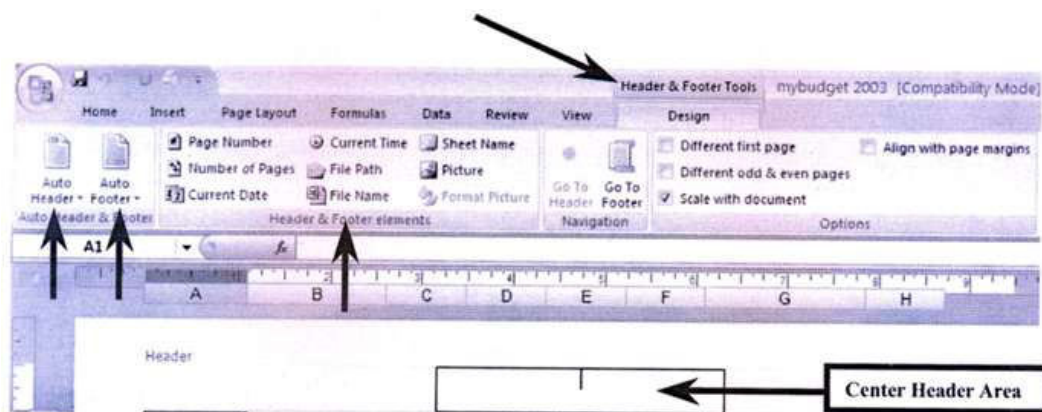
While we're in the Excel View Toolbar we'll mention the new **Page Layout View**.



When you **click the Page Layout View button**, an image similar to the one below will appear. This view is **similar to Print Layout View in Word**. Now you have it in Excel! It's really great as it **not only gives you a "mini" print preview of your spreadsheet, it also allows you to work with your Headers and Footers "interactively" by clicking the Header or Footer area!**



We **clicked** in the **center Header area** and the image below appeared. **Notice** that a **Header and Footers Tools Tab/Ribbon** is available - with all of the Header and Footer Groups - and easy to use buttons.



Notice the Auto Header and Auto Footer buttons to the **left of the Ribbon**. You can use these, or the **Header & Footer elements**, or **simply type your header**. This is really flexible and you see your choices instantly.

You can see that 2007 Office is working with you more than ever. As we create the 2007 Office

tutorials, we'll introduce you to, and show you how to use these Ribbons, Tabs, Groups, and Toolbars unique to each application.

At the moment Word, Excel, PowerPoint, Access and creating messages in Outlook Mail contain these new features. 2007 Publisher is fairly similar to Publisher 2003 - but has Publisher Tasks to assist you in creating Publications. FrontPage has changed its name and moved to a new family called Expression - with a new name - Microsoft Expression Web Designer.

APPENDIX

Table 1 : Significant Values of Student's t (with n degrees of freedom) Level of Significance

	.005	.01	.025	.05	.10	.20	.25
One tail							
Two tail	.01	.02	.05	.10	.20	.40	.50
γ							
1.	63.66	31.82	12.71	6.31	3.08	1.376	1.000
2.	9.92	6.96	4.30	2.92	1.89	1.061	.816
3.	5.84	4.54	3.18	2.35	1.64	.978	.765
4.	4.60	3.75	2.78	2.13	1.53	.941	.741
5.	4.03	3.36	2.57	2.02	1.48	.920	.727
6.	3.71	3.14	2.45	1.94	1.44	.906	.718
7.	3.50	3.00	2.36	1.90	1.42	.896	.711
8.	3.36	2.90	2.31	1.86	1.40	.889	.706
9.	3.25	2.82	2.26	1.83	1.38	.883	.703
10.	3.17	2.76	2.23	1.81	1.37	.879	.700
11.	3.11	2.72	2.20	1.80	1.36	.876	.697
12.	3.06	2.68	2.18	1.78	1.36	.873	.695
13.	3.01	2.65	2.16	1.77	1.35	.870	.694
14.	2.98	2.62	2.14	1.76	1.34	.868	.692
15.	2.95	2.60	2.13	1.75	1.34	.866	.691
16.	2.92	2.58	2.12	1.75	1.34	.865	.690
17.	2.90	2.57	2.11	1.74	1.33	.863	.689
18.	2.88	2.55	2.10	1.73	1.33	.862	.688
19.	2.86	2.54	2.09	1.73	1.33	.861	.688
20.	2.84	2.53	2.09	1.72	1.32	.860	.687
21.	2.83	2.52	2.08	1.72	1.32	.859	.686
22.	2.82	2.51	2.07	1.72	1.32	.858	.686
23.	2.81	2.50	2.07	1.71	1.32	.858	.685
24.	2.80	2.49	2.06	1.71	1.32	.857	.685
25.	2.79	2.48	2.06	1.71	1.32	.856	.684
26.	2.78	2.48	2.06	1.71	1.32	.856	.684
27.	2.77	2.47	2.05	1.70	1.31	.855	.684
28.	2.76	2.47	2.05	1.70	1.31	.855	.683
29.	2.76	2.46	2.04	1.70	1.31	.854	.683
30.	2.75	2.46	2.04	1.70	1.31	.854	.683
40.	2.70	2.42	2.02	1.68	1.30	.851	.681
60.	2.66	2.39	2.00	1.67	1.30	.848	.679
120.	2.62	2.36	1.98	1.66	1.29	.845	.677
∞	2.58	2.33	1.96	1.645	1.28	.842	.674

