

Research Methods and Design

DEGEN531

Edited by:
Dr. Ashish Kumar



LOVELY
PROFESSIONAL
UNIVERSITY



Research Methods and Design

**Edited By
Dr. Ashish Kumar**

CONTENTS

Unit 1:	Basic Introduction to Sheets/Workbook	1
	<i>Dr. Harish Mittu, Lovely Professional University</i>	
Unit 2:	Classification and Tabulation	20
	<i>Mandeep Bhardwaj, Lovely Professional University</i>	
Unit 3:	Data Graphical Presentation and Analysis	38
	<i>Mandeep Bhardwaj, Lovely Professional University</i>	
Unit 4:	Central Tendency	60
	<i>Mandeep Bhardwaj, Lovely Professional University</i>	
Unit 5:	Correlation and Linear Bivariate Regression	76
	<i>Mandeep Bhardwaj, Lovely Professional University</i>	
Unit 6:	Sampling and Sampling Distribution	102
	<i>Mandeep Bhardwaj, Lovely Professional University</i>	
Unit 7:	Design of Experiments	117
	<i>Dr. Harish Mittu, Lovely Professional University</i>	
Unit 8:	Probability	130
	<i>Mandeep Bhardwaj, Lovely Professional University</i>	
Unit 9:	Probability Distribution	144
	<i>Mandeep Bhardwaj, Lovely Professional University</i>	
Unit 10:	Estimation	161
	<i>Dr. Harish Mittu, Lovely Professional University</i>	
Unit 11:	Hypothesis	179
	<i>Dr. Harish Mittu, Lovely Professional University</i>	
Unit 12:	Hypothesis Testing I	190
	<i>Dr. Harish Mittu, Lovely Professional University</i>	
Unit 13:	Hypothesis Testing II	207
	<i>Dr. Harish Mittu, Lovely Professional University</i>	
Unit 14:	Hypothesis Testing III	224
	<i>Dr. Harish Mittu, Lovely Professional University</i>	

Unit 01: Basic Introduction to Sheets/Workbook

CONTENTS

Objectives

Introduction

1.1 Basics of Spreadsheet/Workbook

1.2 Basic Operations

1.3 Excel Options and Add-ins

Summary

Keywords

Self Assessment

Answers for Self Assessment

Review Questions

Further Readings

Objectives

- Understand the basics of MS-excel sheets or workbook.
- Identify the cell, row, and columns in MS-excel sheet.
- Analyze the function of various icons available on MS-excel sheet.
- Use appropriately the options available in MS-excel sheet.
- Understand various basic operations of MS-excel,
- Use various basic operations of MS-excel.
- Explore statistics related options in MS-excel,
- Install add-ins in MS-excel,
- Apply both excel and add-ins options for solving statistical problems.

Introduction

The first unit endeavors to make detailed discussion on basic of MS-excel sheets or workbooks including the topics basics of spreadsheet/workbook, basic operations, excel options and Add-ins. In this unit we must understand what excel is? How it helps an individual for data analysis? A proper understanding of the various functions of excel is essential for a researcher. That is to know about the basic feature of excel. It can be done through the introduction of the basics of excel. Let's start with the basic introduction to MS-excel i.e., sheets/workbook-cell, row, columns, etc.

1.1 Basics of Spreadsheet/Workbook

Sheet or Workbook

A sheet or workbook, in computer, is also known as a spreadsheet and worksheet. Each excel file is called a workbook. A workbook consists of several individual worksheets. The word spreadsheet is used for both workbooks and worksheets. MS excel makes the work of calculation very easy and fast. It is the computerized equivalent of a general/paper ledger (figure 1.1).

Research Methods and Design

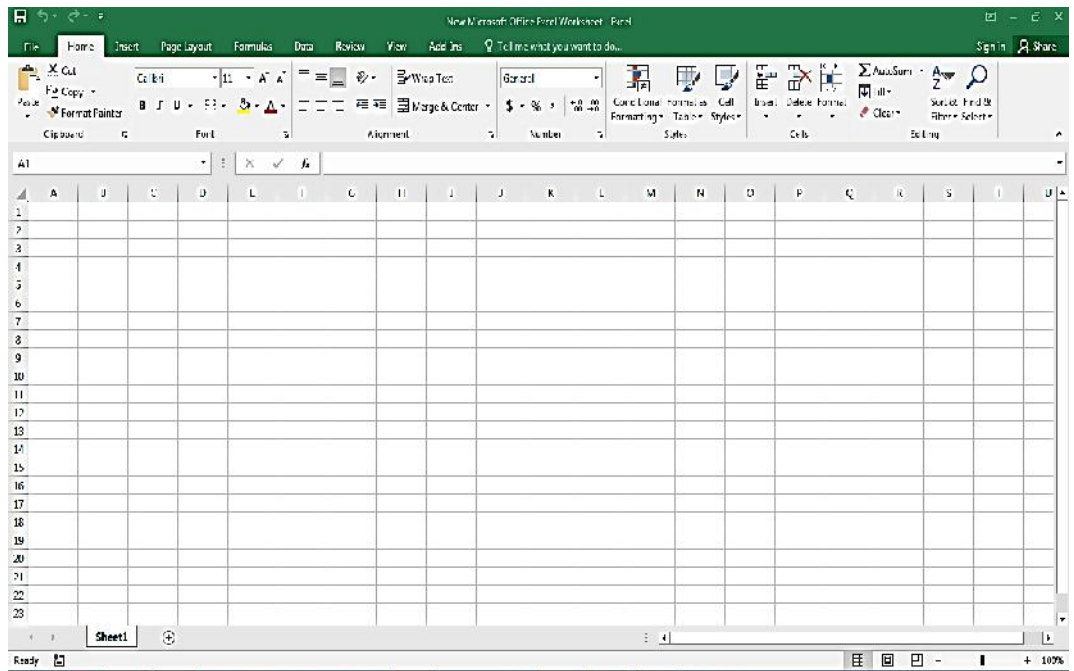


Figure 1.1

A basic spreadsheet consists of a grid of rows and columns of cell. The rows are labeled with numbers 1,2,3, and soon.The columns are labeled with English alphabets A, B,.....and soon, as well as their combinations like AA, AB, and so on (figure 1.2).

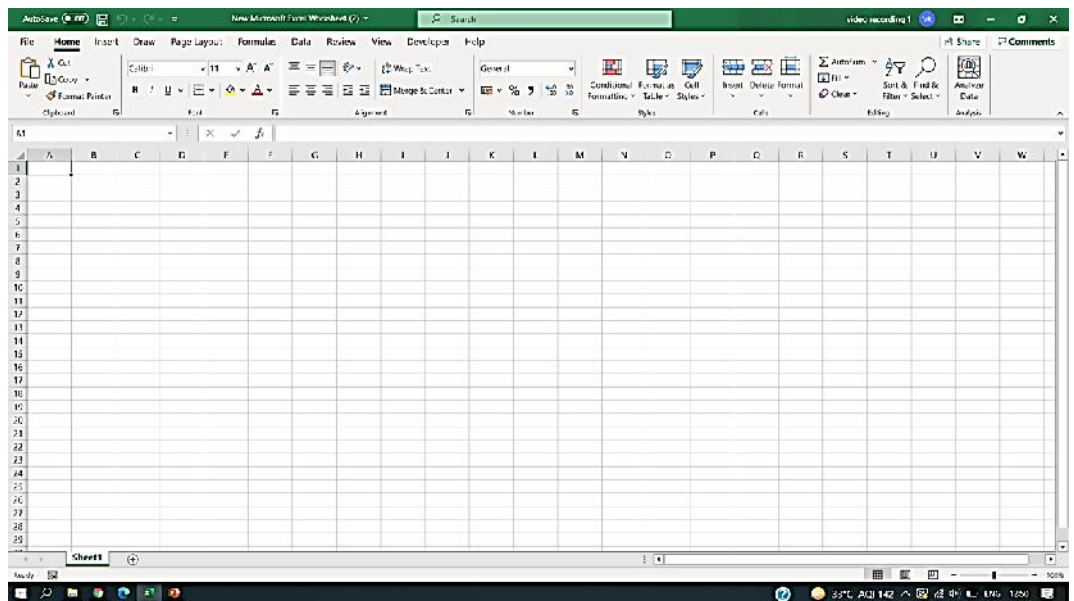


Figure 1.2

There is maximum 16384 columns and 1,048576 rows (figure 1.3).The address of a cell is the intersection of the column and row. For example, A7, D3 and H5 (figure 1.4).

Unit 01: Basic Introduction to Sheets/Workbook

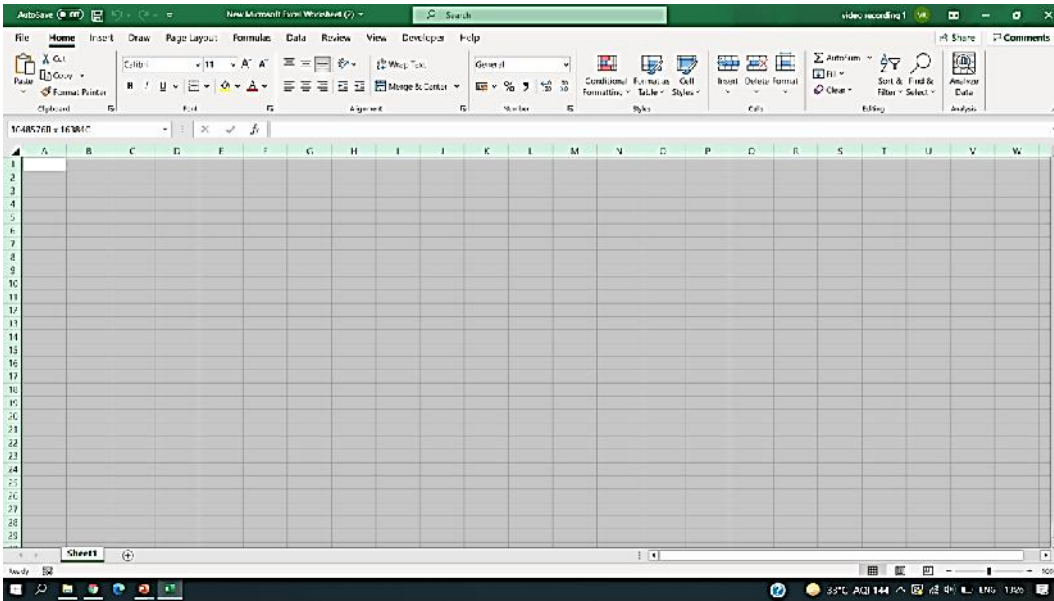


Figure 1.3

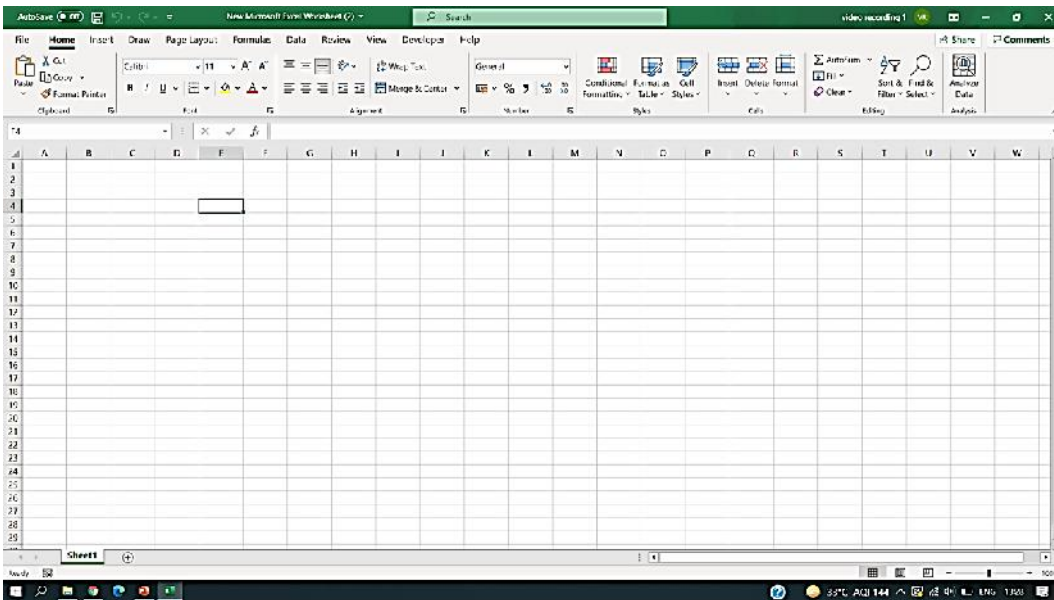


Figure 1.4

MS-Excel-Spreadsheet

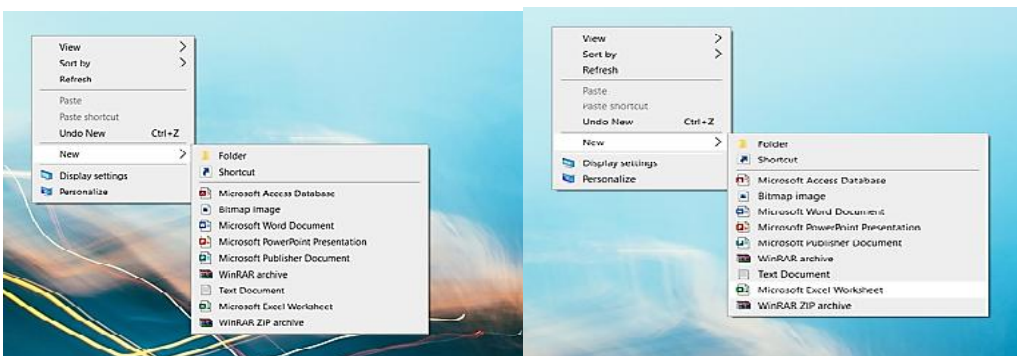


Figure 1.5

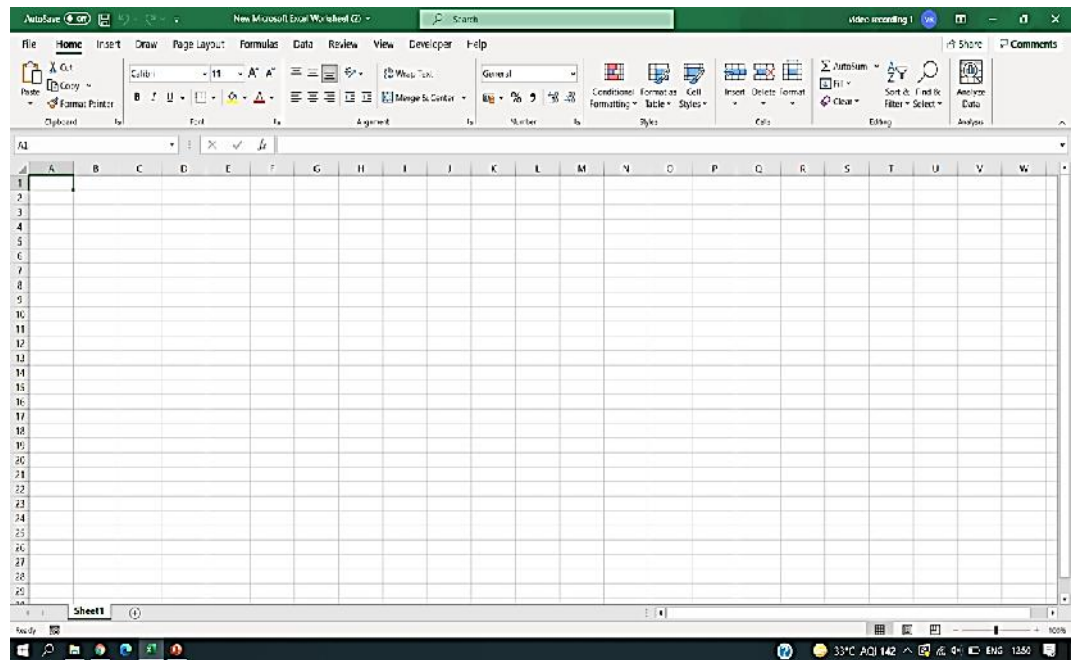


Figure 1.6

Features of Excel Window/Spreadsheet

Quick Access Toolbar

- It is located at the upper left corner of the spreadsheet,
- It provides a set of icons. This set of icons provide shortcuts to frequently used commands.
- This toolbar can be customized by selecting the downward pointing arrow to the right of the icons.

Tabs

- Below the quick access tool bar, the row includes tabs. The main excel commands are organized through this row with the help of tabs.
- Tabs includes – File, Home, insert, Page layout, formulas, Data, review, view, and Add ins.

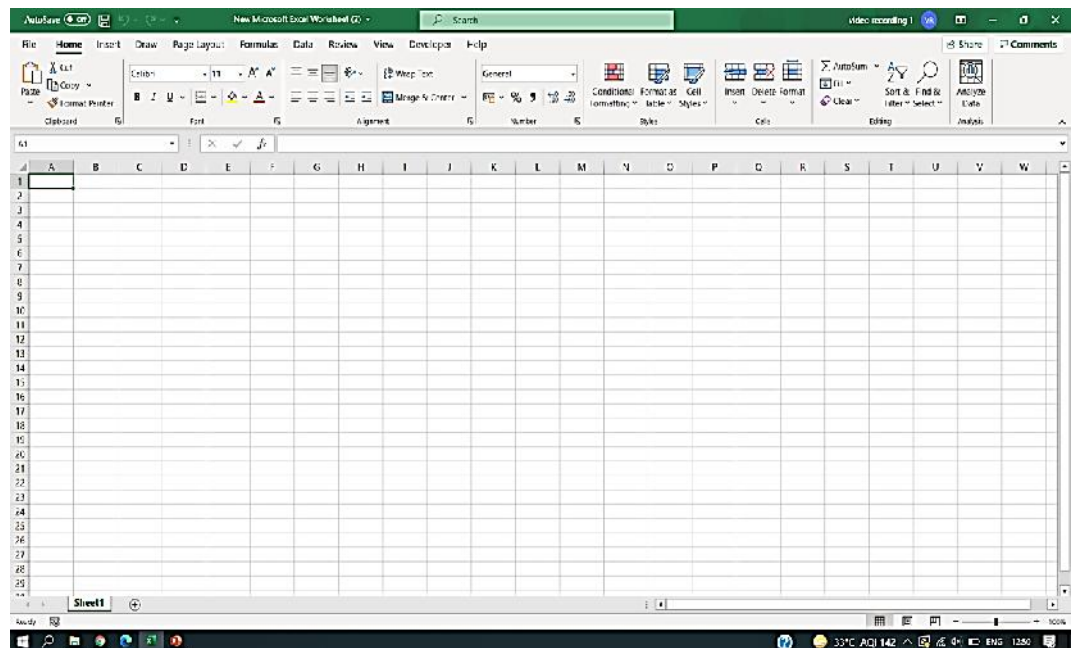


Figure 1.7

File/office Button

- The file option is located at the extreme left corner of the tab row.
- It includes commands like info, new, open, save, save as, print share, export, close, account, and options.

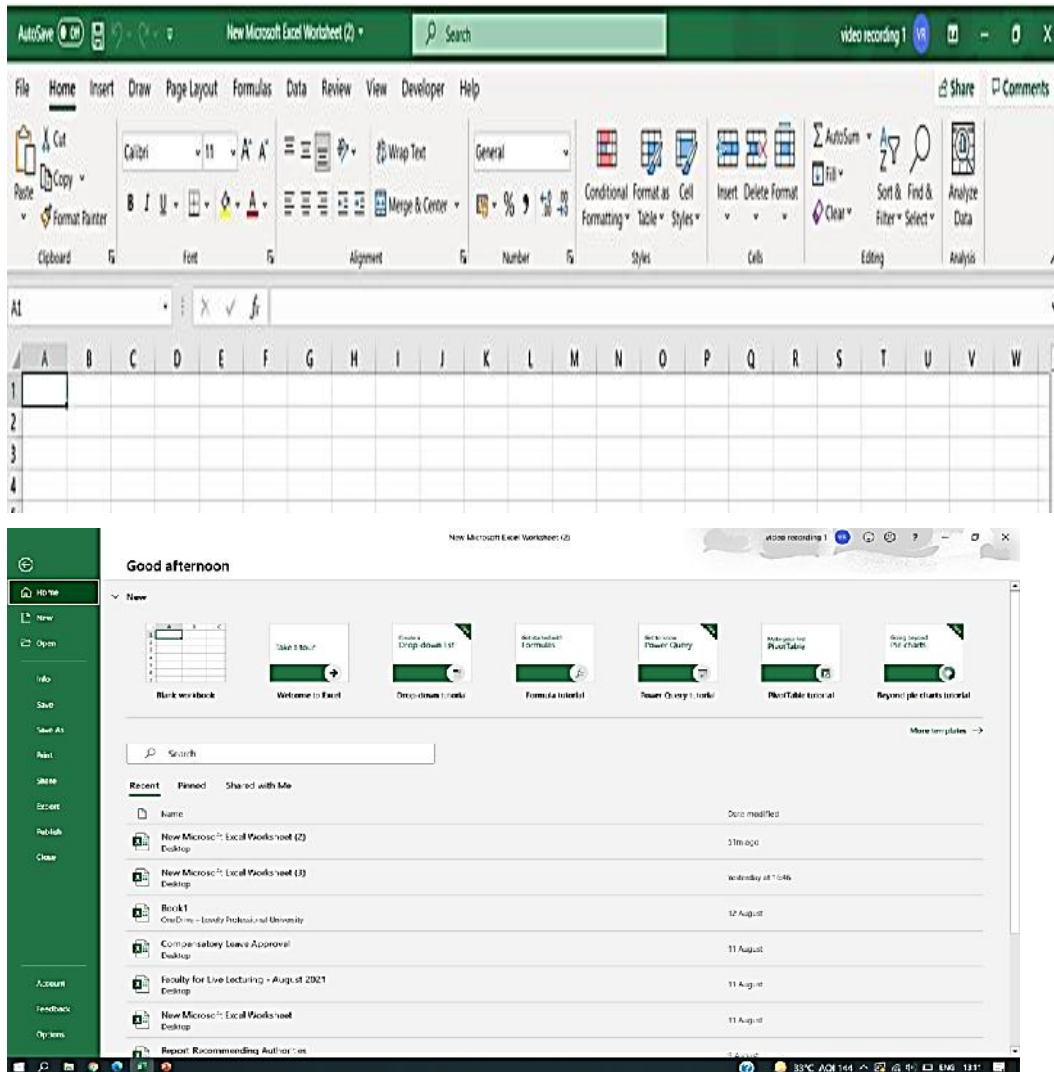


Figure 1.8

Ribbons

Each tab give access to a ribbon below it.

Groups

Various commands are organized into groups

Research Methods and Design

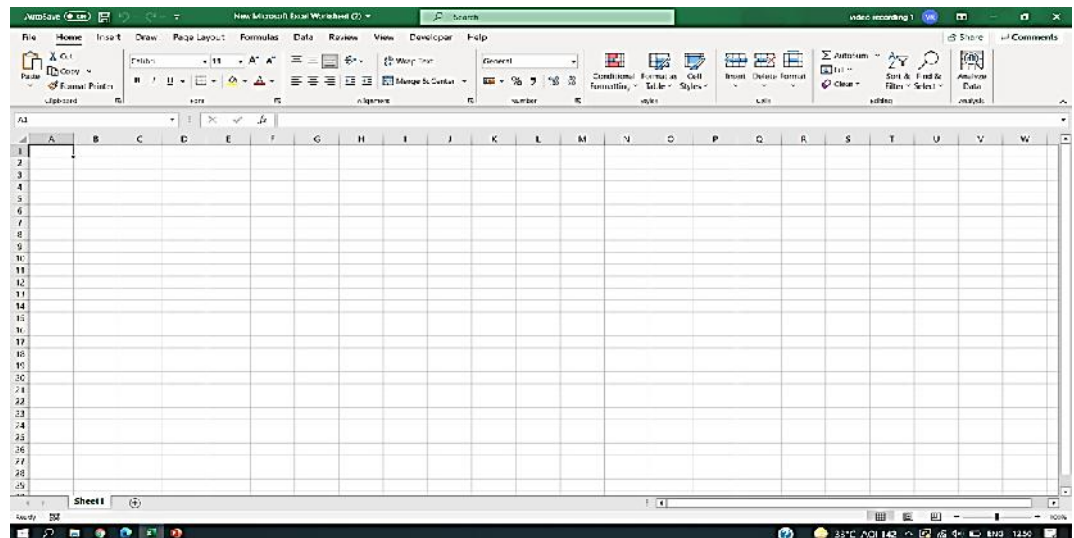


Figure 1.9

Home

Home tab includes the following:

Groups of commands - Clipboard, Font, Alignment, Number, Styles, Cells, Editing.

Clipboard – Cut, copy, Paste, Format Painter

In addition, there is a small downward – pointing arrow icon in case of Clipboard, Font, Alignment and Number.

Clipboard - Click an item to paste.

A format cells window is open in case of Font, Alignment, and Number.

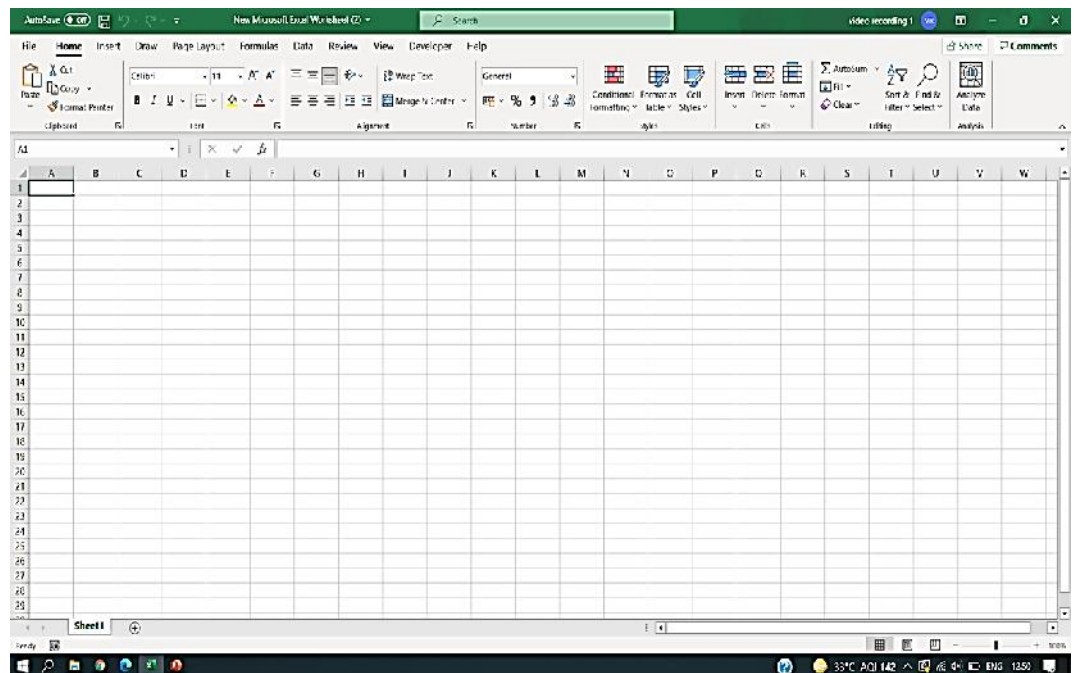


Figure 1.10

Message Area – It is at the bottom left side of the spreadsheet. A message will appear in this area, when excel performs operations (lengthy calculations), giving information on the progress of the procedure/operation.

Scroll Bars – These scroll bars help the individual to change the display portion of the spreadsheet on the screen.

Unit 01: Basic Introduction to Sheets/Workbook

Sheet Tabs – These tabs allow the individual to select a worksheet to display. By doing this an individual can move from one sheet to another within the workbook.

Tab-Scrolling Buttons – The small triangles at the bottom left side of the spreadsheet represent the tab scrolling button. These triangles are used to display different sheets in a workbook when not all the tabs are visible at once/ together.

Name Box – The name box displays the address of the cell as well as the list of any named ranges where the cursor is located.

Formula Bar – The formula bar displays the content of the cell (whether a number, formula, text, etc. where the cursor is located. This is generally the area in which an individual enters information in the cell.

Mouse Cursor – The location of the cursor is shown with and open cross symbol.

Cell Cursor – Let us select a cell (when a cell has been selected), then Cell cursor is an outline with a dark boarder. It is colored blue with a dark border as soon as an individual select a range of cells.

Fill Handle – If an individual wants to copy the content of a cell into the adjacent cell/cells then the cell border, which is a cross can be selected by placing the cursor at the lower right-hand corner of the cell border. The change of mouse cursor to a darken cross ensures the selection of this cross.

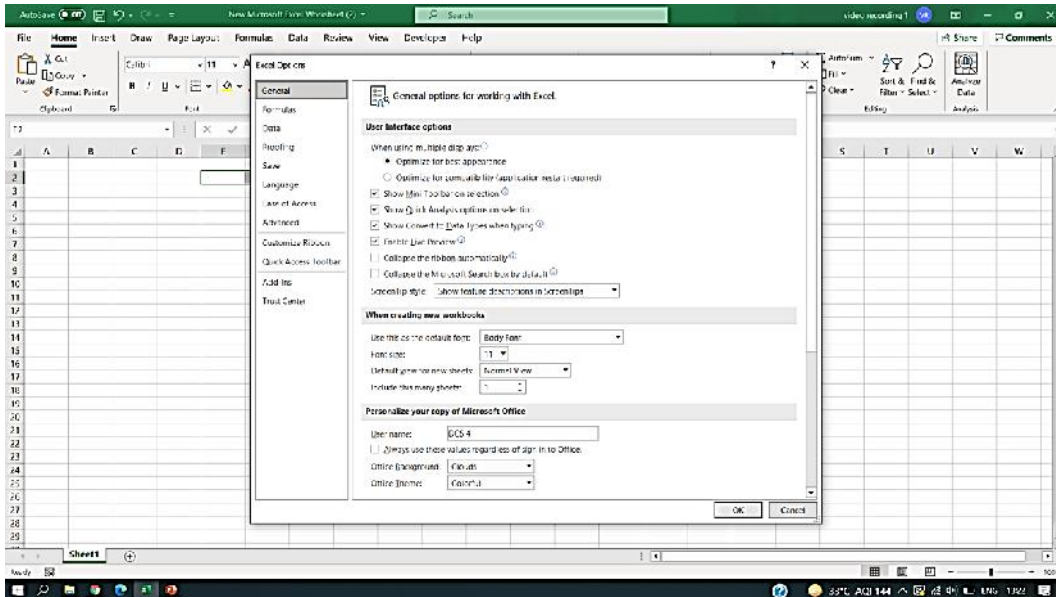


Figure 1.11

You can control the look and behavior of your spreadsheet by setting certain parameters with the 'option' available at last in file tab/office button.

1.2 Basic Operations

The basic operations in MS-Excel are enlisted below:

- Cut, Copy and Paste
- Insert a Column and Row
- Delete a Column and Row
- Clear Contents
- Format Cell/Column/Row
- Adjust Column Width and Row Height
- Hide and Unhide Column/Row
- Sorting

- Arithmetic Precedence
- Entering and Display Formulas

Column or Row

An individual can perform the following commands in column(s) and row(s) of MS-excel:

Cut, copy, paste, paste special, insert, delete, clear content, format cell, column/row width, hide, and unhide

Cut, Copy and Paste - Cell(s), Column(s) and Row(s)

For the command of cut, copy, and paste; an individual must follow the following steps in MS-excel:

Select the cell(s), column(s) or row(s), rightclick, cut and paste, copy and paste, and special paste.

Insert a Column/Row

For the command of insert a column or row, an individual must follow the following steps in MS-excel:

Column/Row header, rightclick, and left click on insert.

Deleting a Column/Row

For the command of delete a column or row, an individual must follow the following steps in MS-excel:

Column/Row header, rightclick, and leftclick on delete.

Clear Contents

For the command of clear content, an individual must follow the following steps in MS-excel:

Column/Row header, right-click, and leftclick on clear contents.

Format - Cell/Column/Row

For the command of format-cell or column or row, an individual must follow the following steps in MS-excel:

Click and drag, right-click, format cells, number tab, and category.

Adjust Column Width/Row Height

For the command of adjust column width or row height, an individual must follow the following steps in MS-excel:

Column/Row header, rightclick, leftclick on column width/row height, and enter a value.

Hide Column/Row

For the command of hide column or row, an individual must follow the following steps in MS-excel:

Column/Row header, rightclick, and leftclick on hide.

Unhide Column/Row

For the command of unhide column or row, an individual must follow the following steps in MS-excel:

Column/Row header, right-click, left and right columns - hidden column or above and below rows - hidden row, and leftclick on unhide.

Sorting

For the command of sorting, an individual must follow the following steps in MS-excel:

Select the data, home, editing group, sort, and sort smallest to largest or sort largest to smallest.

Column, Sort on, Sort by, and Order.

Arithmetic Precedence

Microsoft Excel follows the rules of arithmetic precedence when evaluating formulas. Symbols and their descriptions are given in the following table 1.1

Table 1.1

Symbol	Description
()	Operations enclosed in parentheses are evaluated first; nested parentheses are evaluated from the inside out
^	Exponentiation
* and /	Multiplication and division, evaluated from left to right
+ and -	Addition and subtraction, evaluated from left to right

Entering Formulas

For entering the formulas, an individual must follow the following steps in MS-excel:

Click on the cell, type the formula, and start with.

Displaying Formulas in the Worksheet

For displaying the formulas, an individual must follow the following steps in MS-excel:

CTRL Key, Press the Key (~) - Display Formula, and Press the Key (=) - Display Numerical.

1.3 Excel Options and Add-ins

File, Home, Insert, Draw, Formulae, Data, etc. are the options available in excel. An individual can use excel and add-ins for drawing graphs and calculating descriptive (mean, median, mode etc.) and inferential statistics (ANOVA, regression etc.).

In insert option, there is an option of charts for presenting data graphically with the help of bar chart, pie chart, histogram, line chart etc.

Follow the procedure shown in following figure 1.12 to draw bar graph:

Research Methods and Design

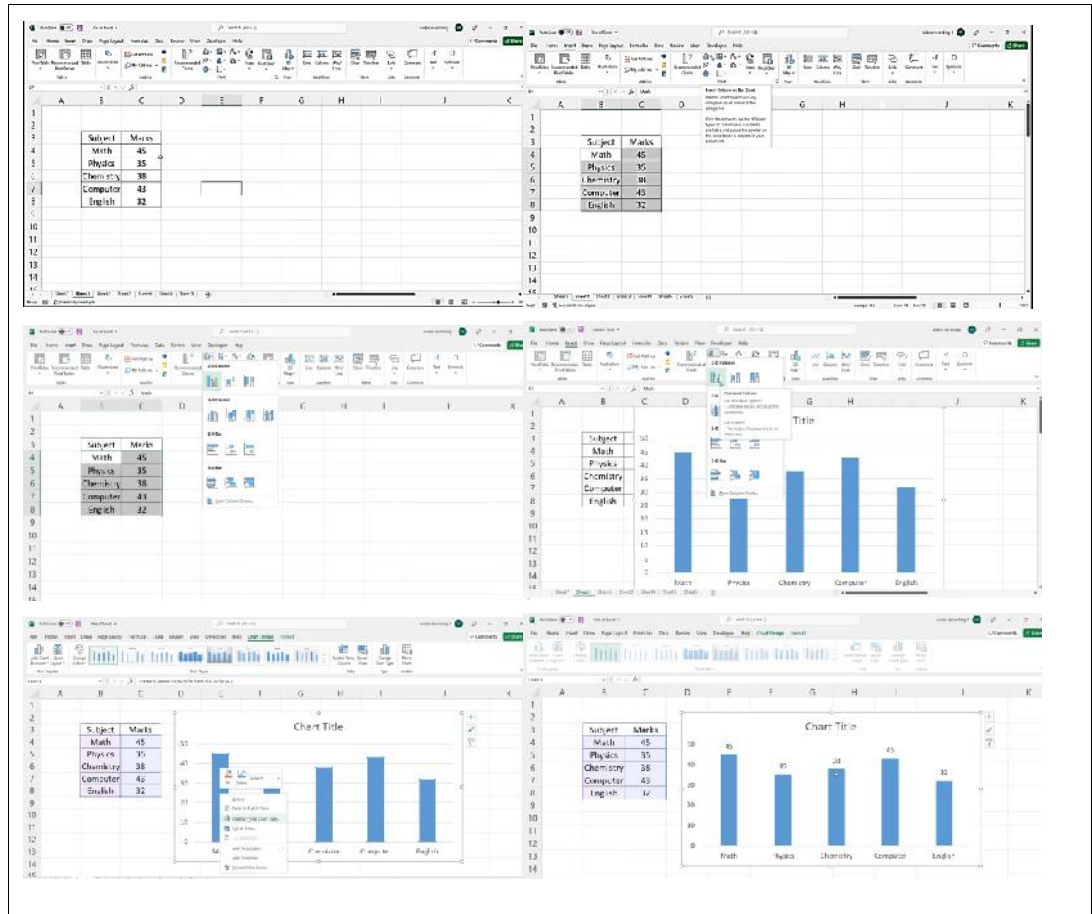


Figure 1.12

Follow the procedure shown in following figure 1.13 to draw bar graph:

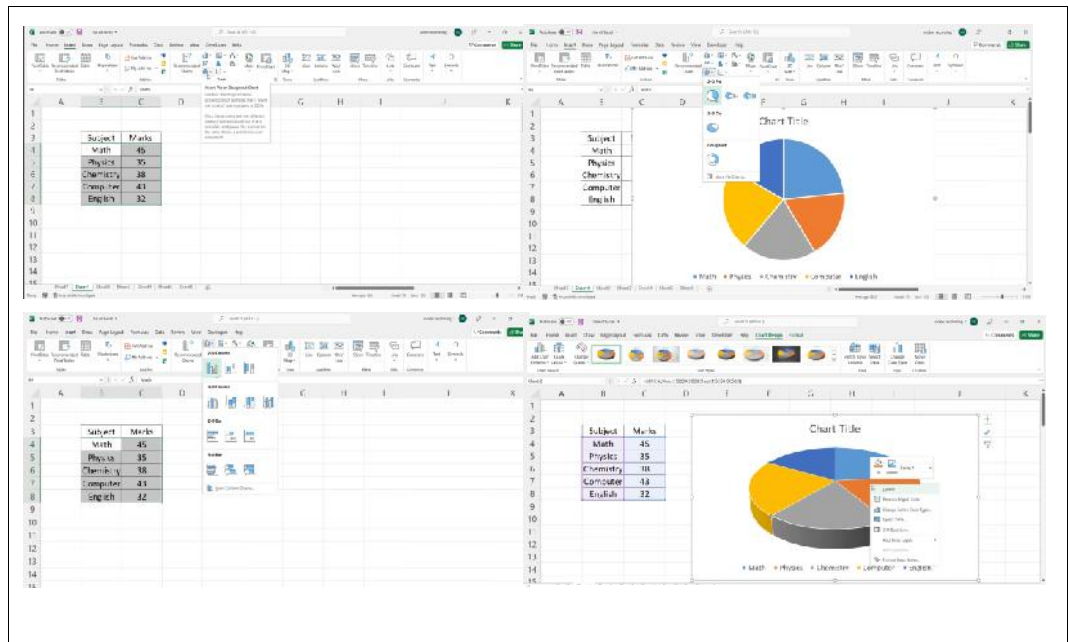


Figure 1.13

Follow the procedure shown in following figure 1.14 to draw line graph:

Unit 01: Basic Introduction to Sheets/Workbook

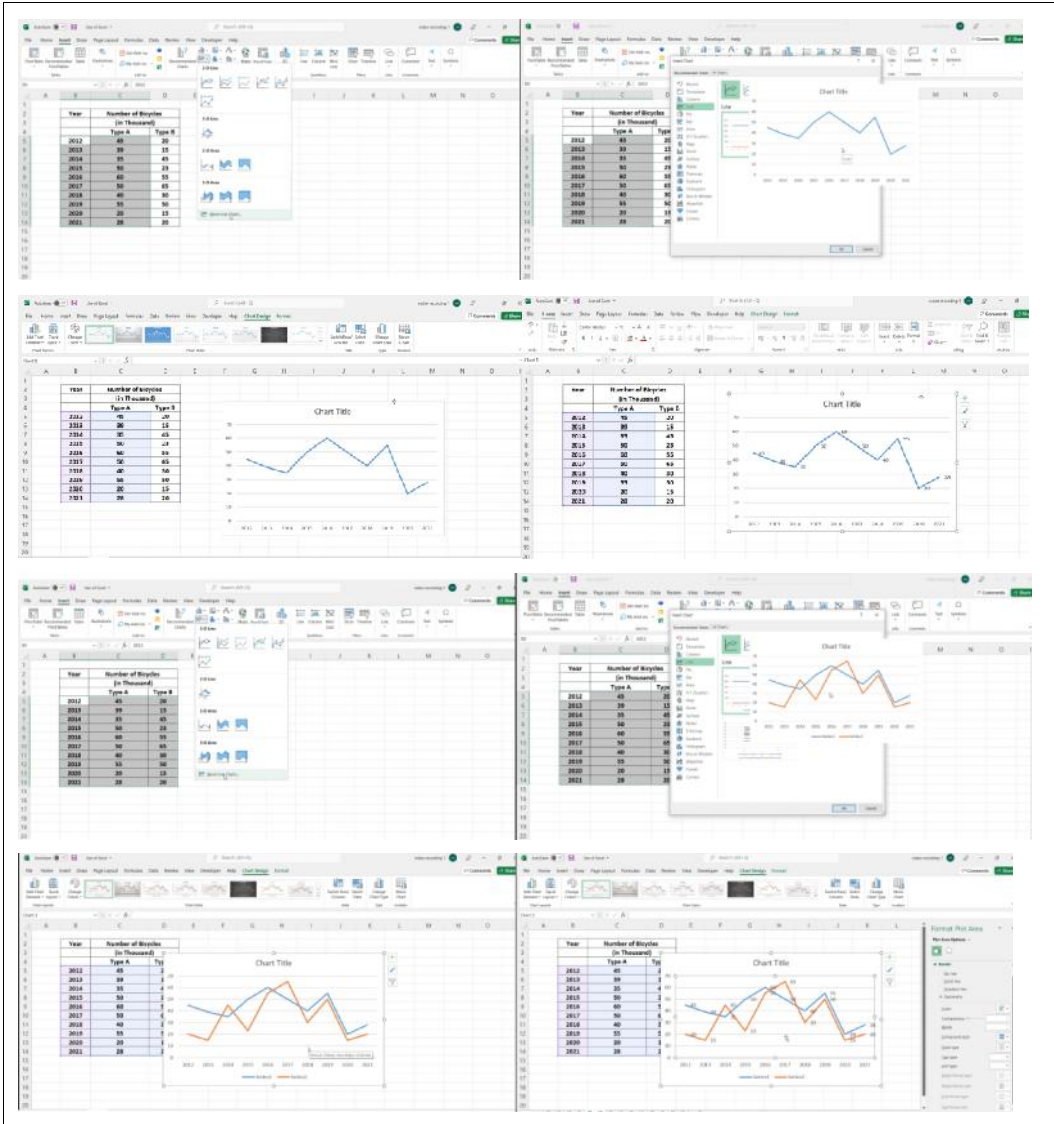
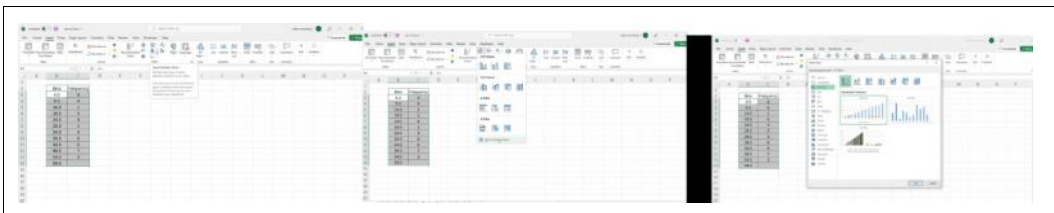


Figure 1.14

Follow the procedure shown in following figure 1.15 to draw histogram graph:



Research Methods and Design

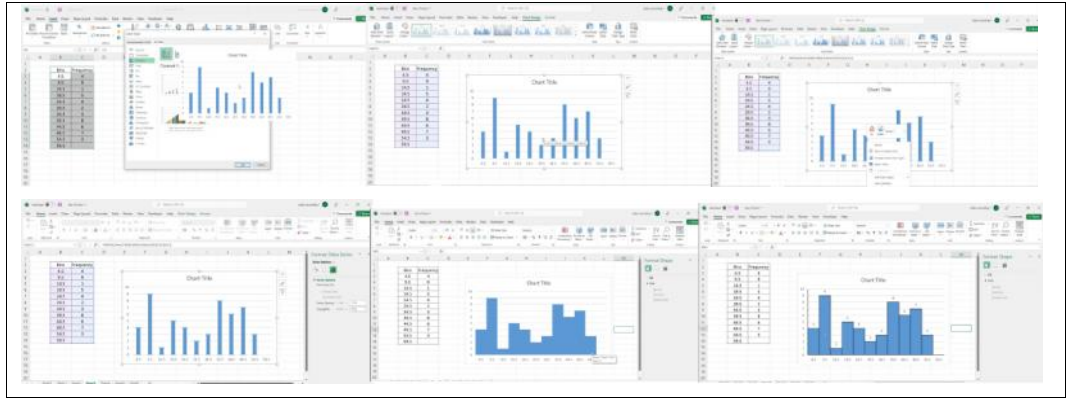


Figure 1.15

Follow the procedure shown in following figure 1.16 to calculate count, min., max., mean, median, mode, and correlation:

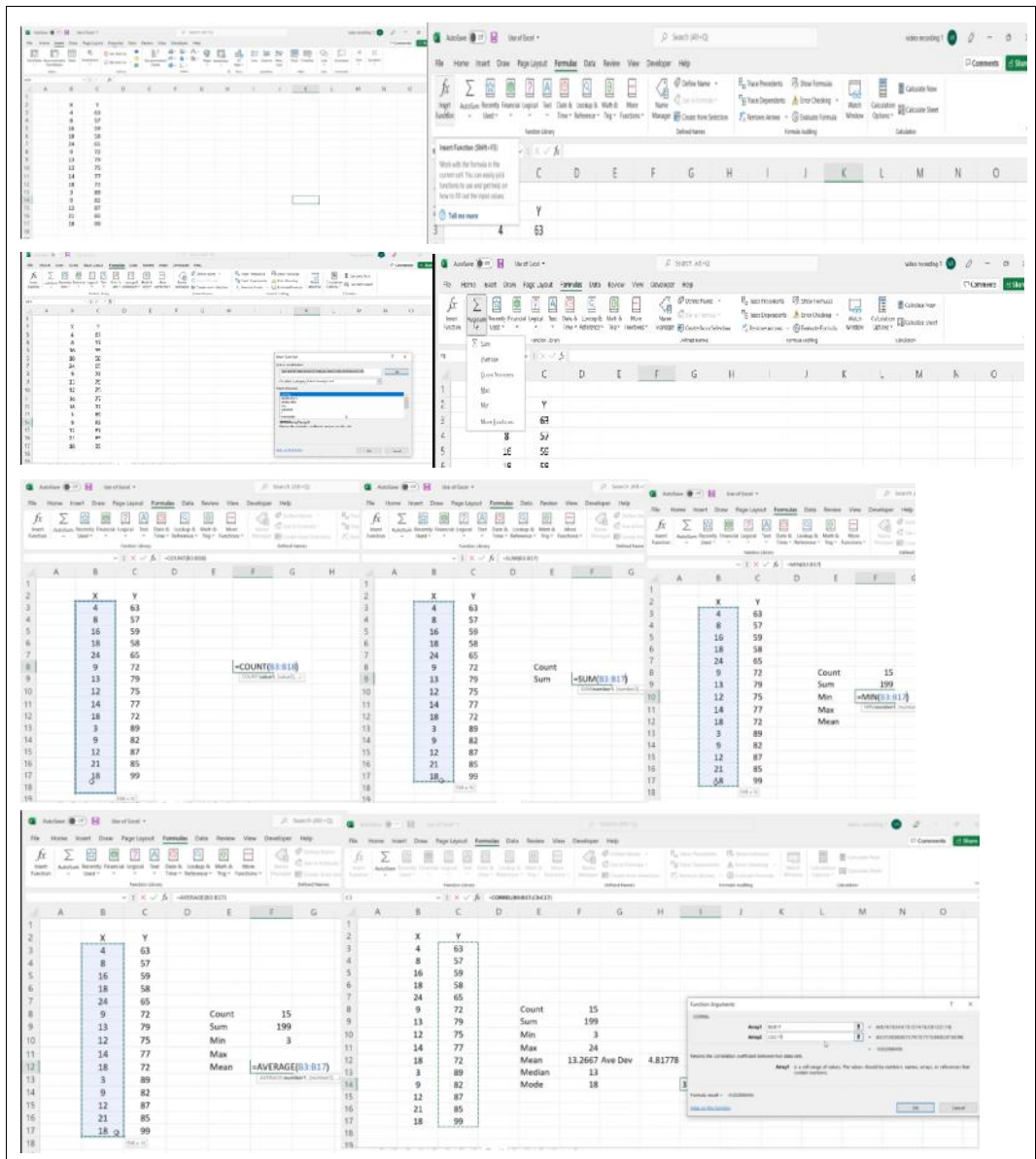


Figure 1.16

Follow the procedure shown in following figure 1.17 to activate add-ins in MS-excel:

Unit 01: Basic Introduction to Sheets/Workbook

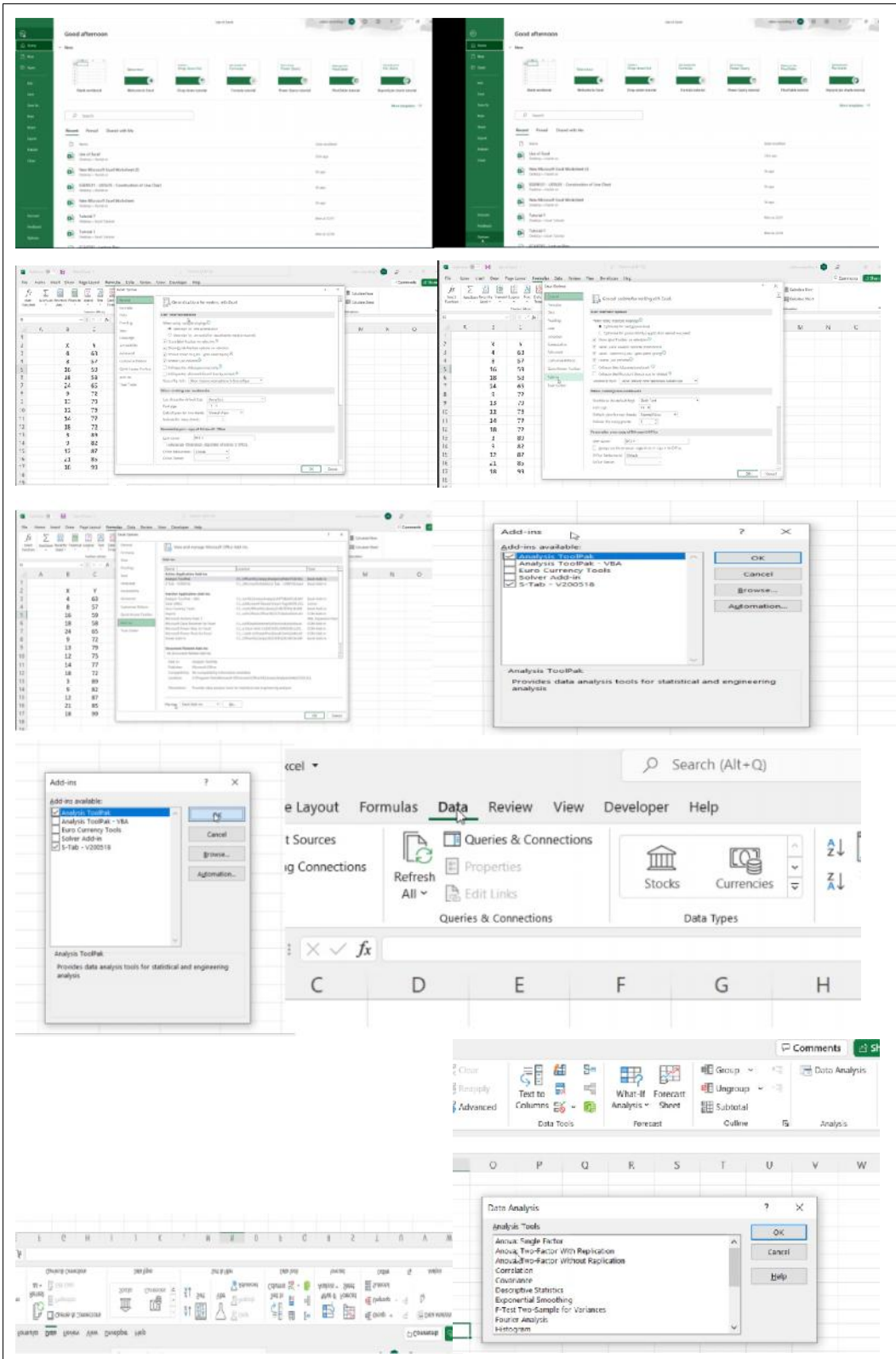


Figure 1.17

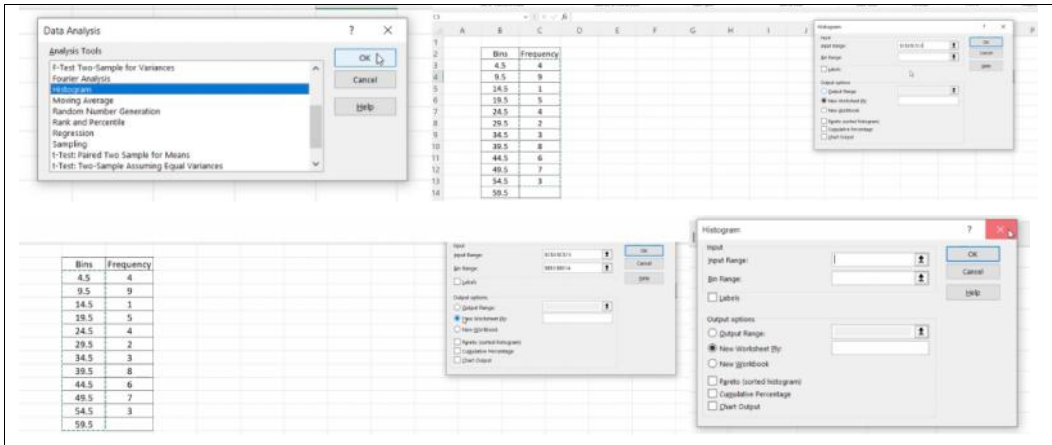
Follow the procedure shown in following figure 1.18 to calculate descriptive statistics, correlation, and draw histogram with the help of add-inns (data analysis) in MS-excel:

The following steps illustrate the process of calculating a correlation coefficient in Excel:

- Data Entry:** A dataset with two columns, X and Y, is entered into a spreadsheet.
- Data Analysis Tool Selection:** The 'Data Analysis' tool is accessed from the 'Data Tools' ribbon.
- Descriptive Statistics Configuration:** The 'Descriptive Statistics' dialog box is configured with the following settings:
 - Input Range: \$C\$3:\$C\$17
 - Grouped By: Columns
 - Labels in first row:
 - Output options:
 - Output Range: \$E\$3:\$I\$12
 - Summary statistics:
 - Confidence Level for Mean: 95%
- Data Analysis Tool Selection:** The 'Data Analysis' dialog box is configured with:
 - Analysis Tools: Correlation
- Correlation Dialog Box Configuration:** The 'Correlation' dialog box is configured with:
 - Input Range: \$B\$3:\$C\$17
 - Grouped By: Columns
 - Labels in first row:
 - Output options:
 - Output Range: (empty)
 - New Worksheet By: New Worksheet
- Output:** The results are displayed in a table:

	Column 1	Column 2
Column 1	1	
Column 2	-0.02299	1

Unit 01: Basic Introduction to Sheets/Workbook



Summary

MS excel makes the work of calculation very easy and fast. It is the computerized equivalent of a general/paper ledger. Basics of MS-Excel.

The Excel window/spreadsheet includes quick access toolbar, tabs, file/office button, ribbons, groups, home etc. Home tab includes groups of commands, clipboard, message area, scroll bars, sheet tabs, tab-scrolling buttons, name box, formula bar, mouse cursor, cell cursor etc.

If an individual wants to copy the content of a cell into the adjacent cell/cells then the cell border, which is a cross can be selected by placing the cursor at the lower right-hand corner of the cell border with the help of fill handle. The change of mouse cursor to a darkened cross ensures the selection of this cross.

You can control the look and behavior of your spreadsheet by setting certain parameters with the 'option' available at last in file tab/office button.

The Basic Operations of MS-Excel includes -cut, copy and paste; insert and delete a column/row; clear and format the content of cell/column/row; adjust column width and row height; hide and unhide column/row; sort the data; arithmetic precedence; and entering and display formulas.

In MS-excel, the option to make a graph is in insert tab, the option of statistical formula's is in formulas tab (insert function and auto sum). An individual can activate the add-ins option in tab data (data analysis) of excel to draw histogram and apply various statistical techniques.

Keywords

A **sheet** or **workbook**, in computer, is also known as a spreadsheet and worksheet.

A **quick access toolbar** provides a set of icons which includes shortcuts to frequently used commands.

Tabs include main excel commands like file, home, insert, page layout, formulas, data, review, view, and add ins.

File/office Button includes commands like info, new, open, save, save as, print share, export, close, account, and options.

Ribbons are there below each tab.

Groups organize various commands.

Message Area is at the bottom left side of the spreadsheet.

Scroll Bars help the individual to change the display portion of the spreadsheet on the screen.

Sheet Tabs allow the individual to select a worksheet to display.

Tab-Scrolling Buttons are the small triangles at the bottom left side of the spreadsheet

Name Box displays the address of the cell as well as the list of any named ranges where the cursor is located.

Formula Bar displays the content of the cell where the cursor is located.

Mouse Cursor's location is shown with an open cross symbol.

Cell Cursor is an outline with a dark border.

Self Assessment

1. There are ____ columns in a spreadsheet.
 - A. 16382
 - B. 16383
 - C. 16384
 - D. 16385

2. The portion of the spreadsheet shown in the figure below includes



- A. Customize Quick Access Toolbar
 - B. Office button
 - C. Formula bar
 - D. Tabs
-
3. _____ allows the individual to select a worksheet to display.
 - A. Sheet Tabs
 - B. Scroll Bars
 - C. Tabs
 - D. Tab-Scrolling Buttons

 4. Borders command is available in
 - A. Alignment
 - B. Editing
 - C. Font group
 - D. Styles

 5. _____ gives access to _____ in which commands are organized to _____.
 - A. Groups; Ribbon; Tab
 - B. Ribbon; Tab; Groups
 - C. Tab; Ribbon; Groups
 - D. Ribbon; Groups; Tab

 6. _____ allows an individual to choose what we want to paste in cells.
 - A. Paste
 - B. Copy
 - C. Copy and Paste
 - D. Paste Special

7. Which of the following symbol is used for multiplication in excel?
- A. x
 - B. /
 - C. *
 - D. ^
8. Sort and filter option is available in
- A. group editing of the home tab
 - B. group arrange of page layout tab
 - C. group home of editing tab
 - D. group page layout of the home tab
9. Which of the following is correct?
- A. (A+B)
 - B. = (A+B)
 - C. = (A plus B)
 - D. = Add (A and B)
10. Which of the following command is correct for displaying formula in Ms-Excel?
- A. Shift and ~
 - B. ~
 - C. Ctrl
 - D. Ctrl and ~
11. Bar graph is available in
- A. chart group of the insert tab
 - B. graph group of the insert tab
 - C. insert chart group of formula tab
 - D. graph group of the data tab
12. The correct expression to calculate count number of observations in MS-Excel is
- A. Count(First value: Last value)
 - B. = (First value: Last value)
 - C. =count(First value: Last value)
 - D. =counting(First value: Last value)
13. Bins represent
- A. Frequency
 - B. Class-intervals
 - C. Cumulative frequency
 - D. Mid-points of class-intervals

14. When an individual tries to construct a histogram through chart group of insert tab, then s/he
- first, obtain a line graph
 - first, obtain a bar graph
 - directly obtain histogram
 - first, obtain bars without gaps
15. Which of the following is the correct option to draw a histogram in MS-Excel?
- Charts only
 - Add-ins only
 - Charts as well as Add-ins
 - Charts as well as Data Analysis

Answers for Self Assessment

- | | | | | |
|-------|-------|-------|-------|-------|
| 1. C | 2. D | 3. A | 4. C | 5. C |
| 6. D | 7. C | 8. A | 9. B | 10. D |
| 11. A | 12. C | 13. B | 14. B | 15. D |

Review Questions

- Discuss the functions of different tabs of MS-excel.
- Construct bar graph by selecting appropriate data in MS-excel.
- Construct histogram by selecting appropriate data in MS-excel.
- Construct line graph by selecting appropriate data in MS-excel.
- Construct line graph by selecting appropriate data in MS-excel.
- Construct pie chart by selecting appropriate data in MS-excel.
- Calculate mean, median and mode by selecting appropriate data with help of analysis tool in MS-excel.
- Construct histogram by selecting appropriate data with help of analysis tool in MS-excel.
- Compute correlation by selecting appropriate data with help of analysis tool in MS-excel.
- Compute descriptive statistics by selecting appropriate data with help of analysis tool in MS-excel.

**Further Readings**

- Jacob, K., 2007, Microsoft Office Excel 2007, The L Line, The Express Line of Learning, John Wiley and Sons, New York.
- Reding, E. and Wermers, L., 2007, Microsoft Office Excel 2007, Illustrated Introductory, MA Course Technology, Cambridge.

**Web Links**

<https://ncert.nic.in/textbook/pdf/lca102.pdf>

<https://nios.ac.in/media/documents/sec229new/Lesson6.pdf>

<https://courses.lumenlearning.com/santaana-informationsystems/chapter/unit-1-excel-fundamental/>

http://182.18.165.51/Fac_File/STUDY180@770517.pdf

<https://corporatfinanceinstitute.com/resources/excel/study/basic-excel-formulas-beginners/>

<https://trumpexcel.com/excel-add-in/>

<https://www.epa.gov/sites/default/files/2015-06/documents/UsingExcelREV.pdf>

<https://www.chem.purdue.edu/gchelp/tools/operate.html>

<https://www.sgul.ac.uk/about/our-professional-services/information-services/library/documents/training-manuals/Excel-Fundamentals-Manual.pdf>

<https://excelchamps.com/blog/tips/>

<https://adminfinance.umw.edu/tess/files/2013/06/Excel-Manual1.pdf>

https://ccsuniversity.ac.in/bridge-library/pdf/DHA_Shikha_BHI_204_Unit4.pdf

<https://www.gacbe.ac.in/pdf/ematerial/18BCS5EL-U5.pdf>

https://ptgmedia.pearsoncmg.com/images/9781509306190/samplepages/9781509306190_Sample.pdf

Unit 02: Classification and Tabulation

CONTENTS

Objectives

Introduction

2.1 Types of Variables

2.2 Classification of Data

2.3 Tabulation of Data

2.4 Constructing a Frequency Distribution

2.5 Cumulative Frequency Distribution

Summary

Keywords

Self Assessment

Answers for Self Assessment

Review Questions

Further Readings

Objectives

- Define the concept of data
- Understand the concept and meaning of classification of data
- Understand the concept and meaning of tabulation of data
- Understand the concept and meaning of frequency distribution
- Construct ungrouped and grouped frequency distributions for given data

Introduction

Statistics is the art of collecting, organizing, presenting, analysing and interpreting data to make informed decisions. It is the science of aggregating the facts and arranging them in a meaningful manner. Data is the backbone of statistics as the analysis is carried out on data. There are two types of Statistics:

Descriptive Statistics

Masses of unorganized data – such as the census of population, the weekly earnings of thousands of computer programmers, and the individual responses of 2,000 registered voters regarding their choice for president of the United States – are of little value as is. However, descriptive statistics can be used to organize data into a meaningful form. Descriptive statistics is the method of organizing, summarizing, and presenting data in an informative way.

Inferential Statistics

Sometimes we must make decisions based on a limited set of data. For example, we would like to know the operating characteristics, such as fuel efficiency measured by miles per gallon, of sport utility vehicles (SUVs) currently in use. If we spent a lot of time, money, and effort, all the owners

POPULATION The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest.

of SUVs could be surveyed. In this case, our goal would be to survey the population of SUV owners. Inferential statistics is the method used to estimate a property of a population on the basis of a sample.

SAMPLE a portion, or part, of the population of interest.

2.1 Types of Variables

There are two basic types of variables: (1) qualitative and (2) quantitative (see fig. 2.1). When an object or individual is observed and recorded as a non-numeric characteristic, it is a qualitative variable or an attribute. Examples of qualitative variables are gender, beverage preference, type of vehicle owned, state of birth, and eye color. When a variable is qualitative, we usually count the number of observations for each category and determine what per cent is in each category. For example, if we observe variable eye color, what per cent of the population has blue eyes and what per cent has brown eyes? If the variable is a type of vehicle, what per cent of the total number of cars sold last month were SUVs? Qualitative variables are often summarized in charts and bar graphs

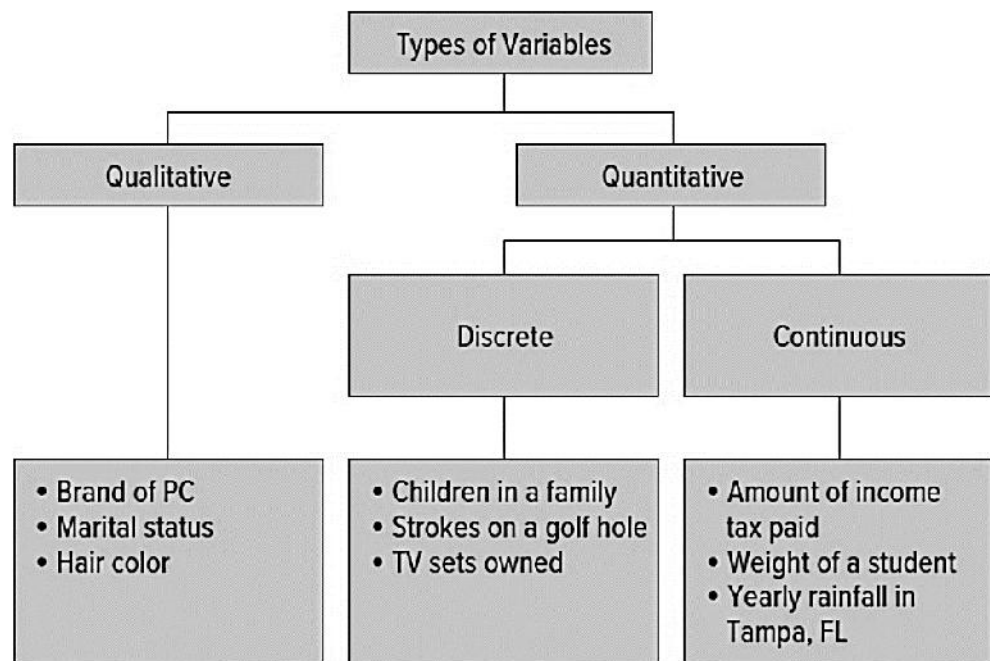


Fig. 2.1 Types of Variables

Source: Basic Statistics for Business and Economics by Lind ET. al (2022)

When a variable can be reported numerically, it is called a quantitative variable. Examples of quantitative variables are the balance in your checking account, the number of gigabytes of data used on your cell phone plan last month, the life of a car battery (such as 42 months), and the number of people employed by a company.

Quantitative variables are either discrete or continuous. Discrete variables can assume only certain values, and there are “gaps” between the values. Examples of discrete variables are the number of bedrooms in a house (1, 2, 3, 4, etc.), the number of cars (326, 421, etc.) arriving at Exit 5 on Indira Gandhi Airport, New Delhi in an hour, and the number of students in each section of a statistics course. We count, for example, the number of cars arriving at Exit 5 on Indira Gandhi Airport, New Delhi, and we count the number of statistics students in each section. Notice that a home can have 3 or 4 bedrooms, but it cannot have 3.85 bedrooms. Thus, there is a “gap” between possible values. Typically, discrete variables are counted. Observations of a continuous variable can assume any value within a specific range. Examples of continuous variables are the air pressure in a tire and the

weight of a shipment of tomatoes. Grade point average (GPA) is a continuous variable. We could report the GPA of a particular student as 7.675426. The usual practice is to round to 3 places—7.675. Typically, continuous variables result from measuring.

2.2 Classification of Data

Classification of data is the process of arranging data in groups/classes on the basis of certain properties. The classification of statistical data serves the following purposes:

- i. It condenses the raw data into a form suitable for statistical analysis.
- ii. It removes complexities and highlights the features of the data.
- iii. It facilitates comparisons and drawing inferences from the data. For example, if university students in a particular course are divided according to sex, their results can be compared.
- iv. It provides information about the mutual relationships among elements of a data set. For example, based on literacy and the criminal tendency of a group of people, it can be established whether literacy has any impact or not on the criminal tendency.
- v. It helps in statistical analysis by separating elements of the data set into homogeneous groups and hence brings out the points of similarity and dissimilarity.

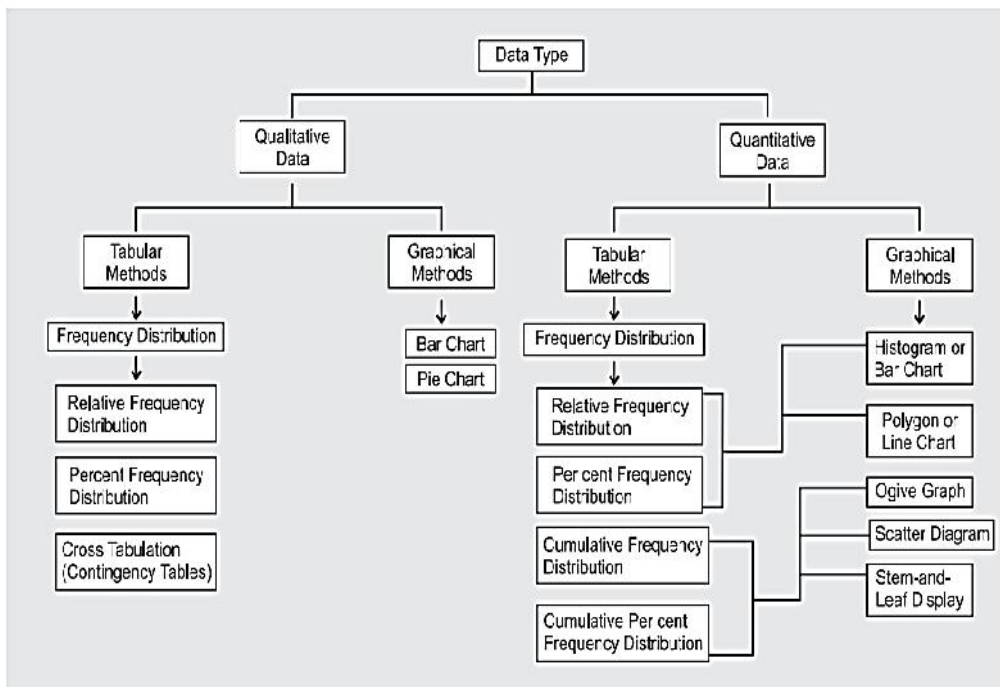


Fig. 2.2 Data Types

Source: Business Statistics, J K Sharma (2019)

Requisites of Ideal Classification

The classification of data is decided after taking into consideration the nature, scope, and purpose of the investigation. However, an ideal classification should have following characteristics:

1. **Unambiguous:** It is necessary that the various classes should be so defined that there is no room for confusion. There must be only one class for each element of the data set. For example, if the population of the country is divided into two classes, say literates and illiterates, and then an exhaustive definition of the terms used would be essential.
2. **Exhaustive and Mutually Exclusive Classes:** Each element of the data set must belong to a class. For this, an extra class can be created with the title 'others' so as to accommodate all the remaining elements of the data set. Each class should be mutually exclusive so that each

element must belong to only one class. For example, the classification of students according to the age: below 25 years and more than 20 years, is not correct because students of age 20 to 25 may belong to both the classes.

3. **Stable:** The classification of a data set into various classes must be done in such a manner that if each time an investigation is conducted, it remains unchanged and hence the results of one investigation may be compared with that of another. For example, the classification of the country's population by a census survey based on occupation suffers from this defect because various occupations are defined in different ways in successive censuses and, as such, these figures are not strictly comparable.
4. **Flexible:** A classification should be flexible so that suitable adjustments can be made in new situations and circumstances. However, flexibility does not mean instability. The data should be divided into a few major classes which must be further subdivided. Ordinarily, there would not be many changes in the major classes. Only small sub-classes may need a change and the classification can thus retain the merit of stability and yet have flexibility.

Basis of Classification

Broadly, data can be classified under following categories:

1. Geographical classification
2. Chronological classification
3. Qualitative classification
4. Quantitative classification

Geographical Classification

In geographical classification, data are classified on the basis of location, region, etc. For example, if we present the data regarding the production of sugarcane or wheat or rice, in view of the four main regions in India, this would be known as geographical classification as given below in Table-2.1. Geographical classification is usually listed in alphabetical order for easy reference. Items may also be listed by size to emphasize the magnitude of the areas under consideration such as ranking the states based on population. Normally, in reference tables, the first approach (that is listing in alphabetical order) is followed.

Table 2.1 Classification of Production of Sugarcane

Region	Production of Sugarcane (in '000 tones)
East	567
South	1458
North	5467
West	2321

Chronological Classification

This is a time-series classification of data which means that the data is arranged as per time. One of the most common examples of time-series data is the census data which is published every 10th year.

Table 2.2 Classification of Population on the Basis of Time

Year	Population (in crores)
1991	89.13 crores

2001	107.5 crores
2011	125.03 crores

Quantitative Classification

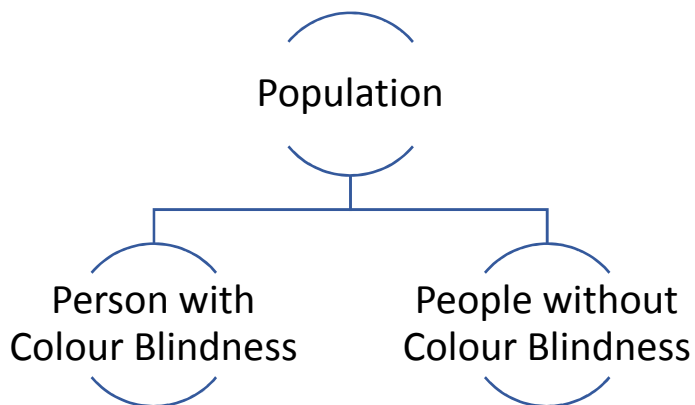
Quantitative classification refers to the classification of data according to some characteristics that can be measured numerically such as height, weight, income, age, sales, etc. For example, the students of MBA can be classified as per their age as follows:

Table 2.3 Quantitative Classification of Students on the Basis of Age

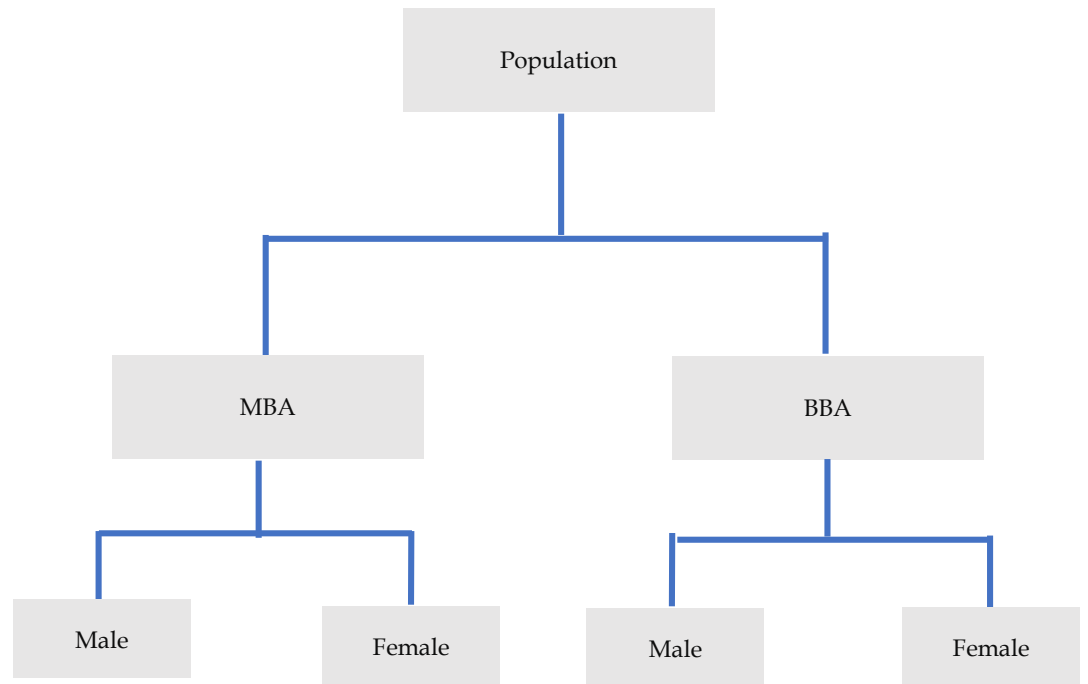
Age	Number of Students
22-25	500
25-28	56
28-31	134
31 and above	18
Total	708

Qualitative Classification

In qualitative classification, data are classified based on some attributes or qualitative characteristics such as sex, color of hair, literacy, religion, etc. You should note that in this type of classification the attribute under study cannot be measured quantitatively. One can only count it according to its presence or absence among the individuals of the population under study. For example, in case of color blindness, we may find out as how many persons are color blind in a given population. It is not possible to measure the degree of color blindness in each case. Thus, when only one attribute is studied, two classes are formed – one for possessing the attribute and the other for not possessing it. This type of classification is known as simple classification. For example, the population under study may be divided into two categories based on the characteristic 'Color blindness' as follows:



This is an example of simple classification. The Qualitative Classification can also be manifold that is there is sub-classification within a group of population. An example is



Methods of Data Classification

There are two ways in which the data can be classified- Inclusive method and Exclusive Method.