Table 2 : Significant Values of χ^2 (with n degrees of freedom) Level of Significance

γ	.005	.01	.025	.05	.10	.25	.50
1	7.88	6.63	5.02	3.84	2.71	1.32	.455
2	10.6	9.21	7.38	5.99	4.61	2.77	1.39
3	12.8	11.3	9.35	7.81	6.25	4.11	2.37
4	14.9	13.3	11.1	9.49	7.78	5.39	3.36
5	16.7	15.1	12.8	11.1	9.24	6.63	4.35
6	18.5	16.8	14.4	12.6	10.6	7.84	5.35
7	20.3	18.5	16.0	14.1	12.0	9.04	6.35
8	22.0	20.1	17.5	15.5	13.4	10.2	7.34
9	23.6	21.7	19.0	16.9	14.7	11.4	8.34
10	25.2	23.2	20.5	18.3	16.0	12.5	9.34
11	26.8	24.7	21.9	19.7	17.3	13.7	10.3
12	28.3	26.2	23.3	21.0	18.5	14.8	11.3
13	29.8	27.7	24.7	22.4	19.8	16.0	12.3
14	31.3	29.1	26.1	23.7	21.1	17.1	13.3
15	32.8	30.6	27.5	25.0	22.3	18.2	14.3
16	34.3	32.0	28.8	26.3	23.5	19.4	15.3
17	35.7	33.4	30.2	27.6	24.8	20.5	16.3
18	37.2	34.8	31.5	28.9	26.0	21.6	17.3
19	38.6	36.2	32.9	30.1	27.2	22.7	18.3
20	40.0	37.6	34.2	31.4	28.4	23.8	19.3
21	41.4	38.9	35.5	32.7	29.6	24.9	20.3
22	42.8	40.3	36.8	33.9	30.8	26.0	21.3
23	44.2	41.6	38.1	35.2	32.0	27.1	22.3
24	45.6	43.0	39.4	36.4	33.2	28.2	23.3
25	46.9	44.3	40.6	37.7	34.4	29.3	24.3
26	48.3	45.6	41.9	38.9	35.6	30.4	25.3
27	49.6	47.0	43.2	40.1	36.7	31.5	26.3
28	51.0	48.3	44.5	41.3	37.9	32.6	27.3
29	52.3	49.6	45.7	42.6	39.1	33.7	28.3
30	53.7	50.9	47.0	43.8	40.3	34.8	29.3
40	66.8	63.7	59.3	55.8	51.8	45.6	39.3
50	79.5	76.2	71.4	67.5	63.2	56.3	49.3
60	92.0	88.4	83.3	79.1	74.4	67.0	59.3
70	104.2	100.4	95.0	90.5	85.5	77.6	69.3
80	116.3	112.3	106.6	101.9	96.6	88.1	79.3
90	128.3	124.1	118.1	113.1	107.6	98.6	89.3
100	140.2	135.8	129.6	124.3	118.5	109.1	99.3

Table 3 : Significant Points of F (5 percent) with n_1 and n_2 degrees of freedom

$n_2 \backslash n_1$	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.00	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.84	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81

(Contd.)