Inclusive Method

When the data are classified in such a way that both lower and upper limits of a class interval are included in the interval itself, then it is said to be the inclusive method of classifying data. This method is shown in table 2.4

Table 2.4 Inclusive Classification of Data

Marks	Number of Students
0-9	36
10-19	78
20-29	132
30-39	67
40-50	45

An important point to be kept in mind while going for an inclusive method of data classification is that discrete variables should be classified in an inclusive method.

Exclusive Method

When the data are classified in such a way that the upper limit of a class interval is the lower limit of the succeeding class interval (i.e. no data point falls into more than one class interval), then it is said to be the exclusive method of classifying data. This method is illustrated in table 2.5

Table 2.5 Exclusive Classification of Data

Marks	Number of Students
0-10	36
10-20	78

20-30	132
30-40	67
40-50	45

2.3 Tabulation of Data

Meaning and Definition Tabulation is another way of summarizing and presenting the given data in a systematic form in rows and columns. Such presentation facilitates comparison by bringing related information close to each other and helps in further statistical analysis and interpretation. Tabulation has been defined by two statisticians as:

The logical listing of related quantitative data in vertical columns and horizontal rows of numbers with sufficient explanatory and qualifying words, phrases and statements in the form of titles, headings and explanatory notes to make clear the full meaning, context and the origin of the data.
– Tuttle

This definition gives an idea of the broad structure of statistical tables and suggests that tabulation helps organize a set of data in an orderly manner to highlight its basic characteristics.

Tables are means of recording in permanent form the analysis that is made through classification and by placing in just position things that are similar and should be compared. – Secrist

This definition defines tabulation as the process of classifying the data in a systematic form which facilitates comparative studies of data sets.

Contents of Table

The various components of a table may vary case to case depending upon the given data. But a good table must contain at least the following components:

1. Table Number
2. Table Heading
3. Caption
4. Stub
5. Body of Table
6. Head Note
7. Foot Note

1. Table Number: A statistical table should be numbered. There are different ways with regard to the place where table number is to be given. The table number may be shown either in the centre at the top above the title or in the left hand side of the table at the top. When there are many columns, it is desirable to number each column so that easy reference to it is possible.

2. Table Heading: A good table should have a suitable heading. The heading is a brief description of the contents of the table. It should be placed above the table. It should answer the following questions: (a) what categories of statistical data are shown? (b) Where the data occurred? (c) When the data occurred? In other words the heading of the table should be clear, brief and self explanatory, but sometimes long title may have to be used for the sake of clarity. The title should be so worded that it permits one and only one interpretation.

3. Caption: Caption refers to the column heading, and explains what information column presents. It may consist of one or more column headings, i.e. under a column heading there may be two or more sub headings. The caption should be clearly defined and placed at the middle of the column. If the different columns are expressed in different units, the unit should be specified along with the captions.

4. Stub: The stubs are row headings. They are placed at the extreme left of the table and perform the same function for the horizontal rows in the table as the captions do for the vertical columns.

5. Body: The body of the table is the central part of table that contains the numerical information presented in table. This is the most vital part of the table.

6. Head Note: Head note is a brief explanatory statement applying to all or a major part of the material presented in the table and is placed below the title entered and enclosed in brackets. It is used to explain certain points relating to the whole table that have not been included in the title nor in the captions or stubs. For example, the unit of measurement is frequently written as the head note such as “in thousands” or “million tons” or “in crores”, etc.

7. Foot Note: Anything in a table which the reader supposed to find difficult to understand should be explained in footnotes. Footnotes may be placed directly below the body of the table. The footnotes are generally used for the following purposes:

(a) Any special circumstances affecting the data, for example, strike, fire, etc.

(b) To clarify anything in the table.

(c) To give the source in case of the secondary data. If any information in the table obtained from some journal, its name, date of publication, page number, table number, etc. should be mentioned so that if the user wishes to check the data from the original source, he could know where to look for the information. After discussing the parts of a table, let us discuss different kinds of tables, through which we can represent or arrange the different types of information's.

Types of Tables

Tables may broadly be classified into following two categories.

1. Simple and Complex Tables
2. General Purpose and Special Purpose Tables

Simple and Complex Tables

The simple and complex tables can be differentiated on the basis of number of characteristics presented and studied. If the data based on one characteristic is presented, the table is known as simple table. The simple table is also known as one way table. On the other hand, in a complex table, two or more characteristics are presented. The complex tables are frequently used in practice because they facilitate to incorporate full information and a proper consideration of all related facts. If the data are tabulated on the basis of only two characteristics, then the table is known as two way tables. If three characteristics are arranged in a table then the table is known as treble table. When four or more characteristics are simultaneously presented it is known as manifold tabulation

The following table presenting the distribution of marks obtained by 100 students in a test is an illustration of a simple table:

Table 2.5 Distribution of Marks Obtained by 100 Students in Statistics

Marks	No. of Students
Below 10	5
10-20	8
20-30	12
30-40	10
40-50	15
50-60	18
60-70	17
70-80	13

Above 80	02
Total	100

Two Way Table

Two way tables show two characteristics and is formed when either the stub or the caption or both are divided into two categories. In the following example the nature of such a table is given and is an illustration of two-way table (a complex table):

Age	Persons Living in the colony		Total
	MALE	FEMALE	
Below 15	12	6	18
15-25	20	14	32
25-35	42	27	69
35-45	25	18	43
45-55	10	8	18
55-65	8	5	13
65- and Above	5	2	07
Total	122	78	200

Higher Order Table

When three or more characteristics are represented in the table then such a table is called higher order table. The need for such a table arises when we are interested in presenting three or more characteristics simultaneously.

It should be remembered that as the number of characteristics increases, the table becomes more and more condensing. It is advised normally not more than four characteristics should be represented in the same table. When more than four characteristics are to be represented, we should form more than one table depicting relationship between different attributes.

General Purpose and Special Purpose Tables

General purpose tables, also known as reference tables or repository tables, and provide the information for general use or reference. They usually contain detailed information and are not used for specific discussion. In other words, these tables serve as a repository of information and are arranged for easy reference such as the tables published by government agencies, the tables contained in the statistical abstract of the Indian Union, tables in the census reports, etc.

The general tables tell facts which are not for particular discussion. If general tables are used by a researcher, they are usually placed in the form of appendix at the end of the report for easy reference.

Special purpose tables, also known as summary tables or analytical tables, provide information for particular discussion. These tables are also called derivative tables since they are often derived from general tables. A special purpose table should be designed in such a way that a reader may easily refer to the table for comparison, analysis or emphasis concerning the specific discussion.

2.4 Constructing a Frequency Distribution

Techniques used to describe a set of data as descriptive statistics. To put it another way, we use descriptive statistics to organize data in various ways to point out where the data values tend to concentrate and help distinguish the largest and the smallest values. The first procedure we use to describe a set of data is a frequency distribution.

A grouping of data into mutually exclusive classes showing the number of observations in each.

How do we develop a frequency distribution? The first step is to tally the data into a table that shows the classes and the number of observations in each class. The steps in constructing a frequency distribution are best described by using an example. Remember, our goal is to construct tables, charts, and graphs that will quickly reveal the shape of the data.

“Frequency distribution shows the number of cases falling within a given class interval or range of scores.” -Chaplin (1975)

As the number of observations obtained gets large, the method discussed above to condense the data becomes quite difficult and time-consuming. Thus, to further condense the data into frequency distribution tables, the following steps should be taken:

- i. Select an appropriate number of non-overlapping class intervals
 - ii. Determine the width of the class intervals (iii) Determine class limits (or boundaries) for each class interval to avoid overlapping.
1. **Decide the number of class intervals:** The decision on the number of class groupings depends largely on the judgment of the individual investigator and/or the range that will be used to group the data, although there are certain guidelines that can be used. As a general rule, a frequency distribution should have at least five class intervals (groups), but not more than fifteen. The following two rules are often used to decide approximate number of classes in a frequency distribution:
 - i) If k represents the number of classes and N the total number of observations, then the value of k will be the smallest exponent of the number 2, so that $2^k \geq N$

If $N=30$

$$2^3 = 8 < 30$$

$$2^4 = 16 < 30$$

$$2^5 = 32 > 30$$

Thus, we may choose $k = 5$ as the number of classes.

a) According to Sturge's rule, the number of classes can be determined by the formula k

$$= 1 + 3.222 \log_e N$$

Where k is the number of classes and $\log_e N$ is the logarithm of the total number of observations.

Applying this rule to the data, we get

$$k = 1 + 3.222 \log 30$$

$$= 1 + 3.222 (1.4771) = 5.759 \cong 5$$

2. **Determine the width of class intervals:** When constructing the frequency distribution, it is desirable that the width of each class interval should be equal in size. The size (or width) of each class interval can be determined by first taking the difference between the largest and smallest numerical values in the data set and then dividing it by the number of class intervals desired.

Width of class interval (h) = (Largest numerical value - Smallest numerical value) / Number of classes desired

The value obtained from this formula can be rounded off to a more convenient value based on the investigator's preference.

3. **Determine class limits (Boundaries):** The limits of each class interval should be clearly defined so that each observation (element) of the data set belongs to one and only one class. Each class has two limits—a lower limit and an upper limit. The usual practice is to let the lower limit of the first class be a convenient number slightly below or equal to the lowest value in the data set.

Methods of Data Classification

There are two ways in which observations in the data set are classified on the basis of class intervals, namely

- i. Exclusive method, and
- ii. Inclusive method

Exclusive Method: When the data are classified in such a way that the upper limit of a class interval is the lower limit of the succeeding class interval (i.e. no data point falls into more than one class interval), then it is said to be the exclusive method of classifying data.

Table 1: Exclusive Method of Data Classification

<i>Dividend Declared in per cent (Class Intervals)</i>	<i>Number of Companies (Frequencies)</i>
0–10	5
10–20	7
20–30	15
30–40	10

Such classification ensures continuity of data because the upper limit of one class is the lower limit of succeeding class. As shown in Table 1, 5 companies declared dividend ranging from 0 to 10 per cent, this means a company which declared exactly 10 per cent dividend would not be included in the class 0–10 but would be included in the next class 10–20.

Inclusive Method:

When the data are classified in such a way that both lower and upper limits of a class interval are included in the interval itself, then it is said to be the inclusive method of classifying data.

Table 2: Inclusive Method of Data Classification

<i>Number of Accidents (Class Intervals)</i>	<i>Number of Weeks (Frequencies)</i>
0– 4	5
5– 9	22
10–14	13
15–19	8
20– 24	2



Example 1: The following set of numbers represents mutual fund prices reported at the end of a week for selected 40 nationally sold funds

10 17 15 22 11 16 19 24 29 18 25 26 32 14 17 20 23 27 30 12 15 18 24 36 18 15 21 28 33 38 34 13 10 16 20 22 29 29 23 31

Research Methods and Design

Arrange these prices into a frequency distribution having a suitable number of classes.

Solution: Since the number of observations are 40, it seems reasonable to choose 6 ($26 > 40$) class intervals to summarize values in the data set. Again, since the smallest value is 10 and the largest is 38, therefore the class interval is given by

$h = \text{Range} / \text{Number of classes} = (38-10)/6 = 28/6 = 4.66 \approx 5$ Now performing the actual tally and counting the number of values in each class, we get the frequency distribution by exclusive method as shown in Table 3:

Table 3: Frequency Distribution

<i>Class Interval</i> (<i>Mutual Fund Prices, Rs</i>)	<i>Tally</i>	<i>Frequency</i> (<i>Number of Mutual Funds</i>)
10 – 15		6
15 – 20		11
20 – 25		9
25 – 30		7
30 – 35		5
35 – 40		2
		40



Example 2: A computer company received a rush order for as many home computers as could be shipped during a six-week period. Company records provide the following daily shipments:

22 65 65 67 55 50 65 77 73 30 62 54 48 65 79 60 63 45 51 68 79 83 33 41 49 28 55 61 65 75 55 75 39 87 45 50 66 65 59 25 35 53

Group these daily shipments figures into a frequency distribution having the suitable number of classes.

Solution: Since the number of observations are 42, it seems reasonable to choose 6 ($26 > 42$) classes. Again, since the smallest value is 22 and the largest is 87, therefore the class interval is given by $h = \text{Range} / \text{Number of classes} = (87-22)/6 = 65/6 = 10.833$ or 11

Now performing the actual tally and counting the number of values in each class, we get the following frequency distribution by inclusive method as shown in Table 4:

Table 4: Frequency Distribution

<i>Class Interval</i> (<i>Number of Computers</i>)	<i>Tally</i>	<i>Frequency</i> (<i>Number of Days</i>)
22 – 32		4
33 – 43		4
44 – 54		9
55 – 65		14
66 – 76		6
77 – 87		5
		42



Example 5: Following are the number of items of similar type produced in a factory during the last 50 days

21 22 17 23 27 15 16 22 15 23 24 25 36 19 14 21 24 25 14 18 20 31 22 19 18 20 21 20 36 18 21 20 31 22 19 18 20 20 24 35 25 26 19 32 22 26 25 26 27 22

Arrange these observations into a frequency distribution with both inclusive and exclusive class intervals choosing a suitable number of classes

Unit 02: Classification and Tabulation

Solution: Since the number of observations are 50, it seems reasonable to choose 6 ($26 > 50$) or less classes. Since smallest value is 14, and the largest is 36 therefore the class interval is given byh = Range/Number of classes = $(36-14) / 6 = 22 / 6 = 3.66$ or 4

Performing the actual tally and counting the number of observations in each class, we get the following frequency distribution with inclusive class intervals as shown in Table 5.

Table 5 Frequency Distribution with Inclusive Class Intervals

<i>Class Intervals</i>	<i>Tally</i>	<i>Frequency (Number of Items Produced)</i>
14 – 17		6
18 – 21		18
22 – 25		15
26 – 29		5
30 – 33		3
34 – 33		3
		50

Converting the class intervals shown in Table 5 into exclusive class intervals is shown in Table 6.

Table 6: Frequency Distribution with Exclusive Class Intervals

<i>Class Intervals</i>	<i>Mid-Value of Class Intervals</i>	<i>Frequency (Number of Items Produced)</i>
13.5 – 17.5	15.5	6
17.5 – 21.5	19.5	18
21.5 – 25.5	23.5	15
25.5 – 29.5	27.5	5
29.5 – 33.5	31.5	3
33.5 – 37.5	34.5	3

2.5 Cumulative Frequency Distribution

Sometimes it is preferable to present data in a cumulative frequency (cf) distribution or simply a distribution which shows the cumulative number of observations below the upper boundary (limit) of each class in the given frequency distribution. A cumulative frequency distribution is of two types: (i) more than type and (ii) less than type.

In a less than cumulative frequency distribution, the frequencies of each class interval are added successively from top to bottom and represent the cumulative number of observations less than or equal to the class frequency to which it relates. But in the more than cumulative frequency distribution, the frequencies of each class interval are added successively from bottom to top and represent the cumulative number of observations greater than or equal to the class frequency to which it relates.

The frequency distribution given in Table 7 illustrates the concept of cumulative frequency distribution:

Table 7: Cumulative Frequency Distribution

Research Methods and Design

<i>Number of Accidents</i>	<i>Number of Weeks (Frequency)</i>	<i>Cumulative Frequency (less than)</i>	<i>Cumulative Frequency (more than)</i>
0– 4	5	5	$45 + 5 = 50$
5– 9	22	$5 + 22 = 27$	$23 + 22 = 45$
10–14	13	$27 + 13 = 40$	$10 + 13 = 23$
15–19	8	$40 + 8 = 48$	$2 + 8 = 10$
20–24	2	$48 + 2 = 50$	2

From Table 7 it may be noted that cumulative frequencies are corresponding to the lower limit and upper limit of class intervals. The 'less than' cumulative frequencies are corresponding to the upper limit of class intervals and 'more than' cumulative frequencies are corresponding to the lower limit of class intervals shown in Table 8 and 9.

Table 8: Less than type Cumulative Frequency Distribution

<i>Upper Limits</i>	<i>Cumulative Frequency (less than)</i>
less than 4	5
less than 9	27
less than 14	40
less than 19	48
less than 24	50

Table 9: More than Type Cumulative Frequency Distributions

<i>Lower Limits</i>	<i>Cumulative Frequency (more than)</i>
0 and more	50
5 and more	45
10 and more	23
15 and more	10
20 and more	2

Summary

Collected data are unorganized and complex mass of figures. To draw some meaningful conclusions, they must be arranged in an orderly manner. This can be done in many ways, such as by forming simple and frequency array, discrete and continuous frequency distributions, etc. Sometimes, it serves a useful purpose to form what is called "less-than" or "morethan" cumulative frequency distributions. The former is arrived at by successive totaling of frequencies from above and the latter by successive totaling from below. After collection and condensation of data, good presentation of data is important. A good presentation helps to highlight important points of the data and makes possible useful comparisons and their intelligent use. This can be done through five statistical tools. These are: i) formal tables - one-way and two-way; ii) line graph histograms, frequency polygon and frequency curves; iii) cumulative distributions- "less-than" and "more-than" ogives; iv) one, two and three-dimensional diagrams such as bar diagrams, rectangles, squares, circles, cubes and pie diagrams; and v) statistical maps.

Keywords

1. Discrete frequency distribution: A discrete distribution or discrete series is formed where the variable can take only discrete values like 1,2,3, Number of children in a family, number of students in a university, etc. is examples of discrete variable.
2. Continuous frequency distribution: A continuous frequency distribution is formed where the variable can take any value between two numbers. For example, height, weight, income and temperature.
3. Inclusive type class interval: A class interval in which all observations lying between and including the class limits are included.
4. Exclusive type class interval: A class interval which includes all observations that are greater than or equal to the lower limit but less than the upper limit.
5. Cumulative frequency distribution: It is obtained by successive totaling of the simple frequencies of a discrete or continuous frequency distribution. This totaling can be done either from above (we get "less-than" cumulative frequency distribution) or from below (we get "more-than" cumulative frequency distribution).

Self Assessment

1. Which of the following is an example of quantitative data?
 - A. Honesty
 - B. Colors
 - C. Achievement
 - D. Ethnicity

2. Based on the measurement, there are _____ types of data.
 - A. three
 - B. four
 - C. five
 - D. Six

3. Measurement of heights is an example of
 - A. Interval data
 - B. Nominal data
 - C. Ordinal data
 - D. Ratio data

4. Human blood groups are an example of
 - A. Interval data
 - B. Continuous data
 - C. Discontinuous data
 - D. Ratio data

5. The qualitative data can be graphically represented by using
 - A. Pie chart and bar graph

- B. Pie chart only
 - C. Bar graph only
 - D. Histogram and bar graph
6. The classification means to divide the data into _____
- A. Homogenous groups.
 - B. Heterogeneous groups.
 - C. Both homogenous and heterogeneous groups.
 - D. Classes.
7. The classification of girls and boys (studying in B.Sc. – Ist year, IInd year and IIIrd year) of science department can be classified according to their
- A. Gender
 - B. Class and gender
 - C. Class
 - D. Department
8. Generally, data can be classified into _____ types.
- A. Two
 - B. Three
 - C. Four
 - D. Five
9. Qualitative data are classified according to the characteristics or attributes
- A. Which an individual can measure.
 - B. Which an individual can measure and only find out the presence or absence.
 - C. Whose presence or absence we can only find out.
 - D. Which an individual cannot measure and only find out the presence or absence.
10. Which of the following is not the primary rule of classification?
- A. The number of classes should not be excessive
 - B. Avoid ambiguity in the definition of classes.
 - C. All the classes should not have equal width.
 - D. The magnitude of the class intervals should be as far as possible in multiples of 5.
11. Tally Mark
- A. Is only represented by upward straight (|).
 - B. Is only represented by slanted stroke (/).
 - C. Is either represented by slanted stroke (/) or upward straight (|).
 - D. is neither represented by slanted stroke (/) nor upward straight (|)
12. Which of the following is correct?
- A. Average of Lower and Upper Class Limit values

- B. (Lower Class Limit - Upper Class Limit)
 C. (Upper Class Limit - Lower Class Limit)
 D. $2(\text{Lower Class Limit} + \text{Upper Class Limit})$
13. Which of the following is an example of true class type?
 A. 20-24, 25-29
 B. Above 20 but less than 24, above 24 but less than 29
 C. Above 0 - 4, 5-9
 D. 19.5-24.5, 24.5-29.5
14. The range of the scores - 72, 61, 63, 65, 62, 68, 69, 64, 65, and 67 is
 A. 8
 B. 9
 C. 10
 D. 11
15. Which of the following is correct?
 A. Width of the class-interval = $\text{Range}/(1 + 3.322\log_{10}N)$
 B. Width of the class-interval = $\text{Range}/(1 - 3.322\log_{10}N)$
 C. Width of the class-interval = $(1+\text{Range})/(3.322\log_{10}N)$
 D. Width of the class-interval = $(\text{no. of class-intervals})/\text{Range}$

Answers for SelfAssessment

1. C 2. B 3. D 4. C 5. A
 6. A 7. B 8. C 9. D 10. C
 11. C 12. A 13. D 14. D 15. A

Review Questions

- What are the advantages of using a frequency distribution to describe a body of raw data?
What are the disadvantages?
- A portfolio contains 51 stocks whose prices are given below:
67 34 36 48 49 31 61 34 43 45 38 32 27 61 29 47 36 50 46 30 40 32 30 33 45 49 48 41 53 36 37 47
47 30 50 28 35 35 38 36 46 43 34 62 69 50 28 44 43 60 39
Summarize these stock prices in the form of a frequency distribution.
- Form a frequency distribution of the following data. Use an equal class interval of 4 where the lower limit of the first class is 10.
10 17 15 22 11 16 19 24 29 18 25 26 32 14 17 20 23 27 30 12 15 18 24 36 18 15 21 28 33 38 34 13
10 16 20 22 29 29 23 31
- The distribution of ages of 500 readers of a nationally distributed magazine is given below:

<i>Age (in Years)</i>	<i>Number of Readers</i>
Below 14	20
15–19	125
20–24	25
25–29	35
30–34	80
35–39	140
40–44	30
45 and above	45

Find the cumulative frequency distributions for this distribution.

5. A. Illustrate two methods of classifying data in class-intervals.
B. Why is it necessary to summarize data? Explain the approaches available to summarize data distributions.
6. Distinguish clearly between a continuous variable and a discrete variable. Give two examples of continuous variables and two examples of discrete variables that might be used by a statistician.
7. What are the objections to unequal class and open class intervals? State the conditions under which the use of unequal class intervals and open class intervals are desirable and necessary.
8. Classify the following data by taking class intervals such that their mid-values are 17, 22, 27, 32, and so on:
30 42 30 54 40 48 15 17 51 42 25 41 30 27 42 36 28 26 37 54 44 31 36 40 36 22 30 31 19 48 16 42 32 21 22 46 33 41 21



Further Readings

- Elhance, D. N. and V. Elhance, 1988, Fundamentals of Statistics, Kitab Mahal, Allahabad.
- Nagar, A. L. and R. K. Dass, 1983, Basic Statistics, Oxford University Press, Delhi
- Mansfield, E., 199 1, Statistics for Business and Economics: Methods and Applications, W. W. Norton and Co.

Unit 03: Data Graphical Presentation and Analysis

CONTENTS

Objectives

Introduction

3.1 Importance Of Graphic Presentation

3.2 General Rules for Drawing Diagrams

3.3 Types of Diagrams

3.4 Advantages and Limitations of Diagrams (Graphs)

Summary

Keywords

Self Assessment

Answers for Self Assessment

Review Questions

Further Readings

Objectives

- Explain the need and significance of visual presentation (diagrams and graphs) of the data in research work,
- Describe various types of diagrams and illustrate how to present the data through an appropriate diagram,
- Describe the principle of preparing a graph,
- Present frequency distribution in the form of histograms, histograms, frequency polygon and ogives to make decisions.

Introduction

It has already been discussed that one of the important functions of statistics is to present complex and unorganized (raw) data in such a manner that they would easily be understandable. All the available numerical data can be represented graphically. A graph is the representation of data by using graphical symbols such as lines, bars, pie diagrams, dots etc. A graph represents a numerical data in the form of a structure and provides important information to the user of the data. When an organized data is graphically represented it not only looks attractive but it is easier to understand. A large amount of data can be presented in a very concise and attractive manner. Graphs are effective and economical as well. They are also easy to interpret and adequately reflect any comparison between two sets of data.

It has already been discussed that one of the important functions of statistics is to present complex and unorganized (raw) data in such a manner that they would easily be understandable. According to King, 'One of the chief aims of statistical science is to render the meaning of masses of figures clear and comprehensible at a glance.' This is often best accomplished by presenting the data in a pictorial (or graphical) form.