Table 3 : Continued

$n_1 \backslash n_2$	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.00	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Significant Points of F (1 per cent) with n_1 and n_2 degrees of freedom

$n_1 \backslash n_2$	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	5000	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.2	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.9	13.8	13.7	13.7	13.6	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49

(Contd.)

$n_2 \backslash n_1$	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
	n_1																		
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.82	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	5.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

19.9 LET US SUM UP

Thus, to conclude in computing, MS Office (Microsoft Office) refers to a suite of productivity software developed by Microsoft that includes applications like Word, Excel, and PowerPoint, used for tasks such as creating documents, managing data, and building presentations. It serves as an essential tool for individuals and businesses to organize, manage, and present information effectively across various computing environments.

19.10 SELF-ASSESSMENT QUESTIONS

1. Discuss in detail about MS-Office.

2. Explain briefly Tabs and Groups used in MS-Office.

19.11 LESSON END EXERCISE

- Q1. Match the following:**

- a. Ctrl+C: Undoes the last action.**
- b. Ctrl+V: Saves the document**
- c. Ctrl+Z: Copies the selected content.**
- d. Ctrl+S: Pastes the content.**

Q2. What does “MS” stand for.

Q3. What are the main components of MS Office.

19.12 SUGGESTED READINGS

1. Argyrous, George. 1997. *Statistics for Social Research*. New York: Mc Millan Press Ltd.
2. Goods, W.J. & Hatt, P.K. 1981. *Methods in Social Research*. New York: Mc Graw Hill.
3. Gupta, S.C. 1981. *Fundamentals of Statistics*. Bombay: Himalayan Publishing House.
4. Gupta, S.P. 2004. *Statistical Methods*. New Delhi: Sultan Chand and Sons.

19.13 ANSWERS TO CHECK YOUR PROGRESS

- a. Microsoft **Office Word** comes with the Microsoft package.
- b. First screen of MS-Word called MS-Office Word **Window**.
- c. The Quick Access tool bar is a **customizable** tool bar.
- d. Formatting refers to change the **appearance** of document.
- e. Saving a document refers to keep the document safely for **future use**.
- f. **Clipboard** collects items after using ‘Copy’ or ‘Cut’ command.
- g. Page **margins** are the blank spaces around the edge of the page.
- h. **Indentation** determines the distance of the paragraph from either left or right margin.
- i. The first line of the paragraph is not indented, but subsequent lines are indented, this is called **hanging indent**.
- j. Ribbon is called **graphical** user interface.

STATISTICAL DATA AND USE OF COMPUTERS**STRUCTURE**

- 20.0 Objectives
- 20.1 Introduction
- 20.2 Statistical Data
- 20.3 Use of Statistical Data with Computers
- 20.4 How Computers Enhance Statistical Analysis
- 20.5 Use of Computers in Social Science Research
- 20.6 Let us sum up
- 20.7 Glossary
- 20.8 Self-Assessment Questions
- 20.9 Lesson End Exercise
- 20.10 Suggested Readings
- 20.11 Answers to check your progress

20.0 OBJECTIVES

After going through this lesson, you should be able to:

- Understand the concept and significance of using computers
- Learn about use of statistical data.
- Learn how Computers Enhance Statistical Analysis

20.1 INTRODUCTION

Computers revolutionize sociology by enabling efficient handling of vast statistical datasets for quantitative analysis and facilitating qualitative data analysis through specialized software. They aid in data collection via online surveys, improve data storage and management, and automate the creation of graphs and charts to visualize trends. This computational power allows sociologists to explore complex social phenomena, run sophisticated models, and derive deeper insights from social data.

Nowadays, Technology is playing a very important role in every field of life and research is also not an exception to it. Researches are highly benefited with the advent of technological advancement like computers. The reason is not confined to the fact that it has made the researchers work more simpler and less expensive but also less time consuming too. With the advent of the computer the calculation and analysis part being made more easier and more representable. Computers are indispensable tools in the field of research. It not only provides space to store huge data but helps in retrieving the same stored data time and again with the same ease and convenience.