The graphical (diagrammatical) presentation of data has many advantages. The following persons rightly observed that

With but few exceptions memory depends upon the faculty of our brains possess of forming visual images and it is this power of forming visual images which lies at the root of the utility of diagrammatic presentation.

— R. L. A. Holmes

Cold figures are uninspiring to most people. Diagrams help us to see the pattern and shape of any complex situation. Just as a map gives us a bird's eye-view of the wide stretch of a country, so diagrams help as visualize the whole meaning of the numerical complex at a single glance.

—M. J. Moroney

Figures are not always interesting, and as their size and number increases, they become uninteresting and confusing to such an extent that nobody would like to study them. The work of a statistician is to understand the data himself, and to put them in such a way that their importance may be known to everyone. According to Calvin F. Schmid, 'Charts and graphs represent an extremely useful and flexible medium for explaining, interpreting and analyzing numerical facts largely by means of points, lines, areas and other geometric forms and symbols. They make possible the presentation of quantitative data in a simple, clear, and effective manner and facilitate comparison of values, trends and relationships.'

3.1 Importance Of Graphic Presentation

Visual presentation of statistical data has become more popular and is often used by the researcher and the statistician in analysis. Visual presentation of data means presentation of Statistical data in the form of diagrams and graphs. In these days, as we know, every research work is supported with visual presentation because of the following reasons.

- 1) ***They relieve the dullness of the numerical data:*** Any list of figures becomes less comprehensible and difficult to draw conclusions from as its length increases. Scanning of the figures from tables causes undue strain on the mind. The data when presented in the form of diagrams and graphs gives a bird's eye-view of the entire data and creates interest and leaves an impression on the mind of readers for a long period.
- 2) ***They make comparison easy:*** This is one of the prime objectives of visual presentation of data. Diagrams and graphs make quick comparison between two or more sets of data simpler, and the direction of curves bring out hidden facts and associations of the statistical data.
- 3) ***They save time and effort:*** The characteristics of statistical data, through tables, can be grasped only after a great strain on the mind. Diagrams and graphs reduce the strain and save a lot of time in understanding the basic characteristics of the data.
- 4) ***They facilitate the location of various statistical measures and establish trends:*** Graph makes it possible to locate several measures of central tendency such as Median, Quartiles, Mode etc. They help in establishing trends of the past performance and are useful in interpolation or extrapolation, line of best fit, establishing correlation etc. Thus, it helps in forecasting.
- 5) ***They have universal applicability:*** It is a universal practice to present the numerical data in the form of diagrams and graphs. In these days, it is an extensively used technique in the field of economics, business, education, health, agriculture etc.
- 6) ***They have become an integral part of research:*** In fact, now days it is difficult to find any research work without visual support. The reason is that this is the most convincing and appealing way of presenting the data. You can find diagrammatic and graphic presentation of data in journals, magazines, television, reports, advertisements etc. After having understood about the importance of visual presentation, we shall move on to discuss about the Diagrams and graphs which are more frequently used in the area of business research.

3.2 General Rules for Drawing Diagrams

As you know, diagrammatic presentation is one of the techniques of visual presentation of statistical data. It is a fact that diagrams do not add new meaning to the statistical facts, but they reveal the facts of the data more quickly and clearly. Because examining the figures from tables becomes laborious and uninteresting to the eye and also confusing. Here, it is appropriate to state the words of M. J. Moroney, "cold figures are uninspiring to most people. Diagrams help us to see the pattern and shape of any complex situation." Thus, the data presented through diagrams are the best way of appealing to the mind visually. Hence, diagrams are widely used in practice to display the structure of the data in research work.

Rules for Preparing Diagrams

To draw useful inferences from graphical presentation of data, it is important to understand how they are prepared and how they should be interpreted. When we say that 'one picture is worth a thousand words', it neither proves (nor disproves) a particular fact, nor is it suitable for further analysis of data. However, if diagrams are properly drawn, they highlight the different characteristics of data. The following general guidelines are taken into consideration while preparing diagrams:

1. **Title:** Each diagram should have a suitable title. It may be given either at the top of the diagram or below it. The title must convey the main theme which the diagram intends to portray.
2. **Size:** The size and portion of each component of a diagram should be such that all the relevant characteristics of the data are properly displayed and can be easily understood.
3. **Proportion of length and breadth:** An appropriate proportion between the length and breadth of the diagram should be maintained. As such there are no fixed rules about the ratio of length to width. However, a ratio of 2:1 or 1.414 (long side): 1 (short side) suggested by Lutz in his book *Graphic Presentation* may be adopted as a general rule.
4. **Proper scale:** There are again no fixed rules for selection of scale. The diagram should neither be too small nor too large. The scale for the diagram should be decided after taking into consideration the magnitude of data and the size of the paper on which it is to be drawn. The scale showing the values as far as possible, should be in even numbers or in multiples of 5, 10, 20, and so on. The scale should specify the size of the unit and the nature of data it represents, for example, 'millions of tonnes', in Rs thousand, and the like. The scale adopted should be indicated on both vertical and horizontal axes if different scales are used. Otherwise, can be indicated at some suitable place on the graph paper.
5. **Footnotes and source note:** To clarify or elucidate any points which need further explanation but cannot be shown in the graph, footnotes are given at the bottom of the diagrams.
6. **Index:** A brief index explaining the different types of lines, shades, designs, or colors used in the construction of the diagram should be given to understand its contents.
7. **Simplicity:** Diagrams should be prepared in such a way that they can be understood easily. To keep it simple, too much information should not be loaded in a single diagram as it may create confusion. Thus, if the data are large, then it is advisable to prepare more than one diagram, each depicting some identified characteristic of the same data.

3.3 Types of Diagrams

There are a variety of diagrams used to represent statistical data. Different types of diagrams, used to describe sets of data, are divided into the following categories:

1. Dimensional diagrams

- (i) One dimensional diagram such as histograms, frequency polygons, and pie chart.
- (ii) Two-dimensional diagrams such as rectangles, squares, or circles.
- (iii) Three dimensional diagrams such as cylinders and cubes.

2. Pictograms or Ideographs

3. Cartographs or Statistical maps

1. Dimensional diagrams

a. One-Dimensional Diagrams:

These diagrams are most useful, simple, and popular in the diagrammatic presentation of frequency distributions. These diagrams provide a useful and quick understanding of the shape of the distribution and its characteristics. According to Calvin F. Schmid, 'The simple bar chart with many variations is particularly appropriate for comparing the magnitude (or size) of coordinate items or of parts of a total. The basis of comparison in the bar is linear or one-dimensional.'

These diagrams are called one-dimensional diagrams because only the length (height) of the bar (not the width) is taken into consideration. Of course, width or thickness of the bar has no effect on the diagram, even then the thickness should not be too much otherwise the diagram would appear like a two-dimensional diagram.

Tips for Constructing a Diagram

The following tips must be kept in mind while constructing one-dimensional diagrams:

- i. The width of all the bars drawn should be same.
- ii. The gap between one bar and another must be uniform.
- iii. There should be a common base to all the bars.
- iv. It is desirable to write the value of the variable represented by the bar at the top end so that the reader can understand the value without looking at the scale.
- v. The frequency, relative frequency, or per cent frequency of each class interval is shown by drawing a rectangle whose base is the class interval on the horizontal axis and whose height is the corresponding frequency, relative frequency, or per cent frequency.
- vi. The value of variables (or class boundaries in case of grouped data) under study are scaled along the horizontal axis, and the number of observations (frequencies, relative frequencies or percentage frequencies) are scaled along the vertical axis.

The one-dimensional diagrams (charts) used for graphical presentation of data sets are as follows:

1. Histogram
2. Frequency polygon
3. Frequency curve
4. Cumulative frequency distribution (Ogive)
5. Pie diagram

1. Histograms (Bar Diagrams)

These diagrams are used to graph both ungrouped and grouped data. In the case of an ungrouped data, values of the variable (the characteristic to be measured) are scaled along the horizontal axis and the number of observations (or frequencies) along the vertical axis of the graph. The plotted points are then connected by straight lines to enhance the shape of the distribution. The height of such boxes (rectangles) measures the number of observations in each of the classes.

- i. **Simple bar charts:** In a Simple bar diagram, the data related to one variable is depicted. Such as, profits, investments, exports, sales, production etc. This type of diagram may be drawn either vertically or horizontally. Both positive and negative values can be presented. In such a case, if bars are constructed vertically, the positive values are taken on the upper side of horizontal axis while the negative values are taken on its lower side. On the other hand, if the bars are constructed horizontally, the positive values are taken on the right-hand side of the vertical axis and the negative values are considered on its left side. This type of construction of bars are also called deviation bar diagram. The simple bar diagram is very easy to prepare and to understand the level of fluctuations from one situation to another. It

should be kept in mind that, only length is taken into account and not width. Width should be uniform for all bars and the gap between each bar is normally identical. Let us consider the following illustrations and learn how to present the given data in the form of simple bar diagrams vertically and horizontally.

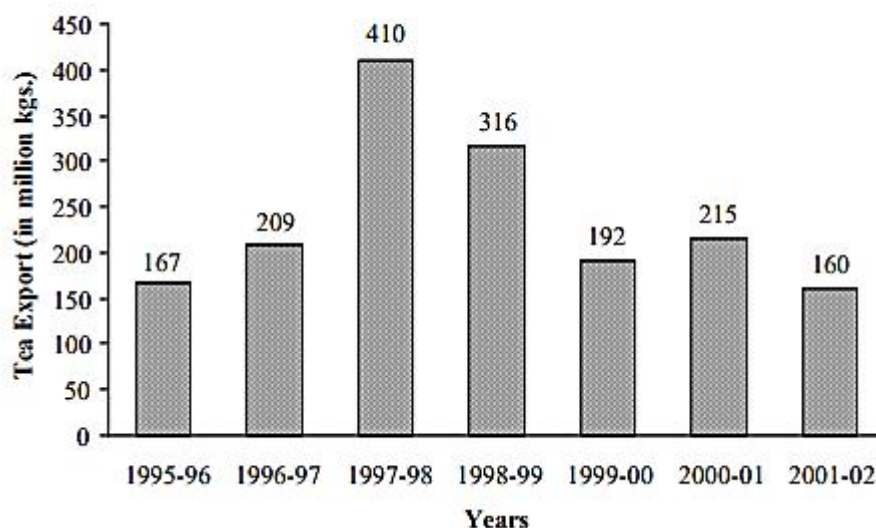


Example 1: Prepare a Simple Bar Diagram from the Following Data Relating to Tea Exports.

Year	1995-96	1996-97	1997-98	1998-99	1999-00	2000-01	2001-02
Exports	167	209	410	316	192	215	160

Solution: The quantity of tea exported is given in million kgs. for different years. A simple bar diagram will be constructed with 7 bars corresponding to the 7 years. Now study the following vertical construction of bar diagram by referring the guide lines for construction of simple bars, as shown in Fig. 3.1

Fig. 3.1: Simple Bar Diagram Showing the Tea Exports in Different Years.

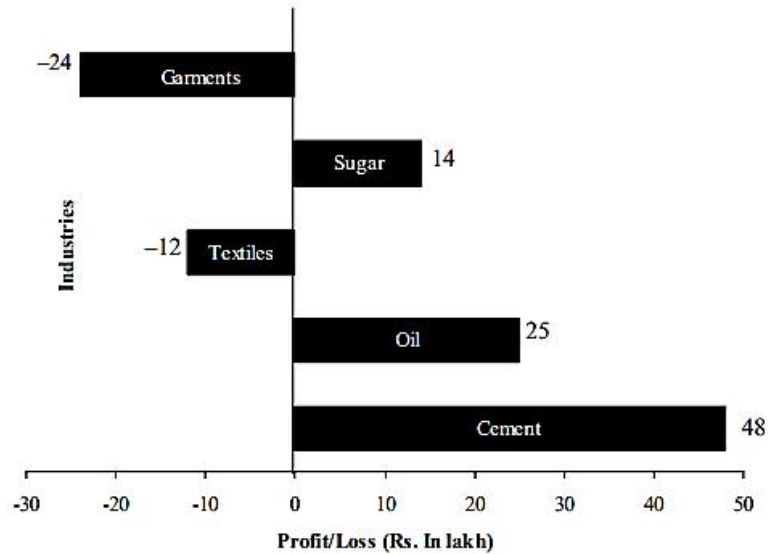


Example 2: The following data relates to the Profit and Loss of different industries in 1999-2002. Present the data through simple bar diagram.

Industry	Cement	Oil	Textile	Sugar	Garments
Profit/Loss	48	25	-12	14	-24

Solution 2: The given data represents positive and negative values i.e., profit and loss. Let us draw the bars horizontally.

Fig. 3.2: Simple Bar Diagram Showing the Profit and Loss of Different Industries during 1999-02



- ii. **Multiple Bar Diagram:** A multiple bar chart is also known as grouped (or compound) bar chart. Such charts are useful for direct comparison between two or more sets of data. The technique of drawing such a chart is same as that of a single bar chart with a difference that each set of data is represented in different shades or colors on the same scale. An index explaining shades or colors must be given.

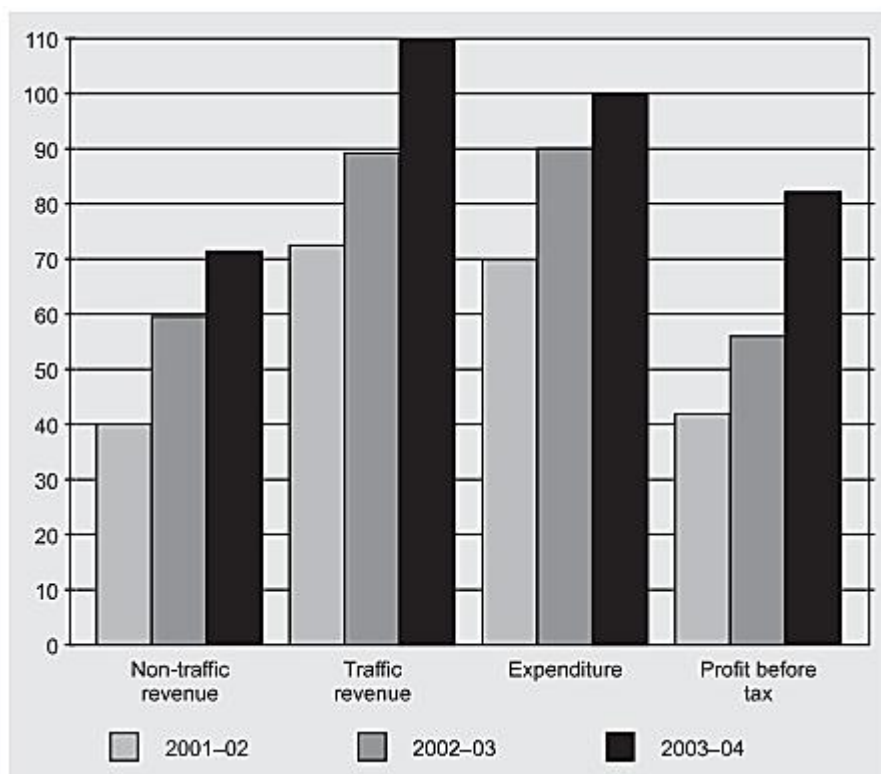


Example 3.3: The data on fund flow (Rs in crore) of an International Airport Authority during financial years 2001-02 to 2003-04 are given below:

	2001-02	2002-03	2003-04
Non-Tariff revenue	40.00	50.75	70.25
Traffic Revenue	70.25	80.75	110.00
Profit before tax	40.15	50.50	80.25

Represent this data by a suitable bar chart.

Solution: The multiple bar chart of the given data is shown in Fig. 3.3



iii. Sub-divided Bar Diagram:

In this diagram one bar is constructed for the total value of the different components of the same variable. Further it is sub-divided in proportion to the values of various components of that variable. This diagram shows the total of the variables as well as the total of its various components in a single bar. Hence, it is clear that the sub-divided bar serves the same purpose as multiple bars. The only difference is that, in case of the multiple bar each component of a variable is shown side by side horizontally, whereas in construction of subdivided bar diagram each component of a variable is shown one upon the other. It is also called a component bar diagram. This method is suitable if the total values of the variables are small, otherwise the scale becomes very narrow to depict the data. To study the relative changes, all components may be converted into percentages and drawn as sub-divided bars. Such a bar construction is called a sub-divided percentage bar. The limitation is that all the parts do not have a common base to enable us to compare accurately the various components of a set.

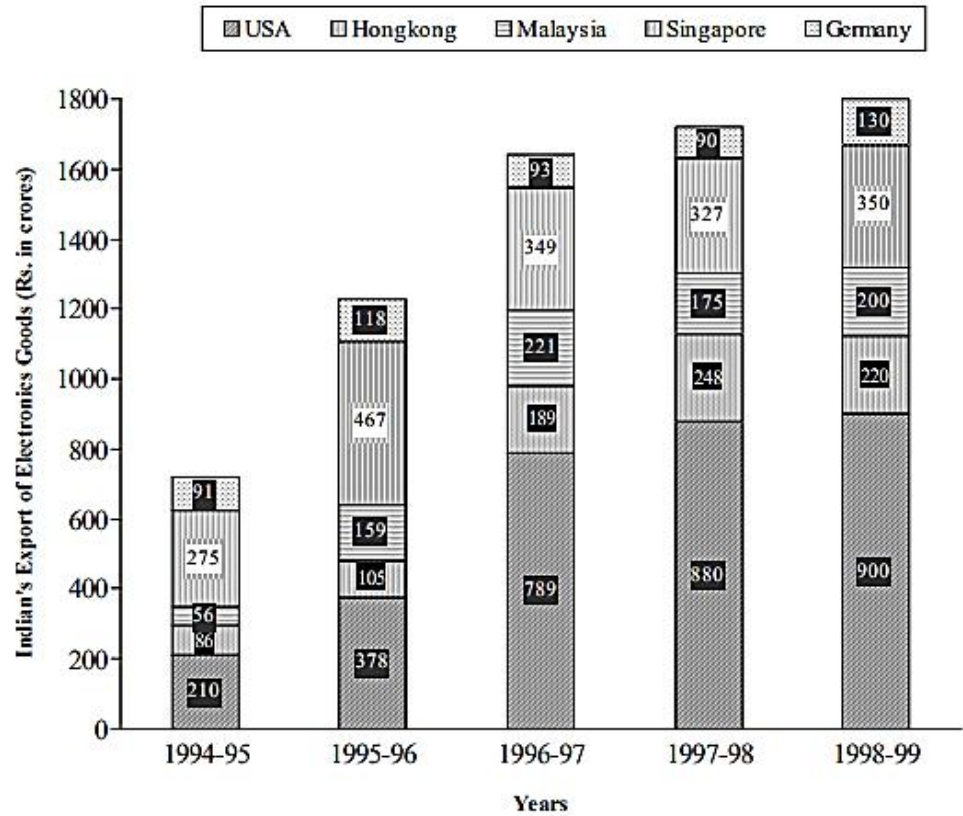


Example 4: The following data relates to India's exports of electronic goods to different countries during 1994-98. Represent the data by sub-divided bar diagram.

Years	Country					Total
	USA	Hong-Kong	Malaysia	Singapore	Germany	
1994-95	210	86	56	275	91	718
1995-96	378	105	159	467	118	1127
1996-97	789	189	221	349	93	1641
1997-98	880	248	175	327	90	1720
1998-99	900	220	200	350	130	1800

Solution: For construction of sub-divided bar diagram, first of all, we must obtain the total export value of the five countries in each year. However, in the above illustration of different countries, total exports in each year are given. Construct sub-divided bar diagram.

Fig. 3.4: Sub-divided Bar Diagram Showing the India's Exports of Electronic Goods to Different Countries During 1994-99.



- iv. **Deviation Bar Charts:** Deviation bar charts are suitable for presentation of net quantities in excess or deficit such as profit, loss, import, or exports. The excess (or positive) values and deficit (or negative) values are shown above and below the base line.



Example 5: The following are the figures of sales and net profits of a company over the last three years.

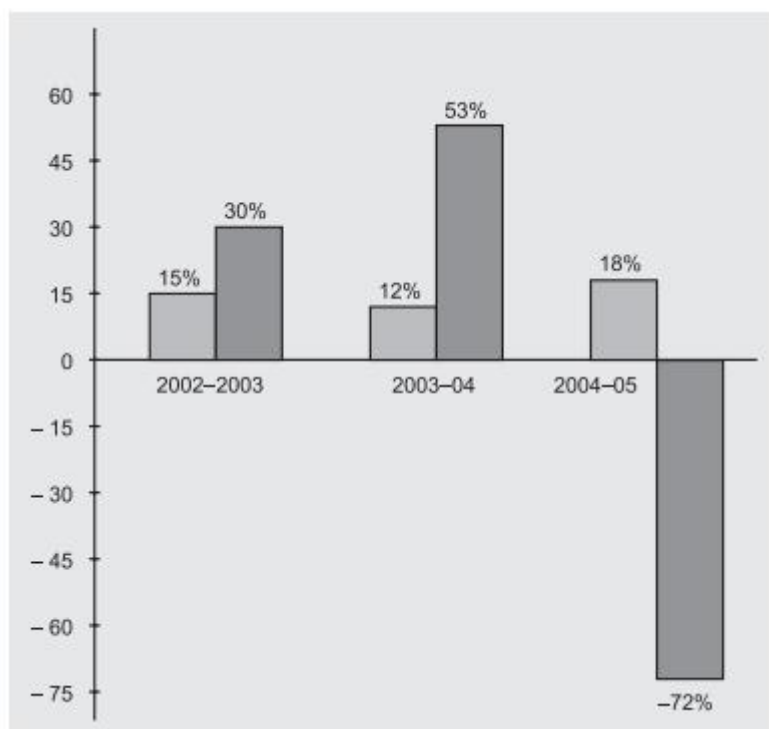
(Per cent change over previous year)

Year	Sales Growth	Net profit
2002-03	15	30
2003-04	12	53
2004-05	18	-72

Present this data by a suitable bar chart.

Solution:

Figure 3.5 depicts deviation bar charts for sales and per cent change in sales over previous year's data.



v. Percentage Bar Charts

When the relative proportions of components of a bar are more important than their absolute values, then each bar can be constructed with same size to represent 100%. The component values are then expressed in terms of percentage of the total to obtain the necessary length for each of these in the full length of the bars. The other rules regarding the shades, index, and thickness are the same as mentioned earlier.

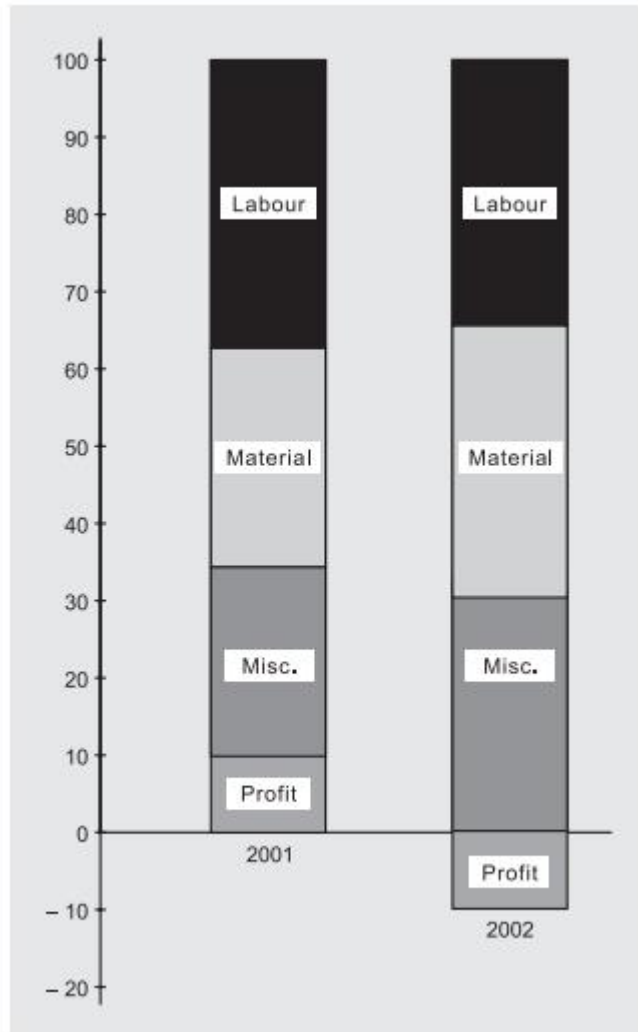


Example 6: The following table shows the data on cost, profit, or loss per unit of a good produced by a company during the year 2003-04.