20.2 STATISTICAL DATA

Data Measurement Statistical data is generally obtained in many formats such as spreadsheets (e.g. MS Excel) or databases (e.g. MS Access). Data may also be received in various open formats such as typically tab-delimited text (*.dat, *.tab, *.txt), comma-separated text or fixed-width text data (*.dat, *.txt). The data could be of two types, qualitative and quantitative. Most of the statistical methods are based on the quantitative data. Quantitative variable is a variable whose values are numbers with real numeric meaning. It consists of mainly two types of data viz. discrete and continuous. A set of data is said to be discrete if the values/observations belonging to it are distinct and separate, i.e. they can be counted e.g. number of books in a library. Whereas, a set of data is said to be continuous if the values/observations belonging to it may take on any value within a finite or infinite interval. There are four well-known levels of measurement scales i.e. nominal, ordinal, interval, and ratio. There is a relationship between the level of measurement and the appropriateness of various statistical procedures. For example, it would be impractical to compute the mean of nominal measurements. Data must be measured on an interval or a ratio scale for the computation of means and other statistics to be valid.

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

- 1) **Nominal Scale:** Nominal measurement consists of assigning items to groups or categories. No quantitative information is conveyed and no ordering of the items is implied. Nominal scales are therefore qualitative rather than quantitative. Religious preference, race, and sex are all examples of nominal scales. Frequency distributions are usually used to analyse data measured on a nominal scale. The main statistic computed is the mode. Variables measured on a nominal scale are often referred to as categorical or qualitative variables. Nominal variables allow for only qualitative

classification. That is, they can be measured only in terms of whether the individual items belong to some distinctively different categories, but we cannot quantify or even rank order those categories.

- 2) **Ordinal Scale:** Measurements with ordinal scales are ordered in the sense that higher numbers represent higher values. However, the intervals between the numbers are not necessarily equal. Ordinal variables allow us to rank order the items we measure in terms of which has less and which has more of the quality represented by the variable, but still they do not allow us to say “how much more.”
- 3) **Interval Scale (Cardinal Scale):** On interval measurement scales, one unit on the scale represents the same magnitude on the trait or characteristic being measured across the whole range of the scale. Interval scales do not have a “true” zero point, however, and therefore it is not possible to make statements about how many times higher one score is than another. True interval measurement is somewhere between rare and nonexistent in the behavioral sciences. A good example of an interval scale is the Fahrenheit scale for temperature. Interval variables allow us not only to rank order the items that are measured, but also to quantify and compare the sizes of differences between them.
- 4) **Ratio Scale:** Ratio scales are like interval scales except they have true zero points. Ratio variables are very similar to interval variables. In addition to all the properties of interval variables, they feature an identifiable absolute zero point, thus they allow for statements such as x is two times more than y. A typical example of ratio scales is measure of time or space. Interval scales do not have the ratio property. Most statistical data analysis procedures do not distinguish between the interval and ratio properties of the measurement scales

- **Functions of Statistics**

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

- 1) **Summarise and Describe data:** One summarises and describes the data in order to view data at a glance. If it is nominal or ordinal data, one makes cross-tabulations and graphs; if it is interval or ratio data then z-scores are calculated.
- 2) **Variance and distribution of the data:** In order to measure the spread of the data and knowing its distributions one makes tables and charts and graphs for nominal/ordinal data and histograms with normal curve or box plots with interquartile range for interval/ratio data.
- 3) **Compare groups:** When one has to compare two or more populations then one makes cross-tabulations for nominal/ordinal data and employ testing of hypothesis for continuous/numeric data divided into groups.
- 4) **Identify relationships:** In order to identify relationships in the data, one uses cross- Statistical

Packages tabulations for nominal/ordinal data; calculate correlation coefficient and scatter plot for Interval/ratio data or go for linear regression/ ANOVA for data with one dependent and 2 or more predictor variables.

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

20.3 Use of Statistical Data with Computers

The use of statistical data with computers is given below:

1. **Quantitative Analysis:** Computers, especially through statistical software like SPSS or Excel, perform complex statistical calculations rapidly, allowing sociologists to test theories and identify patterns in large datasets.
2. **Data Visualization:** Software automatically generates tables, charts, and graphs to display distributions, trends, and relationships within statistical data, making findings easier to interpret and communicate.
3. **Large-Scale Data Management:** Computers store, organize, and retrieve large volumes of statistical data efficiently, which would be incredibly time-consuming to manage manually.
4. **Data Analysis and Preprocessing:** Computers help in cleaning, transforming, and preparing raw data for analysis, a crucial step before applying statistical models.
5. **Statistical Modeling and Algorithms:** They are used to develop and apply statistical models, including regression analysis, classification, and clustering, which are foundational to machine learning and AI.
6. **Simulations and Computer-Intensive Methods:** They enable computer-intensive methods like Monte Carlo simulations, which allow for the testing of complex statistical theories and the generation of data to demonstrate principles.
7. **Handling Large Datasets:** Computers provide the power to efficiently manage and manipulate “Big Data” at an “exascale,” allowing for analyses on datasets that were previously unmanageable. 8. **Exploratory Data Analysis (EDA):** Computers have shifted the emphasis in statistics towards EDA, allowing for a more hands-on and dynamic approach to data exploration rather than solely relying on hypothesis testing.

CHECK YOUR PROGRESS-I

Fill in the Blanks

1. Computers have had a much _____ impact on our societies than just schoolwork, communication, and entertainment ·
2. In the _____ computers made a huge difference in TV programs, radio, and movies ·
3. Editing, voice mixing, _____, animation, and graphics— none of these can be done efficiently without the help of computers ·
4. In the business world, companies use computer _____ to store, organize, and analyze data
5. _____ and video conferences are major communication tools for businesses ·
6. Computers also play a vital role in manufacturing - for example, products are _____ with the help of a Computer-Aided Design, or CAD software ·
7. Many manufacturing processes, as well as are governed by computers ·
8. Computers are a _____ part of our lives and have permanently changed our society and will continue to impact us throughout the future

20.4 How Computers Enhance Statistical Analysis

1. **Efficiency:** Computers automate tedious calculations and data management tasks, significantly speeding up the analysis process.
2. **Accessibility:** With personal computing and user-friendly statistical software, advanced statistical techniques are now accessible to a wider range of researchers and analysts.
3. **Complexity:** Computers make it possible to apply advanced and computationally expensive statistical methods that would be impractical or impossible to perform manually.
4. **Insight and Decision-Making:** By processing vast amounts of data and identifying hidden relationships, computers help analysts gain insights and make data-driven .5 decisions.

20.5 Use of Computers in Social Science Research

Social research is an essential field that helps us understand the dynamics of human behavior, culture, societies, and interactions. Traditionally, social research relied heavily on manual processes for data collection, analysis, and interpretation. However, the advent of computers has transformed the way social research is conducted. Computers have made it possible to handle vast amounts of data, apply

sophisticated analysis techniques, and simulate social phenomena more efficiently. This comprehensive article explores the various ways in which computers are used in social research, from data collection to complex statistical modeling, and how they have revolutionized the field.

1 THE ROLE OF COMPUTERS IN DATA COLLECTION

Data collection is the foundation of any social research project, and computers have significantly improved how this data is gathered.

1.1. Online Surveys and Questionnaires

Computers have made online surveys one of the most common methods of data collection in social research. Tools like Google Forms, SurveyMonkey, and Qualtrics enable researchers to create, distribute, and manage surveys easily. These platforms also offer analytics features that allow researchers to track response rates and analyze the data in real time.

Advantages:

- **Cost-Effective:** Online surveys eliminate printing and mailing costs.
- **Reach:** Researchers can collect data from diverse geographical locations.
- **Automation:** Responses are automatically saved in databases, reducing the likelihood of human error.

1.2. COMPUTER-ASSISTED PERSONAL INTERVIEWS (CAPI)

Computers are also used for conducting face-to-face interviews through CAPI systems. In these systems, the interviewer inputs responses directly into a computer or tablet, streamlining the data collection process and improving data accuracy.

Advantages:

- **Real-Time Data Entry:** Reduces the need for manual data entry post-interview.
- **Complex Survey Logic:** CAPI allows the use of branching questions, which adapt to previous answers.

1.3. BIG DATA AND SOCIAL MEDIA MINING

The rise of social media and other digital platforms has introduced new opportunities for data collection. Social media mining, or the process of analyzing data from social networks, is a powerful tool in social research. Researchers use social media platforms like Twitter, Facebook, and Instagram to study user behavior, opinions, and trends on a large scale.

Example: Researchers might analyze tweets related to a political event to assess public sentiment in real time.

2 COMPUTERS IN DATA ANALYSIS AND STATISTICAL COMPUTATION

Data analysis is another area where computers have greatly enhanced the capabilities of social researchers. The ability to process large datasets and apply complex statistical methods is one of the biggest advantages of using computers in social research.