Particulars	2003			2004		
	Amount (Rs)	Percentage	Cumulative Percentage	Amount (Rs)	Percentage	Cumulative Percentage
Cost per unit						
(a) Labour	25	41.67	41.67	34	40.00	40.00
(b) Material	20	33.33	75.00	30	35.30	75.30
(c) Miscellaneous	15	25.00	100.00	21	24.70	100.00
Total cost	60	100		85	100	
Sales proceeds per unit	80	110		80	88	
Profit (+) or loss (-) per item	+ 20	+ 10		- 5	- 12	

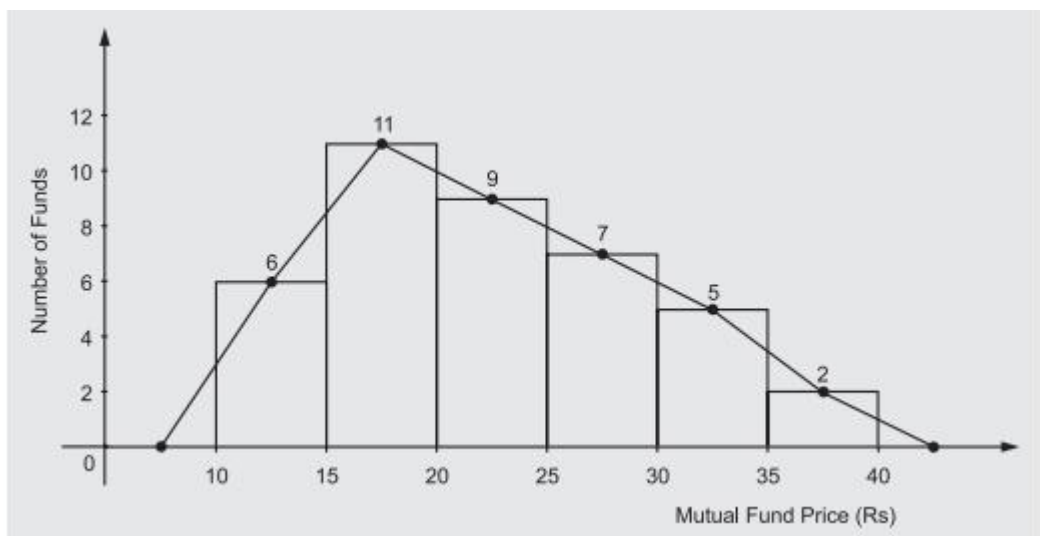
Solution 6: The cost, sales, and profit/loss data expressed in terms of percentages have been represented in the bar chart as shown in Fig. 3.6

Fig. 3.6: Percentage bar Chart Pertaining to Cost, Sales, and Profit/Loss



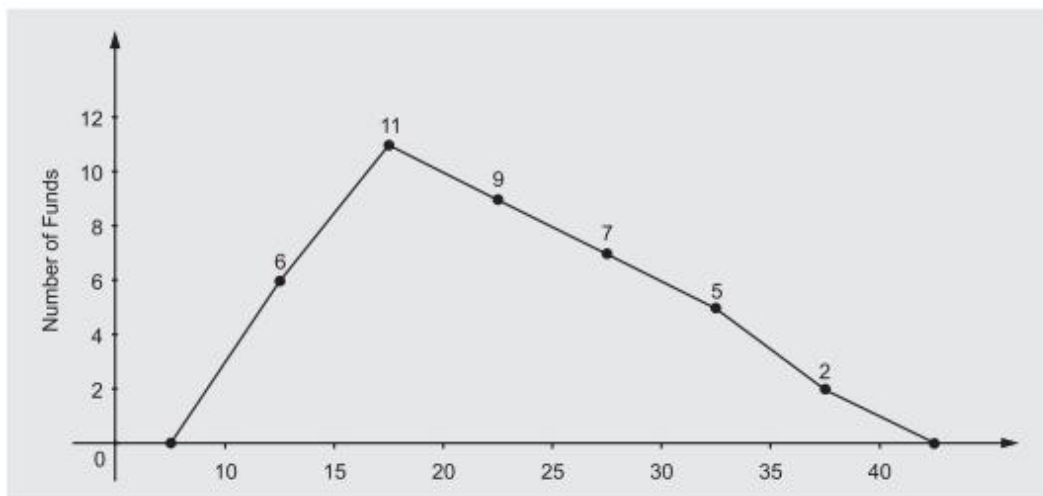
vi. Frequency Polygon

It has been derived from the word "polygon" which means many sides. In statistics, it means a graph of a frequency distribution. A frequency polygon is obtained from a histogram by joining the mid-points of the top of various rectangles with the help of straight lines, as shown in Fig. 3.7. In order that total area under the polygon remains equal to the area under histogram, two arbitrary classes, each with zero frequency, are added on both ends, as shown below Fig 3.7.



vii. Frequency Curve:

It is described as a smooth frequency polygon as shown in Fig. 3.8. A frequency curve is described in terms of its (i) symmetry (skewness) and its (ii) degree of peakedness (kurtosis).



Two frequency distributions can also be compared by superimposing two or more frequency curves provided the width of their class intervals and the total number of frequencies are equal for the given distributions. Even if the distributions to be compared differ in terms of total frequencies, they still can be compared by drawing per cent frequency curves where the vertical axis measures the per cent class frequencies and not the absolute frequencies.

viii. Cumulative Frequency Distribution (Ogive)

It enables us to see how many observations lie above or below certain values rather than merely recording the number of observations within intervals. Cumulative frequency distribution is another method of data presentation that helps in data analysis and interpretation. Table shows the cumulative number of observations below and above the upper boundary of each class in the distribution.

A cumulative frequency curve popularly known as Ogive is another form of graphic presentation of a cumulative frequency distribution. The ogive for the cumulative frequency distribution given in Table is presented in Fig. 3.9.

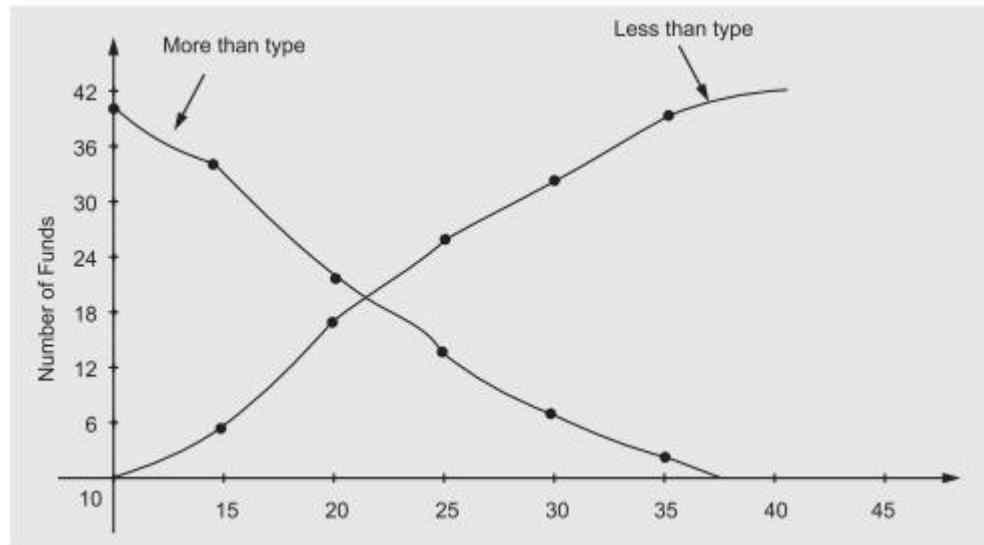
Once cumulative frequencies are obtained, the remaining procedure for drawing curve, called ogive is as usual. The only difference being that the y-axis now has to be so scaled that it accommodates the total frequencies. The x-axis is labelled with the upperclass limits in the case of less than ogive, and the lower class limits in case of more than ogive.

**Example 10**

Mutual Fund price (Rs)	Upper Class Boundary	Number of Funds (f)	Less than cumulative frequency	More than cumulative frequency
10-15	15	6	6	40
15-20	20	11	6+11=17	40-6=34
20-25	25	9	17+9=26	34-11=23
25-30	30	7	26+7=33	23-9=14
30-35	35	5	33+5=38	14-7=7

35-40	40	2	$38+2=40$	$7-5=2$
-------	----	---	-----------	---------

Fig.3.9: Ogive for Mutual Fund Prices



To draw a cumulative 'less than ogive', points are plotted against each successive upper-class limit and the corresponding less than cumulative frequency value. These points are then joined by the series of straight lines and the resultant curve is closed at the bottom by extending it so as to meet the horizontal axis at the real lower limit of the first class interval.

To draw a cumulative 'more than ogive', points are plotted against each successive lower class limit and the corresponding more than cumulative frequency. These points are joined by the series of straight lines and the curve is closed at the bottom by extending it to meet the horizontal axis at the upper limit of the last class interval. Both the types of ogives so drawn are shown in Fig. 3.9.

It may be mentioned that a line drawn parallel to the vertical axis through the point of intersection of the two types of ogives will meet the x-axis at its middle point, and the value corresponding to this point will be the median of the distribution. Similarly, the perpendicular drawn from the point of intersection of the two curves on the vertical axis will divide the total frequencies into two equal parts.

Two ogives, whether less than or more than type, can be readily compared by drawing them on the same graph paper. The presence of unequal class intervals poses no problem in their comparison, as it does in the case of comparison of two frequency polygons. If the total frequencies are not the same in the two distributions, they can be first converted into per cent frequency distributions and then ogives drawn on a single graph paper to facilitate comparison.

ix. Pie Diagram or Pie Chart:

These diagrams are normally used to show the total number of observations of different types in the data set on a percentage basis rather than on an absolute basis through a circle. Usually the largest percentage portion of data in a pie diagram is shown first at 12 o'clock position on the circle, whereas the other observations (in per cent) are shown in clockwise succession in descending order of magnitude. The steps to draw a pie diagram are summarized below:

- i. Convert the various observations (in per cent) in the data set into corresponding degrees in the circle by multiplying each by 3.6 ($360 \div 100$).
- ii. Draw a circle of appropriate size with a compass.
- iii. Draw points on the circle according to the size of each portion of the data with the help of a protractor and join each of these points to the center of the circle.

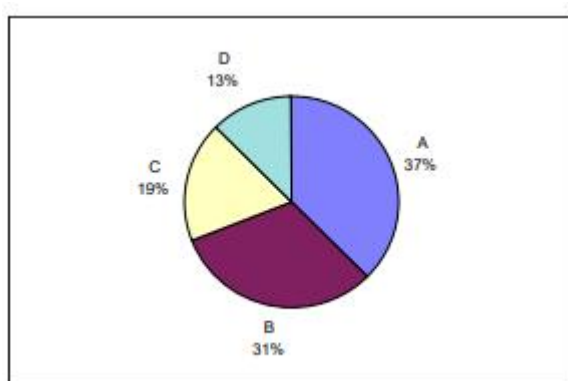
The pie chart has two distinct advantages: (i) it is aesthetically pleasing and (ii) it shows that the total for all categories or slices of the pie adds to 100%



Example 11: Exports of X to A, B, C and D in 1990

Country	Exports	Percentage Share	Degree
A	300	$(300 \times 100) / 800 = 37.50$	$(37.5 \times 360^\circ) / 100 = 135^\circ$
B	250	$(250 \times 100) / 800 = 31.25$	$(31.25 \times 360^\circ) / 100 = 112.5^\circ$
C	150	150	$(18.75 \times 360^\circ) / 100 = 67.5^\circ$
D	100	$(100 \times 100) / 800 = 12.50$	$(12.5 \times 360^\circ) / 100 = 45^\circ$
Total	800	100	360°

Fig 3.10: Pie Diagram Representing Exports of X



2) Two-Dimensional Diagram

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

- Rectangles, where area equals width (or base) multiplied by the length (or height) of the rectangle.
- Squares where area equals square of side (or base).
- Circles where area equals πr^2 , with $\pi = 22/7$ and $r =$ radius.

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

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

Country	Production
USA	130.1
USSR	44.0
UK	16.4
India	3.3

Solution: The given data can be represented graphically by square diagrams. For constructing the sides of the squares, the necessary calculations are shown in table

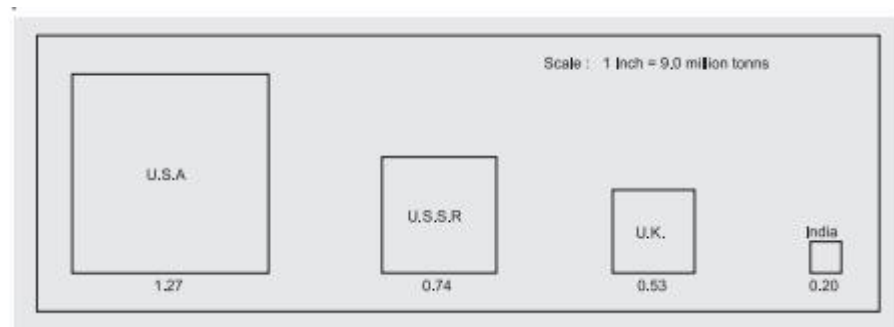
Table: Side of a Square Pertaining to Production of Coal

Country	Production (Million hectare)	Square Root of Production Amount	Side of s Square (One Square inch)
USA	130.1	11.406	1.267
USSR	44.0	6.633	0.737
UK	16.4	4.049	0.449
India	3.3	1.816	0.201

The squares representing the amount of coal production by various countries are shown in Fig.

Now take a scale of 2 cm = 100, so that the first rectangle has dimensions of 2 cm. × 5 cm, the second one has the dimensions of 2 cm × 3.2 cm and the third one has the dimensions of 2 cm × 1.8 cm. After this, we are in a position to draw the rectangles as area diagrams (Fig. 3.11)

Fig. 3.11: Coal Production in Different Countries



Circles:

Circles are alternative to squares to represent data graphically. The circles are also drawn such that their areas are in proportion to the figures represented by them. The circles are constructed in such a way that their centers lie on the same horizontal line and the distance between the circles are equal.

Since the area of a circle is directly proportional to the square of its radius, therefore the radii of the circles are obtained in proportion to the square root of the figures under representation. Thus, the lengths which were used as the sides of the square can also be used as the radii of circles.

Example 12: The following data represent the land area in different countries. Represent this data graphically using suitable diagram.

Country	Land Area (crore acres)
USSR	590.4
China	320.5
USA	19.5
India	81.3

Solution: The data can be represented graphically using circles. The calculations for constructing radii of circles are shown in

Table Radii of Circles Pertaining to Land Area of Countries

Country	Land Acres (crore)	Square Root of land area	Radius of Circles (Inches)
USSR	590.4	24.3	0.81
China	320.5	17.9	0.60
USA	190.5	13.8	0.46
India	81.3	9.0	0.30

The various circles representing the land area of respective countries are shown in Fig. 3.12



iii) Rectangles

Since area of a rectangle is equal to the product of its length and width, therefore while making such type of diagrams both length and width are considered.

Rectangles are suitable for use in cases where two or more quantities are to be compared and each quantity is sub-divided into several components.



Example 12: The following data represent the income of two families A and B. Construct a rectangular diagram.

Item of expenditure	Family A (Monthly income Rs 30,000)	Family B (Monthly income Rs 40,000)
Food	5550	7280
Clothing	5100	6880
House Rent	4800	6480
Fuel and light	4740	6320
Education	4950	6640
Miscellaneous	4860	6640
Total	30,000	40,000

Solution: Converting individual values into percentages taking total income as equal to 100 as shown in Table below

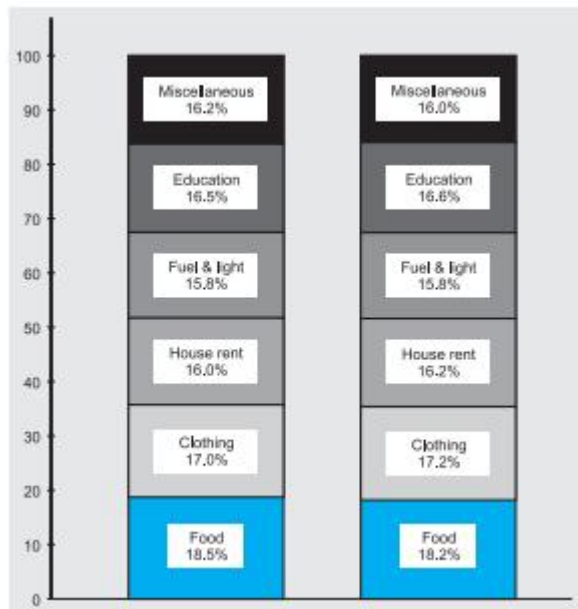
Table Percentage Summary Table Pertaining to Expenses Incurred by Two Families

Item of expenditure	Family A (Monthly income Rs 30,000)			Family B (Monthly income Rs 40,000)		
	Actual	Percentage of	Cumulative	Actual	Percentage of	Cumulative

	Expenses	expenses	percentage	Expenses	expenses	percentage
Food	5550	18.50	18.50	7280	18.20	18.20
Clothing	5100	17.00	33.50	6880	17.20	35.40
House Rent	4800	16.00	51.50	6480	16.20	51.60
Fuel and light	4740	15.80	6.78	6320	15.80	67.40
Education	4950	16.50	83.80	6640	16.60	84.00
Miscellaneous	4860	16.20	100.0	6640	16.00	100.00
Total	30,000		100.00	40,000		100.00

The height of the rectangles shown in Fig. 3.12 is equal to 100. The difference in the total income is represented by the difference on the base line which is in the ratio of 3: 4.

Fig. 3.12: Percentage of Expenditure by Two Families

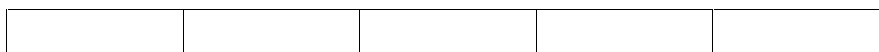


3) Three-Dimensional Diagrams

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

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

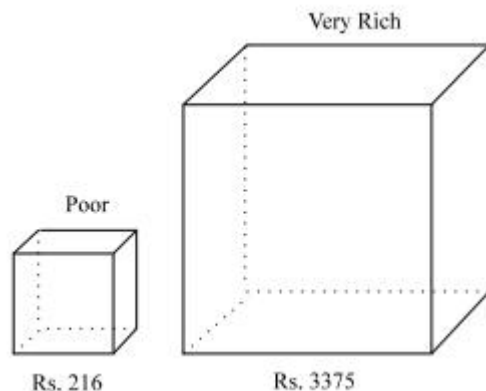
S.NO.	Income Class	Income (rs)	Cube-Root	Side of Cube
1	Poor	216	$(216)^{1/3}=6$	1.5cms
2	Very Rich	3375	$(3375)^{1/3}$	3.75cms



Scale: 1 cm. = 4 units.

- iv. Now draw two cubes with sides equal to 1.5 cms. and 3.75 cms. respectively.

Fig. 3.13: Income levels of Poor and Very Rich people (Rs.)



3.4 Advantages and Limitations of Diagrams (Graphs)

According to P. Maslov, 'Diagrams are drawn for two purposes (i) to permit the investigator to graph the essence of the phenomenon he is observing, and (ii) to permit others to see the results at a glance, i.e. for the purpose of popularization.'

Advantages few of the advantages and usefulness of diagrams are as follows:

- i. **Diagrams give an attractive and elegant presentation:** Diagrams have greater attraction and effective impression. People, in general, avoid figures, but are always impressed by diagrams. Since people see pictures carefully, their effect on the mind is more stable. Thus, diagrams give delight to the eye and add to the spark of interest.
- ii. **Diagrams leave good visual impact:** Diagrams have the merit of rendering any idea readily. The impression created by a diagram is likely to last longer in the mind of people than the effect created by figures. Thus diagrams have greater memorizing value than figures.
- iii. **Diagrams facilitate comparison:** With the help of diagrams, comparisons of groups and series of figures can be made easily. While comparing absolute figures the significance is not clear but when these are presented by diagrams, the comparison is easy. The technique of diagrammatic representation should not be used when comparison is either not possible or is not necessary.
- iv. **Diagrams save time:** Diagrams present the set of data in such a way that their significance is known without loss of much time. Moreover, diagrams save time and effort which are otherwise needed in drawing inferences from a set of figures.
- v. **Diagrams simplify complexity and depict the characteristics of the data:** Diagrams, besides being attractive and interesting, also highlight the characteristics of the data. Large data can easily be represented by diagrams and thus, without straining one's mind, the basic features of the data can be understood and inferences can be drawn in a very short time.

Limitations

We often find tabular and graphical presentations of data in annual reports, newspapers, magazines, bulletins, and so on. But, in spite their usefulness, diagrams can also be misused. A few limitations of these as a tool for statistical analysis are as under:

- i. They provide only an approximate picture of the data.

- ii. They cannot be used as alternative to tabulation of data.
- iii. They can be used only for comparative study.
- iv. They are capable of representing only homogeneous and comparable data.

Summary

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

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

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

Keywords

1. Frequency: It is the number of times a particular variable/ individual or observation (obtained marks in our context) occurs in raw data.
2. Frequency Curve: A curve constructed to smoothen the frequency polygon.
3. Frequency Polygon: A graph of frequency distributed constructed by plotting the frequencies density against the mid-points of the class.
4. Histogram: A graph of frequency distribution where rectangles are awn with area proportionate to the frequency of a class interval and the class interval as the base.
5. Ogive: A graph of frequency distribution depicting cumulative frequencies.

SelfAssessment

1. A pie diagram is also called _____ diagram.
 - A. Bar
 - B. Angular
 - C. multiple bar
 - D. none of the above

2. Which of following is not an example of compressed data?
 - A. data array
 - B. frequency distribution
 - C. histogram
 - D. ogive

3. A graph of a cumulative frequency distribution is called

- A. ogive
 - B. frequency polygon
 - C. frequency curve
 - D. pie diagram
4. A histogram is
- A. A frequency graphs
 - B. A time-series plot
 - C. A graph-plotting mean against standard deviation
 - D. A correlative frequency charts
5. Bar diagram is a
- A. Two-dimensional diagram
 - B. One dimensional diagram
 - C. Three-dimensional diagram
 - D. None of the above
6. The diagram which represents information in a circle is
- A. Bar Diagram
 - B. Pie diagram
 - C. Multiple diagrams
 - D. Sub-divided bar diagram
7. What will be the degree measure of an angle in the pie diagram if a household spends 80% of his income on a good?
- A. 180
 - B. 288
 - C. 90
 - D. 72
8. Histogram represents _____ series.
- A. Individual series
 - B. Discrete series
 - C. Continuous series
 - D. None of the above
9. Frequency polygon is obtained by joining _____.
- A. Mid points of tops of rectangles
 - B. Two end points
 - C. End Points of first-class interval
 - D. End points of last class interval

10. Ogive represents _____ on a graph.
- Individual frequencies
 - Cumulative frequencies
 - Frequency polygon
 - Frequency curve
11. Less than ogive can be used to calculate _____.
- Range
 - Arithmetic mean
 - Mode
 - Median
12. A histogram is a graphical representation of which of the following:
- An ogive
 - A frequency distribution
 - A cumulative relative frequency distribution
 - A stem and leaf plot
13. The two graphical techniques that can be used to represent nominal data are:
- bar chart and histogram
 - pie chart and ogive
 - bar chart and pie chart
 - histogram and ogive
14. All 616 members of a sports club in India were contacted via email and asked whether they thought that Karate should be added to the list of sports currently offered by the club. 146 members said yes, 91 said no, 58 said that they were not sure and 321 did not respond. To represent this information graphically, we could use a:
- histogram
 - box and whisker plot
 - bar graph
 - stem and leaf plot
15. The difference between a histogram and a bar chart is that:
- The histogram reflects qualitative data while the bar chart represents quantitative data.
 - The adjacent rectangles/bars in a histogram have a gap while those for a bar chart do not.
 - The histogram reflects both qualitative and quantitative data while the bar chart represents only qualitative data.
 - the adjacent rectangles/bars in a bar chart have a gap while those for a histogram do not

Answers for SelfAssessment

1. B 2. A 3. A 4. A 5. B

6. B 7. B 8. C 9. A 10. B
 11. D 12. B 13. B 14. C 15. D

Review Questions

- Charts are more effective in attracting attention than other methods of presenting data. Do you agree? Give reasons for your answer
- Diagrams are meant for a rapid view of the relation of different data and their comparisons. Discuss
- The distribution of heights of all students in the commerce department of the university has two peaks or is bimodal. The distribution of the IQ's of the same students, however, has only one peak. How is this possible since the same students are considered in both cases? Explain.
- What are the advantages of using a graph to describe a frequency distribution?
- The following data represent the gross income, expenditure (in Rs lakh), and net profit (in Rs lakh) during the years 1999 to 2002.

	1999-2000	2000-2001	2001-02
Gross income	570	592	632
Gross Expenditure	510	560	610
Net Income	60	32	22

- The following data represent the income and dividend for the year 2000.

Year	Income per Share (in Rs)	Dividend per share (in Rs.)
1995	5.89	3.20
1996	6.49	3.60
1997	7.30	3.85
1998	7.75	3.95
1999	8.36	3.25
2000	9.00	4.45

- Construct a line graph that indicates the income per share for the period 1995–2000.
 - Construct a component bar chart that depicts dividends per share and retained earning per share for the period 1995–2000.
 - Construct a percentage pie chart depicting the percentage of income paid as dividend. Also construct a similar percentage pie chart for the period 1998–2000. Observe any difference between the two pie charts.
- Find a business or economic related data set of interest to you. The data set should be made up of at least 100 quantitative observations.
 - Show the data in the form of a standard frequency distribution.

- (b) Using the information obtained from part (i) briefly describe the appearance of your data.
8. Explain the following terms:
- Line graph
 - Bar diagram
 - Sub-divided or component bar diagram
 - Multiple bar diagram
9. Draw a histogram for the following data relating to the financial outlay on education during the Various Five-Year Plans.

Plans	I	II	III	IV	V	VI	VII
Percentage Outlay	7.6	5.9	6.9	4.9	3.3	2.6	3.6

10. The following data relates to the population of India:

Years	1931	1941	1951	1961	1971	1981	1991
Population (in crores)	27.9	31.9	36.1	43.9	54.8	68.5	72.5

Draw a suitable graph for the above data.



Further Readings

- Gupta, S.P. and M.P. Gupta, 2000. Business Statistics, Sultan Chand & Sons: New Delhi.
- Sinha, S.C. and Dhiman, A.K. 2002. Research Methodology, Vol. 1. EssEss Publication, New Delhi.
- George Argyrions. 2000. Statistics for Social and Health Research with a Guide to SPSS. Sate Publications. New Delhi.

Unit 04: Central Tendency

CONTENTS

Objectives

Introduction

4.1 Concepts of Central Tendency

4.2 Properties of a Good Measure of Central Tendency

4.3 Arithmetic Mean

4.4 Median

4.5 Mode

Summary

Keywords

Self Assessment

Answers for Self Assessment

Review Questions

Further Readings

Objectives

- Describe what central tendency is.
- Define and compute the arithmetic mean, median & mode etc.
- Explain the merits and limitations of different measures of central tendency or averages.

Introduction

As we know that after the classification and tabulation of data one often finds too much detail for many uses that may be made of information available. We, therefore, need further analysis of the tabulated data to draw inference. In this unit we are going to discuss about measures of central tendencies. For the purpose of analysis, very important and powerful tool is a single average value that represents the entire mass of data.