2.1. Statistical Software Packages

Statistical software like **SPSS**, **R**, **Stata**, and **SAS** are essential tools in social research for processing quantitative data. These programs allow researchers to perform advanced analyses, such as regression models, correlation tests, factor analyses, and more.

Example: A researcher studying the relationship between income and educational attainment could use SPSS to run a regression analysis, identifying significant predictors of income variation.

2.2. Qualitative Data Analysis

While quantitative data analysis has always been integral to social research, computers have also made it easier to handle qualitative data, such as interview transcripts or open-ended survey responses. Programs like **NVivo** and **ATLAS.ti** are used for coding and categorizing themes, enabling researchers to find patterns in large volumes of text.

Example: A sociologist might use NVivo to analyze interview transcripts from a focus group discussing gender roles in the workplace.

2.3. Data Visualization Tools

Computers allow researchers to visualize data effectively using tools like **Tableau**, **Google Data Studio**, and **Power BI**. These tools can transform raw data into interactive graphs, charts, and dashboards, making it easier to interpret findings and present them to a wider audience.

Example: A researcher studying migration trends might use Tableau to create an interactive map that visually represents population movements across regions.

3 SIMULATION AND MODELING IN SOCIAL RESEARCH

Simulations and models allow researchers to test hypotheses and predict outcomes by creating virtual environments where different social behaviors can be studied.

3.1. Agent-Based Models (ABM)

Agent-Based Modeling is a powerful tool used in social research to simulate interactions between individuals (agents) within a system. These models can be used to study phenomena like the spread of diseases, voting behavior, or social inequality.

Example: An ABM simulation might be used to model how different education policies impact income inequality over several generations.

3.2. Social Network Analysis (SNA)

Social Network Analysis uses computers to study relationships between individuals or organizations. Tools like **Gephi** and **UCINET** are used to map networks and analyze the interactions between actors, helping researchers understand the dynamics of relationships within a community.

Example: Researchers could use SNA to map how information spreads within a social group and identify key influencers who play a pivotal role in disseminating information.

4 THE USE OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Artificial Intelligence (AI) and Machine Learning (ML) have opened up new possibilities in social research. These technologies can be used to analyze vast datasets, detect patterns, and even predict future trends.

4.1. Natural Language Processing (NLP)

NLP is a branch of AI that helps computers understand and interpret human language. This can be especially useful in analyzing social media posts, interview transcripts, and other text-based data.

Example: NLP could be used to analyze tweets during an election season, categorizing them based on sentiment to predict voter behavior.

4.2. Predictive Analytics

Machine learning algorithms can be used to make predictions based on past data. For instance, researchers can use predictive models to forecast social trends, such as changes in public opinion on a controversial issue.

Example: A researcher might use machine learning to analyze public sentiment about climate change, predicting how opinions will shift in response to new legislation.

5. ETHICAL CONSIDERATIONS IN USING COMPUTERS FOR SOCIAL RESEARCH

While computers have brought many benefits to social research, they also present ethical challenges, especially concerning data privacy and informed consent.

5.1. Data Privacy

When using digital tools to collect and analyze data, researchers must ensure that they are protecting the privacy of their participants. This is especially important when dealing with sensitive information like health data or financial records.

5.2. Informed Consent

It's essential that participants understand how their data will be used, especially when it's collected through online platforms or social media. Researchers must ensure that participants give informed consent before collecting data.

5.3. Algorithmic Bias

AI and machine learning models can sometimes introduce biases, particularly if the data used to train these models is not representative. Researchers must be careful to check for potential biases and correct them.

• OTHER IMPORTANT USE OF COMPUTERS:

Computers are essential in social science research for literature review, data collection (online surveys), data management, analysis (using software like SPSS and R for quantitative data and CAQDAS for qualitative data), and disseminating findings through reports and online platforms. They facilitate the analysis of large datasets, enable simulations, and improve communication among researchers via the internet.

1. LITERATURE REVIEW AND INFORMATION SYNTHESIS

- **Searching:** Computers and the internet provide access to vast databases, academic journals, and online resources, making literature searches more efficient.
- **Synthesis:** Researchers use computers to organize and synthesize information from various sources, helping to build a comprehensive theoretical framework.

2. DATA COLLECTION

- **Online Surveys:** Computers enable the creation and distribution of online questionnaires to large and diverse populations, improving the speed and reach of data collection.
- **Data Entry:** Computers streamline the process of entering data collected via traditional methods (like paper surveys) into a format suitable for analysis.