The term average in Statistics refers to a one figure summary of a distribution. It gives a value around which the distribution is concentrated. For this reason, that average is also called the measure of central tendency. For example, suppose Mr. X drives his car at an average speed of 60 km/hr. We get an idea that he drives fast (on Indian roads of course!). To compare the performance of two classes, we can compare the average scores in the same test given to these two classes. Thus, calculation of average condenses a distribution into a single value that is supposed to represent the distribution. This helps both individual assessments of a distribution as well as in comparison with another distribution.

4.1 Concepts of Central Tendency

Measures of central tendency i.e., condensing the mass of data in one single value, enable us to get an idea of the entire data. For example, it is impossible to remember the individual incomes of millions of earning people of India. But if the average income is obtained, we get one single value that represents the entire population. Measures of central tendency also enable us to compare two or more sets of data to facilitate comparison. For example, the average sales figures of April may be compared with the sales figures of previous months.

According to **Professor Bowley**, averages are “statistical constants which enable us to comprehend in a single effort the significance of the whole”. They throw light as to how the values are concentrated in the central part of the distribution.

According to **King and Minium (2013)** described measures of central tendency as a summary figure that helps in describing a central location for a certain group of scores.

According to **Tate (1955)** defined measures of central tendency as “a sort of average or typical value of the items in the series and its function is to summarize the series in terms of this average value”.

For a proper appreciation of various statistical measures used in analyzing a frequency distribution, it is necessary to note that most of the statistical distributions have some common features. If we move from lowest value to the highest value of a variable, the number of items at each successive stage increases till we reach a maximum value, and then as we proceed further, they decrease. The statistical data which follow this general pattern may differ from one variable to another in the following three ways:

- 1) They may differ in the values of the variables around which most of the items cluster (i.e., Average)
- 2) They may differ in the extent to which items are dispersed (i.e., Dispersion).
- 3) They may differ in the extent of departure from some standard distributions called normal distribution (i.e., Skewness and Kurtosis).

Accordingly, there are three sets of statistical measures to study these three kinds of characteristics. At present, however, we are confined to the first set of measures which are called Averages or Measures of Central Tendency or Measures of Location.

In the general pattern of distribution, in the data we may identify a value around which many other items 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 of the data is towards the central part of the data, it is known as Central Tendency. As it indicates the location of the clustering of items, it is also called a measure of location. Just as the title of an essay gives the central theme of the essay, the central tendency of the numerical data gives the central idea of the entire data. Look at Figure 10.1 carefully. It shows the central locations of three different curves A, B and C. You must have noticed that the central locations of curve A and curve C are equal. The central location of curve B lies to the right of those of curves A and C.

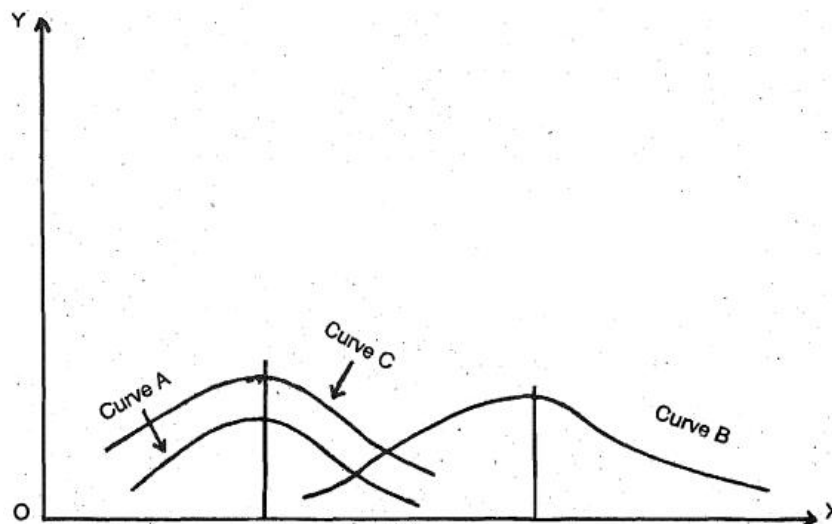


Fig 1: Central Location of Different Curves

4.2 Properties of a Good Measure of Central Tendency

A good measure of central tendency should possess, as far as possible, the following properties,

- i. It should be easy to understand.
- ii. It should be simple to compute.

- iii. It should be based on all observations.
- iv. It should be uniquely defined.
- v. It should be capable of further algebraic treatment.
- vi. It should not be unduly affected by extreme values.

Following are some of the important measures of central tendency which are commonly used in business and industry.

- Arithmetic Mean
- Weighted Arithmetic
- Mean Median
- Quantiles
- Mode
- Geometric
- Mean
- Harmonic
- Mean

4.3 Arithmetic Mean

The arithmetic mean is commonly known as mean. It is a measure of central tendency because other figures of the data congregate around it. Arithmetic mean is obtained by dividing the sum of the values of all observations in the given data set by the number of observations in that set. It is the most commonly used statistical average in the disciplines such as commerce, management, economics, finance, production, etc. The arithmetic mean is also called as simple Arithmetic Mean.

Mean for sample is denoted by symbol 'M or \bar{x} ('x-bar')' and mean for population is denoted by ' μ ' (mu). It is one of the most commonly used measures of central tendency and is often referred to as average. It can also be termed as one of the most sensitive measures of central tendency as all the scores in a data are taken in to consideration when it is computed (Bordens and Abbott, 2011). Further statistical techniques can be computed based on mean, thus, making it even more useful.

Mean is a total of all the scores in data divided by the total number of scores. For example, if there are 100 students in a class and we want to find mean or average marks obtained by them in a psychology test, we will add all their marks and divide by 100, (that is the number of students) to obtain mean.

For Ungrouped Data

Mean (M) is the most familiar and useful measure used to describe the central tendency average of a distribution of scores for any group of individuals, objects or events. It is computed by dividing the sum of the scores by the total number of scores.

Any data that has not been categorized in any way is termed as an ungrouped data. For example, we have an individual who is 25 years old, another who is 30 years old and yet another individual who is 50 years old. These are independent figures and not organized in any way, thus they are ungrouped data.

$$M = \Sigma X / N$$

Where, M = Mean

ΣX = Summation of scores in the distribution

N = Total number of scores.



Example 1:

The scores obtained by 10 students on psychology test are as follows:

58 34 32 47 74 67 35 34 30 39

Solution:

Research Methods and Design

Step 1: In order to obtain mean for the above data we will first add the marks to obtain

$$\Sigma X: 58+ 34+ 32+ 47+ 74+ 67+ 35+ 34+ 30+ 39 = 450$$

Step 2: Now using the formula, we will compute mean

$$M = \Sigma X / N \quad \Sigma X = 450, N = 10 \text{ (Total number of students)}$$

$$\text{Thus, } M = 450 / 10 = 45$$

Thus, the mean obtained for the above data is 45

Computation of Mean for Grouped Data

The formula for computing mean for grouped data is

$$M = \Sigma fX / N$$

Where, M= Mean

Σ = Summation

X= Midpoint of the distribution

f = The respective frequency

N = Total number of scores.

**Example 2:**

A class of 30 students was given a psychology test and the marks obtained by them were categorized in to six categories. The lowest marks obtained were 10 and highest marks obtained were 35. A class interval of 5 was employed. The data is given as follows:

Marks	Frequencies (f)	Midpoint (X)	fX
35- 39	5	37	185
30-34	7	32	224
25-29	5	27	135
20-24	6	22	132
15-19	4	17	68
10-14	3	12	12
	N=30		$\Sigma fX = 780$

The steps followed for computation of mean with grouped data are as follows:

Step 1: The data is arranged in a tabular form with marks grouped in categories with class interval of 5.

Step 2: Once the categories are created, the marks are entered under frequency column based on which category they fall under.

Step 3: The midpoints of the categories are computed and entered under X.

Step 4: fX is obtained by multiplying the frequencies and midpoints for each category.

Step 5: fX for all the categories are added to obtain ΣfX , in case of our example it is obtained as 780

Step 6: The formula $M = \Sigma fX / N$ is used, N is equal to 30.

$$M = \Sigma fX / N \quad M = 780 / 30 = 26$$

Thus, the mean obtained is 26.

Computation of Mean by Shortcut Method (with Assumed mean)

In certain cases data is very large and it is not possible to compute each fX . In such situations, a short cut method with the help of assumed mean can be computed. A real mean can thus be computed with application of correction.

The formula is

$$M = AM + (\Sigma fx' / N \times i)$$

Where,

AM = Assumed mean,

Σ = Summation

i = Class interval

$x' = \{(X - AM) / i\}$, X the midpoint of the scores in the interval

f = the respective frequency of the midpoint

N = The total number of frequencies or students.

Marks	Frequencies (f)	Midpoint (X)	$x' = \{(X - AM) / i\}$	$f x'$
35- 39	5	37	3	15
30-34	7	32	2	14
25-29	5	27	1	5
20-24	6	22	0	0
15-19	4	17	-1	-4
10-14	3	12	-2	-6
	N=30			$\Sigma fx' = 24$

Step 1: We will assume mean (AM) as 22.

Step 2: Difference is obtained between each of the midpoints and the assumed mean and then the same is divided by 'i' that is the class interval (5 in this case), these are then entered under column with heading $x' = \{(X - AM) / i\}$. The x' for 22 will be 0.

Step 3: Frequency (f) is then multiplied with x' to obtain $f x'$.

Step 4: All $f x'$ are added to obtain $\Sigma f x'$, in the present example it is 24.

Step 5: The formula for mean is now applied

$$M = AM + (\Sigma fx' / N \times i)$$

$$M = 22 + (24 / 30 \times 5)$$

$$= 22 + 4 = 26$$

Thus, mean is obtained as 26.

And if you refer to the mean obtained by the direct method and mean obtained with the shortcut method, the mean is the same that is 26.

Computation of Mean by Shortcut Method (with Assumed mean)

Class-Interval	0-8	8-16	16-24	24-32	32-40	40-48
Frequency	10	20	14	16	18	22

Solution:

Here, the intervals are of equal size. So we can apply the step-deviation method, in which

$$A = a + l \cdot \frac{\sum di'fi}{\sum fi}$$

Where a = assumed mean,

l = common size of class intervals

fi = frequency of the ith class interval

di' = m1 - a, mi being the class mark of the ith class interval.

Class-Intervals	Class Marks	Frequency	di = mi - a = mi - 28	di' = di/l = di/8	di'fi
0-8	4	10	-24	-3	-30
8-16	12	20	-16	-2	-40
16-24	20	14	-8	-1	-14
24-32	28	16	0	0	0
32-40	36	18	8	1	18
4-48	44	22	16	2	44
		$\sum fi = 100$			$\sum di'fi = -22$

Therefore, $A = a + l \cdot \frac{\sum di'fi}{\sum fi}$

$$= 28 + 8 \cdot \frac{-22}{100}$$

$$= 28 - 1.76$$

$$= 26.24.$$

Merits and Demerits of Arithmetic Mean

Merits of Arithmetic Mean

1. It utilizes all the observations;
2. It is rigidly defined;
3. It is easy to understand and compute; and
4. It can be used for further mathematical treatments.

Demerits of Arithmetic Mean

1. It is badly affected by extremely small or extremely large values;
2. It cannot be calculated for open end class intervals; and
3. It is generally not preferred for highly skewed distributions.

4.4 Median

Median is that value of the variable which divides the whole distribution into two equal parts. Here, it may be noted that the data should be arranged in ascending or descending order of magnitude. When the number of observations is odd then the median is the middle value of the data. For even number of observations, there will be two middle values. So we take the arithmetic mean of these two middle values. Number of the observations below and above the median, are same. Median is not affected by extremely large or extremely small values (as it corresponds to the middle value) and it is also not affected by open end class intervals. In such situations, it is preferable in comparison to mean.

Median for Ungrouped Data

Mathematically, if x_1, x_2, \dots, x_n are the n observations then for obtaining the median first of all we have to arrange these n values either in ascending order or in descending order. When the observations are arranged in ascending or descending order, the middle value gives the median if n is odd. For even number of observations there will be two middle values. So we take the arithmetic mean of these two values.

$M_d = (n+1/2)^{\text{th}}$ observation ; when n is odd

$$M_d = \frac{\left(\frac{n}{2}\right)^{\text{th}} \text{ observation} + \left(\frac{n}{2}+1\right)^{\text{th}} \text{ observation}}{2} ; \text{When } n \text{ is even}$$



Example 6: Find median of following observations:

Solution: 6, 4, 3, 7, 8

Solution: First we arrange the given data in ascending order as

3, 4, 6, 7, 8

Since, the number of observations i.e. 5, is odd, so median would be the middle value that is 6.



Example 7: Calculate median for the following data:

Solution: 7, 8, 9, 3, 4, 10

Solution: First we arrange given data in ascending order as

3, 4, 7, 8, 9, 10

Here, Number of observations (n) = 6 (even). So we get the median by

$$M_d = \frac{\left(\frac{n}{2}\right)^{\text{th}} \text{ observation} + \left(\frac{n}{2}+1\right)^{\text{th}} \text{ observation}}{2}$$

$$= \frac{\left(\frac{6}{2}\right)^{\text{th}} \text{ observation} + \left(\frac{6}{2}+1\right)^{\text{th}} \text{ observation}}{2}$$

$$= (3^{\text{rd}} \text{ observation} + 4^{\text{th}} \text{ observation})/2$$

$$= (7+8)/2 = 15/2 = 7.5$$

For Ungrouped Data (when frequencies are given)

If x_i is the different value of variable with frequencies f_i then we calculate cumulative frequencies from f_i then median is defined by

$$M_d = \text{Value of variable corresponding to } \left(\frac{\sum f}{2}\right)^{\text{th}} = \left(\frac{N}{2}\right)^{\text{th}} \text{ cumulative frequency}$$

Note: If $N/2$ is not the exact cumulative frequency then value of the variable corresponding to next cumulative frequencies is the median.



Example 8: Find Median from the given frequency distribution

Research Methods and Design

X	F
20	7
40	5
60	4
80	3

Solution: First we find cumulative frequency

X	F	c.f
20	7	7
40	5	12
60	4	16
80	3	19
	$\sum_{i=1}^n f_i = 19$	

Md = Value of the variable corresponding to the

(19/2)th cumulative frequency

=Value of the variable corresponding to 9.5 since 9.5 is not among c.f

So, the next cumulative frequency is 12 and the value of variable against 12 cumulative frequency is 40. So median is 40.

Median for Grouped Data

For class interval, first we find cumulative frequencies from the given frequencies and use the following formula for calculating the median:

$$\text{Median} = L + (N/2 - C) / f * h$$

Where, L = lower class limit of the median class, N = total frequency, C = cumulative frequency of the pre-median class, f = frequency of the median class, and h = width of the median class.

Median class is the class in which the (N/2) th observation falls. If N/2 is not among any cumulative frequency then next class to the N/2 will be considered as median class.

Example 9:

X	F	c.f
0-10	3	3
10-20	5	8
20-30	7	15
30-40	9	24
40-50	4	28

$$\sum_{i=1}^n f_i = 28 = N/2 = 28/2 = 14$$

Since 14 is not among the cumulative frequency so the class with next cumulative frequency i.e. 15, which is 20-30, is the median class.

We have L = lower class limit of the median class = 20 N = total frequency = 28 C = cumulative frequency of the pre median class = 8 f = frequency of the median class = 7 h = width of median class = 10

Now substituting all these values in the formula of Median

$$\begin{aligned}\text{Median} &= L + (N/2 - C)/f * h \\ &= 20 + (14 - 8)/7 * 10 = 28.57\end{aligned}$$

Therefore, median is 28.57.

Merits and Demerits of Median

Merits of Median

1. It is rigidly defined;
2. It is easy to understand and compute;
3. It is not affected by extremely small or extremely large values; and
4. It can be calculated even for open end classes (like "less than 10" or "50 And above").

Demerits of Median

1. In case of even number of observations we get only an estimate of the median by taking the mean of the two middle values. We don't get its exact value;
2. It does not utilize all the observations. The median of 1, 2, 3 is 2. If the observation 3 is replaced by any number higher than or equal to 2 and if the number 1 is replaced by any number lower than or equal to 2, the median value will be unaffected. This means 1 and 3 are not being utilized;
3. It is not amenable to algebraic treatment; and
4. It is affected by sampling fluctuations.

4.5 Mode

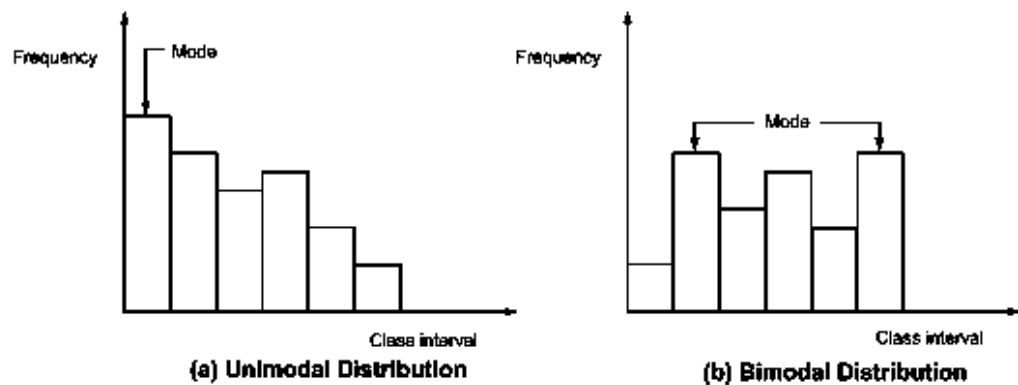
The mode is that value of an observation which occurs most frequently in the data set, that is, the point (or class mark) with the highest frequency.

The concept of mode is of great use to large scale manufacturers of consumable items such as ready-made garments, shoe-makers, and so on. In all such cases it is important to know the size that fits most persons rather than 'mean' size.

There are many practical situations in which arithmetic mean do not always provide an accurate characteristic (reflection) of the data due to the presence of extreme values. For example, in all such statements like 'average man prefers . . . brand of cigarettes', 'average production of an item in a month', or 'average service time at the service counter'. The term 'average' means majority (i.e., mode value) and not the arithmetic mean. Similarly, the median may not represent the characteristics of the data set completely owing to an uneven distribution of the values of observations. For example, suppose in a distribution the values in the lower half vary from 10 to 100 (say), while the same number of observations in the upper half vary from 100 to 7000 (say) with most of them close to the higher limit. In such a distribution, the median value of 100 will not provide an indication of the true nature of the data. Such shortcomings stated above for mean and median are removed by the use of mode, the third measure of central tendency.

The mode is a poor measure of central tendency when most frequently occurring values of an observation do not appear close to the centre of the data. The mode need not even be unique value. Consider the frequency distributions shown in Fig. 3.3(a) and (b). The distribution in Fig. 3.3(a) has its mode at the lowest class and certainly cannot be considered representative of central location. The distribution shown Fig. 3.3(b) has two modes. Obviously neither of these values appear to be representative of the central location of the data. For these reasons the mode has limited use as a

measure of central tendency for decision-making. However, for descriptive analysis, mode is a useful measure of central tendency.



For Ungrouped Data

Mathematically, if x_1, x_2, \dots, x_n are the n observations and if some of the observation are repeated in the data, say i x is repeated highest times then we can say the i x would be the mode value.



Example 10: Find mode value for the given data

2, 2, 3, 4, 7, 7, 7, 7, 9, 10, 12, 12

Solution: First we prepare frequency table as

X	F
2	2
3	1
4	1
7	4
9	1
10	1
12	2

This table shows that 7 have the maximum frequency. Thus, mode is 7.

For Grouped Data:

Data where several classes are given, following formula of the mode is used

$$M_0 = L + \frac{[f_1 - f_0]}{[f_1 - f_0] + [f_1 - f_2]} * h$$

Where, L = lower class limit of the modal class, f_1 = frequency of the modal class, f_0 = frequency of the pre-modal class, f_2 = frequency of the post-modal class, and h = width of the modal class



Example 11:

X	F
0-10	3
10-20	5

30-40	7
40-50	9
50-60	4

Corresponding to highest frequency 9 modal classes is 40-50 and we have

$$L = 40, f_1 = 9, f_2 = 7, f_3 = 4, h = 10$$

Applying the formula,

$$\text{Mode} = 40 + \frac{(9-7)}{(2 \times 9 - 7 - 4)} \times 10$$

$$= 42.86$$

Merits and Demerits of Mode

Merits of Mode

1. Mode is the easiest average to understand and also easy to calculate;
2. It is not affected by extreme values;
3. It can be calculated for open end classes;
4. As far as the modal class is concerned the pre-modal class and the post modal class are of equal width; and
5. Mode can be calculated even if the other classes are of unequal width.

Demerits of Mode

1. It is not rigidly defined. A distribution can have more than one mode;
2. It is not utilizing all the observations;
3. It is not amenable to algebraic treatment; and
4. It is greatly affected by sampling fluctuations.

Summary

In the present unit, we discussed the concept of central tendency. The measures of central tendency were explained as summary figures that help in describing a central location for a certain group of scores. It was further explained as providing information about the characteristics of the data by identifying the value at or near the central location of the data. The functions of measures of tendency besides the characteristics of good measures of central tendency were also discussed. Further, the unit focused on the three measures of central tendency, namely, mean, median and mode. Mean is a total of all the scores in data divided by the total number of scores. It is one of the most frequently used measure of central tendency and is often referred to as an average. It can also be termed as one of the most sensitive measures of central tendency as all the scores in a data are taken in to consideration when it is computed. Median is the middle score in an ordered distribution. Median is a point in any distribution below and above which lie half of the scores. Mode is the score in a distribution that occurs most frequently. Certain distributions are bimodal, where there are two modes. When there are three modes, the term used is trimodal and when there are four or more modes, we use the term multimodal. Though, if the scores in a distribution greatly vary, then it is possible that there is no mode. The properties, advantages and limitations of mean, median and mode were also discussed in detail. Further, the computation of each of these measures of central tendency was also discussed for both ungrouped and grouped data with stepwise explanation.

Keywords

1. **Measures of Central Tendency:** Measures of central tendency can be explained as a summary figure that helps in describing a central location for a certain group of scores.

2. **Mean:** Mean is a total of all the scores in data divided by the total number of scores.
3. **Median:** Median is a point in any distribution below and above which lie half of the scores.
4. **Mode:** Mode is the score in a distribution that occurs most frequently.
5. **Central tendency:** A single value that has a tendency to be somewhere at the center and within range of all values.
6. **Geometric Mean:** For n numbers (X_1, X_2, \dots, X_n) the geometric mean (GM) is defined as the n th root of the product of these n numbers.
7. **Harmonic Mean:** If $x_1, x_2, x_3 \dots x_n$ be a set of n observations then the harmonic mean is defined as the reciprocal of the (arithmetic) mean of the reciprocals of the quantities.

SelfAssessment

1. Which measure of central tendency takes into account the magnitude of scores?
 - A. Median
 - B. Range
 - C. Mode
 - D. Mean

2. Which of the following is not a common measure of central tendency?
 - A. Mean
 - B. Mode
 - C. Median
 - D. Range

3. Which of the following is a characteristic of a mean?
 - A. The sum of deviations from the mean is zero
 - B. It minimizes the sum of squared deviations
 - C. It is affected by extreme scores
 - D. All of the above

4. The total of all the observations divided by the number of observations is called:
 - A. Arithmetic mean
 - B. Geometric mean
 - C. Median
 - D. Harmonic mean

5. The values of extreme items do not influence the average for _____.
 - A. Mean
 - B. Mode
 - C. Median
 - D. None of the above

6. To calculate the median, all the items of a series have to be arranged in a/an _____.
 - A. Descending order

- B. Ascending order
 - C. Ascending or descending order
 - D. None of the above
7. The values of extreme items do not influence the average for_____.
- A. Mean
 - B. Mode
 - C. Median
 - D. None of the above
8. The number of observations smaller than _____ is the same as the number of observations larger than it.
- A. Median
 - B. Mode
 - C. Mean
 - D. None of the above
9. The most frequent observation in a data set is called
- A. Mode
 - B. Median
 - C. Range
 - D. Mean
10. A measurement that corresponds to largest frequency in a set of data is called:
- A. Mean
 - B. Median
 - C. Mode
 - D. Percentile
11. Mode of the series 0, 0, 0, 2, 2, 3, 3, 8, 10 is:
- A. 0
 - B. 2
 - C. 3
 - D. No mode
12. A distribution with two modes is called:
- A. Unimodal
 - B. Bimodal
 - C. Multimodal
 - D. Normal
13. When the values in a series are not of equal importance, we calculate the:
- A. Arithmetic mean

- B. Geometric mean
- C. Weighted mean
- D. Mode

14. Taking the relevant root of the product of all non-zero and positive values are called:

- A. Arithmetic mean
- B. Geometric mean
- C. Harmonic mean
- D. Combined mean

15. The ratio among the number of items and the sum of reciprocals of items is called:

- A. Arithmetic mean
- B. Geometric mean
- C. Harmonic mean
- D. Mode

Answers for Self Assessment

- | | | | | |
|-------|-------|-------|-------|-------|
| 1. D | 2. D | 3. D | 4. A | 5. C |
| 6. C | 7. C | 8. A | 9. A | 10. C |
| 11. A | 12. B | 13. C | 14. C | 15. C |

Review Questions

1. Give a brief description of the different measures of central tendency. Why is arithmetic mean so popular?
2. How would you explain the choice of arithmetic mean as the best measure of central tendency? Under what circumstances would you deem fit the use of median or mode?
3. Suppose the average amount of cash (in pocket, wallet, purse, etc.) possessed by 60 students attending a class is Rs 125. The median amount carried is Rs 90.
 - a) What characteristics of the distribution of cash carried by the students can be explained. Why is mean larger than the median?
 - b) Identify the process or population to which inferences based on these results might apply.
4. What is a statistical average? What are the desirable properties for an average to possess? Mention the different types of averages and state why arithmetic mean is most commonly used amongst them.
5. Given below is the distribution of profits (in '000 rupees) earned by 94 per cent of the retail grocery shops in a city.