3. DATA MANAGEMENT

- **Organization:** Computers help manage large and complex datasets, keeping them organized and accessible for analysis.
- **Monitoring:** Software can monitor data for accuracy and completeness during the collection process.

4. DATA ANALYSIS

- **Quantitative Analysis:** Specialized statistical software such as SPSS, R, and Stata is used to perform statistical tests, create charts, and visualize data, making complex findings understandable.

- **Qualitative Analysis:** Computer-Assisted Qualitative Data Analysis Software (CAQDAS) helps researchers analyze textual data like interview transcripts, identify themes, and interpret patterns.
- **Modeling and Simulation:** Computers allow researchers to create models and simulate social phenomena, helping to understand complex dynamics.

5. RESEARCH PUBLICATION AND DISSEMINATION

- **Report Writing:** Word processing software and tools for creating tables and graphs aid in the formal write-up and presentation of research findings.
- **Communication:** The internet facilitates collaboration, data sharing, and communication between researchers, and the publication of research online.

20.6 LET US SUM UP

Thus, to sum up the use of computers in social research has revolutionized the field, making it easier to collect, analyze, and interpret data on a much larger scale than ever before. From online surveys and big data mining to statistical analysis and AI-driven insights, computers have transformed how researchers approach social issues. However, it is essential to balance these technological advances with ethical considerations, ensuring that research is conducted responsibly and with respect for participants' privacy and consent.

Whether you are a seasoned researcher or just beginning your journey in social research, understanding how to use computers effectively can significantly enhance your research capabilities. By leveraging the power of computers, you can gain deeper insights into the complex social dynamics that shape our world.

20.7 GLOSSARY

1. **Application (App):** A software program designed for a specific task, like a word processor or web browser.
2. **Binary Code:** The base-two number system (0s and 1s) that computers use to represent data and instructions.
3. **CPU (Central Processing Unit):** The computer's "brain," responsible for executing instructions and performing calculations.
4. **Database:** An organized collection of information that can be searched and managed.
5. **Email:** Electronic mail, a method of sending and receiving messages over a network.
6. **Hardware:** The physical components of a computer, such as the keyboard, monitor, and hard drive.
7. **Internet:** A global network of interconnected computers.

8. **Motherboard:** The main circuit board of a computer that connects its principal components.
9. **RAM (Random Access Memory):** Temporary, high-speed memory that holds data currently being used by the CPU.
10. **Software:** A set of instructions or programs that tell a computer what to do.
11. **Spam:** Unsolicited and unwanted email or other digital communication.

20.8 SELF-ASSESSMENT QUESTIONS

Q1. How do computers help students with their studies.

Q2. Where do people use computers for research.

Q3. Which professionals use computers to design buildings and machines.

Q4. How are computers used in the creation of new products.

20.9 LESSON END EXERCISE

Q1. Briefly discuss the uses of statistical data in computers.

Q2. How statistical data is helping the researchers in doing research.

20.9 SUGGESTED READINGS

1. Box, George E.P., William G. Hunter, and J. Stuart Hunter (1978). Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building. New York:
2. John Wiley and Sons. Johnson R.A., Wichern D.W. (2003). Business Statistics: Decision making with Data, John Wiley & Sons (Asia) Pvt. Ltd.: Singapore, 2003.
3. Levin R.I., Rubin, D.S. (2001). Statistics for Management. Prentice Hall of India Private Limited: New Delhi.

20.11 ANSWERS TO CHECK YOUR PROGRESS

1. Computers have had a much **greater** impact on our societies than just schoolwork, communication, and entertainment ·
2. In the **media industry**, computers made a huge difference in TV programs, radio, and movies ·
3. Editing, voice mixing, **sound recording**, animation, and graphics— none of these can be done efficiently without the help of computers ·
4. In the business world, companies use **computer information systems** to store, organize, and analyze data -
5. **E-mails** and video conferences are major communication tools for businesses ·
6. Computers also play a vital role in manufacturing, for example, products are **designed** with the help of a Computer-Aided Design, or CAD software ·
7. Many manufacturing processes, as well as **quality control**, are governed by computers ·
8. Computers are a **fundamental** part of our lives and have permanently changed our society and will continue to impact us throughout the future