Profits	Number of Shops
0-10	0

Unit 04: Central Tendency

10-20	5
20-30	14
30-40	27
40-50	48
50-60	68
60-70	83
70-80	91
80-90	94

6. The management of Doordarshan holds a preview of a new programme and asks viewers for their reaction. The following results by age groups, were obtained.

Age group	Liked the program	Dislike the program
Under 20	140	60
20-39	75	50
40-59	50	50
60 and above	40	20

7. The following are the profit figures earned by 50 companies in the country

Profit (Rs in lakh)	Number of Companies
10 or less	4
20 or less	10
30 or less	30
40 or less	40
50 or less	47
60 or less	50

Calculate

- The median, and
 - The range of profit earned by the middle 80 per cent of the companies. Also verify your results by graphical method.
8. Write a short criticism of the following statement: 'Median is more representative than mean because it is relatively less affected by extreme values'.



Further Readings

- Pagano, R. (2004). *Understanding Statistics in the Behavioural Sciences* (7th edition). Pacific grove, ca: brooks/cole publishing co.
- Guilford, J.P...(1956). *Fundamental Statistics in Psychology and Education*. Mcgraw-hill book company. NY



Web Links

- Black, Thomas R. 1999. *Doing Quantitative Research in the Social Sciences. An Integrated Approach to Research Design, Measurement and Statistics*
- Nachmias, David and ChavaNachmias 1981. *Research Methods in Social Sciences*. St. Martin Press: New York.

Unit 05: Correlation and Linear Bivariate Regression

CONTENTS

Objectives

Introduction

- 5.1 Meaning of Correlation
- 5.2 Significance of Measuring Correlation
- 5.3 Types of Correlations
- 5.4 Degrees Of Correlation
- 5.5 Properties of Correlation
- 5.6 Methods of Correlation
- 5.7 Types of Regression Models
- 5.8 Ordinary Least Squares (OLS)
- 5.9 Assumptions for a Simple Linear Regression Model
- 5.10 Least Squared Methods
- 5.11 Deviations Method

Summary

Keywords

Self Assessment

Answers for Self Assessment

Review Questions

Further Readings

Objectives

- Understand the meaning of correlation
- Recognize the various types of correlation.
- Measure the strength of the linear relationship between two variables.
- Calculate a regression line that allows to predict the value of one of the variables if the value of the other variable is known

Introduction

The statistical methods, discussed so far, are used to analyze the data involving only one variable. Often an analysis of data concerning two or more quantitative variables is needed to look for any statistical relationship or association between them that can describe specific numerical features of the association. The knowledge of such a relationship is important to make inferences from the relationship between variables in a given situation. Examples of correlation problems are found in the study of the relationship between IQ and aggregate percentage marks obtained by a person in SSC examination, blood pressure and metabolism or the relation between height and weight of individuals. In these examples both variables are observed as they naturally occur, since neither variable is fixed at predetermined levels.

5.1 Meaning of Correlation

Correlation refers to the associations between variables. When an association exists between two variables, it means that the average value of one variable changes as there is a change in the value of the other variable. A correlation is the simplest type of association. When a correlation is weak, it means that the average value of one variable change only slightly (only occasionally) in response to changes in the other variable. If there is no association, it means that there is no change in the value of one variable in response to the changes in the other variable. In some cases, the correlation may be positive or it may be negative. A positive correlation means that as one variable increases the other variable increases, e.g. Height of a child and age of the child. Negative correlation implies as one variable increases the other variable decrease, e.g. value of a car and age of the car.

A statistical technique that is used to analyze the strength and direction of the relationship between two quantitative variables is called correlation analysis. A few definitions of correlation analysis are:

An analysis of the relationship of two or more variables is usually called correlation.

– A. M. Tuttle

When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation.

– Croxton and Cowden

The coefficient of correlation, is a number that indicates the strength (magnitude) and direction of statistical relationship between two variables.

- The strength of the relationship is determined by the closeness of the points to a straight line when a pair of values of two variables is plotted on a graph. A straight line is used as the frame of reference for evaluating the relationship.
- The direction is determined by whether one variable generally increases or decreases when the other variable increases.

The importance of examining the statistical relationship between two or more variables can be divided into the following questions and accordingly requires the statistical methods to answer these questions:

- i. Is there an association between two or more variables? If yes, what is form and degree of that relationship?
- ii. Is the relationship strong or significant enough to be useful to arrive at a desirable conclusion?
- iii. Can the relationship be used for predictive purposes, that is, to predict the most likely value of a dependent variable corresponding to the given value of independent variable or variables?

5.2 Significance of Measuring Correlation

The objective of any scientific and clinical research is to establish relationships between two or more sets of observations or variables to arrive at some conclusion which is also near to reality. Finding such relationships is often an initial step for identifying causal relationships. Few advantages of measuring an association (or correlation) between two or more variables are as under:

1. Correlation analysis contributes to the understanding of economic behavior, aids in locating the critically important variables on which others depend, may reveal to the economist the connections by which disturbances spread and suggest to him the paths through which stabilizing forces may become effective. – W. A. Neiswanger
2. The effect of correlation is to reduce the range of uncertainty of our prediction. The prediction based on correlation analysis will be more reliable and near to reality. – Tippett
3. In economic theory we come across several types of variables which show some kind of relationship. For example, there exists a relationship between price, supply, and quantity

Unit 05: Correlation and Linear Bivariate Regression

demanded; convenience, amenities, and service standards are related to customer retention; yield of a crop related to quantity of fertilizer applied, type of soil, quality of seeds, rainfall, and so on. Correlation analysis helps in quantifying precisely the degree of association and direction of such relationships.

4. Correlations are useful in the areas of healthcare such as determining the validity and reliability of clinical measures or in expressing how health problems are related to certain biological or environmental factors. For example, correlation coefficient can be used to determine the degree of inter-observer reliability for two doctors who are assessing a patient's disease.

5.3 Types of Correlations

There are three broad types of correlations:

1. Positive and negative,
2. Linear and non-linear,
3. Simple, partial, and multiple.

1. **Positive Correlation:** If two variables change in the same direction (i.e. if one increases the other also increases, or if one decreases, the other also decreases), then this is called a positive correlation. For example: Advertising and sales. Some other examples of series of positive correlation are:

- i. Price and supply of commodities;
- ii. Amount of rainfall and yield of crops.
- iii. Heights and weights;
- iv. Household income and expenditure.

1.1 **Negative Correlation:** If two variables change in the opposite direction (i.e. if one increases, the other decreases and vice versa), then the correlation is called a negative correlation. For example: T.V. registrations and cinema attendance.

Some other examples of series of negative correlation are:

- i. Volume and pressure of perfect gas;
- ii. Current and resistance [keeping the voltage constant]
- iii. Price and demand for goods.

2. **Linear Correlation:** A linear correlation implies a constant change in one of the variable values with respect to a change in the corresponding values of another variable. In other words, a correlation is referred to as linear correlation when variations in the values of two variables have a constant ratio. The following example illustrates a linear correlation between two variables x and y.

x : 10 20 30 40 50
y : 40 60 80 100 120

When these pairs of values of x and y are plotted on a graph paper, the line joining these points would be a straight line.

In general, two variables x and y are said to be linearly related, if there exists a relationship of the form.

$$y = a + bx$$

Where 'a' and 'b' are real numbers. This is nothing but a straight line when plotted on a graph sheet with different values of x and y and for constant values of a and b. Such relations generally occur in physical sciences but are rarely encountered in economic and social sciences.

2.2 **Non-linear Relationship:** A non-linear (or curvi-linear) correlation implies an absolute change in one of the variable values with respect to changes in values of another variable. In other words, a correlation is referred to as a non-linear correlation when the amount of change in the values of one variable does not bear a constant ratio to the amount of change in the corresponding values of

Research Methods and Design

another variable. The following example illustrates a non-linear correlation between two variables x and y.

x: 8 9 9 10 10 28 29 30
y: 80 130 170 150 230 560 460 600

When these pair of values of x and y are plotted on a graph paper, the line joining these points would not be a straight line, rather it would be curvi-linear.

In other words, the relationship between two variables is said to be non – linear if corresponding to a unit change in one variable, the other variable does not change at a constant rate but changes at a fluctuating rate. In such cases, if the data is plotted on a graph sheet we will not get a straight line curve. For example, one may have a relation of the form.

$$y = a + bx + cx^2$$

3. **Simple Correlation:** The distinction between simple, partial, and multiple correlation is based upon the number of variables involved in the correlation analysis. If only two variables are chosen to study correlation between them, then such a correlation is referred to as simple correlation. A study on the yield of a crop with respect to only amount of fertilizer, or sales revenue with respect to amount of money spent on advertisement, are a few examples of simple correlation.

3.1 **Partial Correlation:** In partial correlation, two variables are chosen to study the correlation between them, but the effect of other influencing variables is kept constant. For example (i) yield of a crop is influenced by the amount of fertilizer applied, rainfall, quality of seed, type of soil, and pesticides, (ii) sales revenue from a product is influenced by the level of advertising expenditure, quality of the product, price, competitors, distribution, and so on. In such cases an attempt to measure the correlation between yield and seed quality, assuming that the average values of other factors exist, becomes a problem of partial correlation.

3.2 **Multiple Correlations:** In multiple correlations, the relationship between more than three variables is considered simultaneously for study. For example, employer-employee relationship in any organization may be examined with reference to, training and development facilities; medical, housing, and education to children's facilities; salary structure; grievances handling system; and so on.

5.4 Degrees Of Correlation

Through the coefficient of correlation, we can measure the degree or extent of the correlation between two variables. On the basis of the coefficient of correlation we can also determine whether the correlation is positive or negative and also its degree or extent.

1. Perfect correlation: If two variables change in the same direction and in the same proportion, the correlation between the two is perfect positive. According to Karl Pearson the coefficient of correlation in this case is +1. On the other hand, if the variables change in the opposite direction and in the same proportion, the correlation is perfect negative. Its coefficient of correlation is -1. In practice we rarely come across these types of correlations.

2. Absence of correlation: If two series of two variables exhibit no relations between them or change in one variable does not lead to a change in the other variable, then we can firmly say that there is no correlation or absurd correlation between the two variables. In such a case the coefficient of correlation is 0.

3. Limited degrees of correlation: If two variables are not perfectly correlated or there is a perfect absence of correlation, then we term the correlation as Limited correlation.

Thus Correlation may be positive, negative or zero but lies with the limits ± 1 . i.e. the value of r is such that $-1 \leq r \leq +1$. The + and – signs are used for positive linear correlations and negative linear correlations, respectively.

1. If x and y have a strong positive linear correlation, r is close to +1. An r value of exactly +1 indicates a perfect positive correlation.

- If x and y have a strong negative linear correlation, r is close to -1 . An r value of exactly -1 indicates a perfect negative correlation.
- If there is no linear correlation or a weak linear correlation, r is close to 0 .

Table 1: Degree and Types of Correlation

Degrees	Positive	Negative
Absence of correlation →	Zero	Zero
Perfect correlation →	+ 1	-1
High degree →	+ 0.75 to + 1	- 0.75 to -1
Moderate degree →	+ 0.25 to + 0.75	- 0.25 to - 0.75
Low degree →	0 to 0.25	0 to - 0.25

Note that r is a dimensionless quantity; that is, it does not depend on the units employed.

5.5 Properties of Correlation

- Coefficient of Correlation lies between -1 and $+1$:

The coefficient of correlation cannot take value less than -1 or more than one $+1$. Symbolically,

$$-1 \leq r \leq +1 \text{ or } |r| < 1$$

- Coefficients of Correlation are independent of Change of Origin:

This property reveals that if we subtract any constant from all the values of X and Y , it will not affect the coefficient of correlation.

- Coefficients of Correlation possess the property of symmetry: The degree of relationship between two variables is symmetric as shown below:

- Coefficient of Correlation is independent of Change of Scale:

This property reveals that if we divide or multiply all the values of X and Y , it will not affect the coefficient of correlation.

- Co-efficient of correlation measures only linear correlation between X and Y .

- If two variables X and Y are independent, coefficient of correlation between them will be zero.

5.6 Methods of Correlation

The following methods of finding the correlation coefficient between two variables x and y are discussed:

- Scatter Diagram method
- Karl Pearson's Coefficient of Correlation method
- Spearman's Rank Correlation method

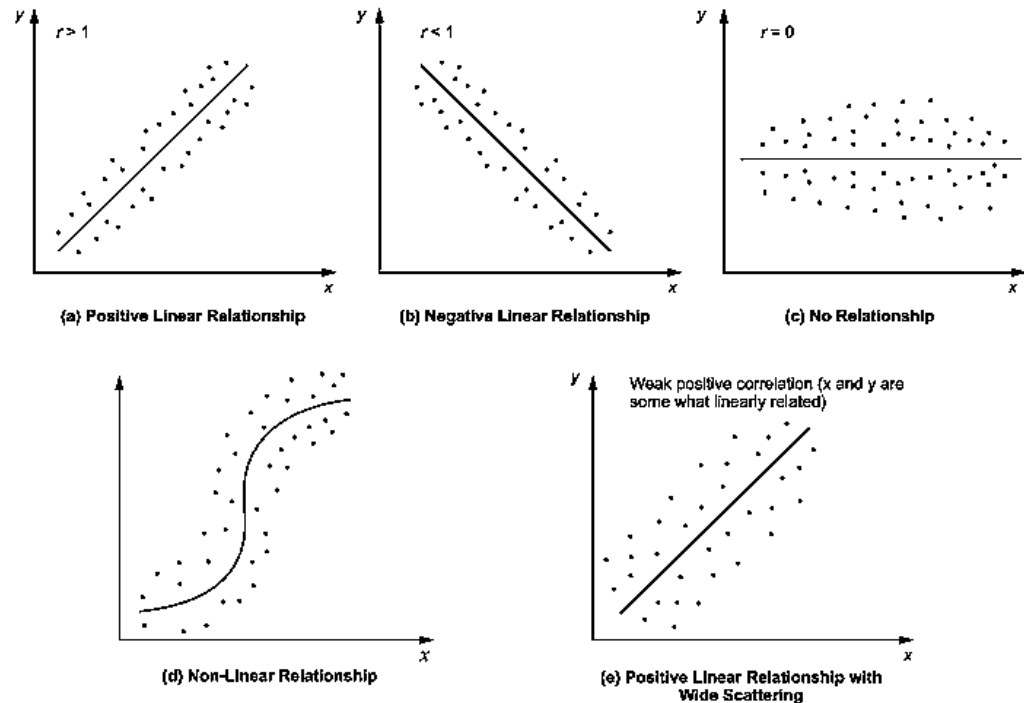
1. Scatter Diagram Method

The scatter diagram method is a quick at-a-glance method of determining of an apparent relationship between two variables, if any. A scatter diagram (or a graph) can be obtained on a graph paper by plotting observed (or known) pairs of values of variables x and y , taking the independent variable values on the x -axis and the dependent variable values on the y -axis.

It is common to try to draw a straight line through data points so that an equal number of points lie on either side of the line. The relationship between two variables x and y described by the data points is defined by this straight line.

In a scatter diagram the horizontal and vertical axes are scaled in units corresponding to the variables x and y , respectively. The pattern of data points in the diagram indicates that the variables are related. If the variables are related, then the dotted line appearing in each diagram describes relationship between the two variables.

Fig 1: Various types of Correlation Coefficient



The patterns depicted in Fig. 1 (a) and (b) represent linear relationships since the patterns are described by straight lines. The pattern in Fig. 1 (a) shows a positive relationship since the value of y tends to increase as the value of x increases, whereas pattern in Fig. 1(b) shows a negative relationship since the value of y tends to decrease as the value of x increases. The pattern depicted in Fig. 1(c) illustrates very low or no relationship between the values of x and y , whereas Fig. 1(d) represents a curvilinear relationship since it is described by a curve rather than a straight line. Figure 13.2(e) illustrates a positive linear relationship with a widely scattered pattern of points. The wider scattering indicates that there is a lower degree of association between the two variables x and y than there is in Fig. 1

2. Karl Pearson's coefficient of correlation

It gives the precise numerical expression for the measure of correlation. It is denoted by ' r '. The value of ' r ' gives the magnitude of correlation and its sign denotes its direction. The mathematical formula for computing r is:

$$r = \frac{\sum xy}{N\sigma_x\sigma_y} \quad 1)$$

Where $x = (X - \bar{X})$, $y = (Y - \bar{Y})$, $\sigma_x = \text{s.d. of } X$, $\sigma_y = \text{s.d. of } Y$

And N number of pairs of observations

$$\text{Since } \sigma_x = \sqrt{\frac{\sum x^2}{N}} \text{ and } \sigma_y = \sqrt{\frac{\sum y^2}{N}}$$

So, Equation 1 can be written as

$$r = \frac{\sum xy}{\sqrt{\frac{\sum y^2}{N}} \sqrt{\frac{\sum x^2}{N}}}$$

By Using Actual Mean

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2} \sqrt{\sum (y - \bar{y})^2}}$$

Unit 05: Correlation and Linear Bivariate Regression

By Assumed Mean Method

$$r = \frac{\sum dxdy - \frac{\sum dxdy}{N}}{\sqrt{\sum dx^2 - \frac{(\sum dx)^2}{N}} \sqrt{\sum dy^2 - \frac{(\sum dy)^2}{N}}}$$

By direct method

$$r = \frac{N\sum XY - \sum X\sum Y}{\sqrt{N\sum X^2 - (\sum X)^2} \cdot \sqrt{N\sum Y^2 - (\sum Y)^2}}$$

Now covariance of X and Y is defined as

$$\text{cov}(X, Y) = (\sum(X - \bar{X})(Y - \bar{Y}))/N$$

Therefore, $r = \text{cov}(X, Y) / \sigma_x \sigma_y$

Where N is the number of pairs of data.

$$d_x = X - A_x$$

$$d_y = Y - A_y$$



Example 1: Calculate the coefficient of correlation between the expenditure on advertising and sales of the company from the following data

Advertising Expenditure	165	166	167	168	167	169	170	172
Sales (in Lakh)	167	168	165	172	168	172	169	171

Solution: N = 8 (pairs of observations)

Advertising Expenditure	Sales (in Lakh)	$x = X_i - \bar{X}$	$y = Y_i - \bar{Y}$	xy	x^2	y^2
165	167	-3	-2	6	9	4
166	168	-2	-1	2	4	1
167	165	-1	-4	4	1	16
167	168	-1	-1	1	1	1
168	172	0	3	0	0	9
169	172	1	3	3	1	9
170	169	2	0	0	4	0
172	171	4	2	8	16	4

Calculation:

$$\bar{X} = \frac{\sum X_i}{N} = 1344/8 = 168 \text{ cm}$$

$$\sigma_x = \sqrt{\frac{\sum x^2}{N}} = \sqrt{\frac{36}{8}}$$

$$\bar{Y} = \frac{\sum Y_i}{N} = 1352/8 = 169 \text{ cm}$$

$$\sigma_y = \sqrt{\frac{\sum y^2}{N}} = \sqrt{\frac{44}{8}}$$

Research Methods and Design

Now, $r = \text{cov}(X, y) / \sigma_x \sigma_y$

$$= 24/8 * \sqrt{\frac{36}{8}} * \sqrt{\frac{44}{8}} = \frac{24}{\sqrt{36 \cdot 44}} = +0.6029$$

Since r is positive and 0.6. This shows that the correlation is positive and moderate (i.e. direct and reasonably good).



Example 2: The following data relates to the Cost and Sales of a Company for the past 10 months

Cost (in 000 rupees)	165	166	167	168	167	169	170	172
Sales (in 000 rupees)	167	168	165	172	168	172	169	171

Find the coefficient of correlation between the two

Solution: Here A = 60, h = 4, B = 60 and k = 3

Cost (in 000 rupees)	Sales (in 000 rupees)	$u = (X_i - A)/h$	$v = (Y_i - B)/d$	uv	u^2	v^2
44	48	-4	-4	16	16	16
80	75	5	5	25	25	25
76	54	4	-2	-8	16	4
48	60	-3	0	-2	4	1
52	63	-2	1	-2	4	1
72	69	3	3	9	9	9
68	72	2	4	8	4	16
56	51	-1	-3	3	1	9
60	57	0	-1	0	0	1
64	66	1	2	2	4	4
		$\Sigma u = 5$	$\Sigma v = 5$	$\Sigma uv = 53$	$\Sigma u^2 = 85$	$\Sigma v^2 = 85$

Calculation

$$r = \frac{\Sigma dx dy - \frac{\Sigma dx \Sigma dy}{N}}{\sqrt{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}} \sqrt{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}}}$$

$$r = \frac{53 - \frac{5 \cdot 5}{10}}{\sqrt{85 - \frac{(5)^2}{10}} \sqrt{85 - \frac{(5)^2}{10}}}$$

$$r = \frac{53 - 2.5}{\sqrt{82.5} \sqrt{82.5}}$$

$$r = 50.5/82.5 = 0.61$$

Assumptions of Using Pearson's Correlation Coefficient

- i. Pearson's correlation coefficient is appropriate to calculate when both variables x and y are measured on an interval or a ratio scale.
- ii. Both variables x and y are normally distributed, and that there is a linear relationship between these variables.
- iii. The correlation coefficient is largely affected due to truncation of the range of values in one or both of the variables. This occurs when the distributions of both the variables greatly deviate from the normal shape.
- iv. There is a cause-and-effect relationship between two variables that influences the distributions of both the variables. Otherwise, correlation coefficient might either be extremely low or even zero.

Advantage and Disadvantages of Pearson's Correlation Coefficient

The correlation coefficient is a numerical number between - 1 and 1 that summarizes the magnitude as well as direction (positive or negative) of association between two variables. The chief limitations of Pearson's method are:

- i. The correlation coefficient always assumes a linear relationship between two variables, whether it is true or not.
- ii. Great care must be exercised in interpreting the value of this coefficient as very often its value is misinterpreted.
- iii. The value of the coefficient is unduly affected by the extreme values of two variable values.
- iv. As compared with other methods the computational time required to calculate the value of r using Pearson's method is lengthy.

3. Spearman's Rank Correlation Coefficient

This method is based on the ranks of the items rather than on their actual values. The advantage of this method over the others is that it can be used even when the actual values of items are unknown. For example, if you want to know the correlation between honesty and wisdom of the boys of your class, you can use this method by giving ranks to the boys. It can also be used to find the degree of agreements between the judgments of two examiners or two judges. The formula is:

$$R = 1 - \frac{6\sum D^2}{N(N^2 - 1)}$$

Where,

R = Rank correlation coefficient

D = Difference between the ranks of two items

N = the number of observations.

Note: $-1 \leq R \leq 1$.

- i. When $R = +1 \Rightarrow$ Perfect positive correlation or complete agreement in the same direction
- ii. When $R = -1 \Rightarrow$ Perfect negative correlation or complete agreement in the opposite direction.
- iii. When $R = 0 \Rightarrow$ No Correlation.



Example 3: Calculate 'R' of 6 students from the following data

Student No.:	1	2	3	4	5	6	7	8	9	10
Rank in Maths	1	3	7	5	4	6	2	10	9	8

Research Methods and Design

Rank in Stata	3	1	4	5	6	9	7	8	10	2
---------------	---	---	---	---	---	---	---	---	----	---

Solution:

Student no.	Rank in Maths (R ₁)	Rank in Stats(R ₂)	D= (R ₁ - R ₂)	D ²
1	1	3	-2	4
2	3	1	2	4
3	7	4	3	9
4	5	5	0	0
5	4	6	-2	4
6	6	9	-3	9
7	2	7	-5	25
8	10	8	2	4
9	9	10	-1	1
10	8	2	6	36
N=10			ΣD = 0	ΣD ² = 96

$$R \approx 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$$

$$R = 1 - \frac{6 * 96}{100(100^2 - 1)}$$

$$R = 1 - \frac{6 * 96}{10 * 99}$$

$$= 0.4181$$



Example 4: The value of Spearman's rank correlation coefficient for a certain number of pairs of observations was found to be 2/3. The sum of the squares of difference between the corresponding ranks was 55. Find the number of pairs.

Solution: We have

$$R = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)} \text{ but } R=2/3 \text{ and } \Sigma D^2=55$$

$$\text{Therefore: } 2/3 = 1 - \frac{6 * 55}{N(N^2 - 1)}$$

$$-1/3 = -\frac{6 * 55}{N(N^2 - 1)}$$

$$N(N^2 - 1) = 6 * 55$$

$$\text{Now } N(N^2 - 1) = 990$$

$$\therefore N(N^2 - 1) = 10 * 99 = 10(100 - 1)$$

$$\therefore N(N^2 - 1) = 10(102 - 1) \Rightarrow N = 10$$

Therefore, there were 10 students.

Advantages and Disadvantages of Spearman's Correlation Coefficient Method

Advantages

- i. This method is easy to understand and its application is simpler than Pearson's method.
- ii. This method is useful for correlation analysis when variables are expressed in qualitative terms like beauty, intelligence, honesty, efficiency, and so on.
- iii. This method is appropriate to measure the association between two variables if the data type is at least ordinal scaled (ranked)
- iv. The sample data of values of two variables is converted into ranks either in ascending order or descending order for calculating degree of correlation between two variables.

Disadvantages

- i. Values of both variables are assumed to be normally distributed and describing a linear relationship rather than non-linear relationship.
- ii. A large computational time is required when number of pairs of values of two variables exceed 30.
- iii. This method cannot be applied to measure the association between two variable grouped data.

Concept of Regression

The statistical technique that expresses the relationship between two or more variables in the form of an equation to estimate the value of a variable, based on the given value of another variable, is called regression analysis. The variable whose value is estimated using the algebraic equation is called dependent (or response) variable and the variable whose value is used to estimate this value is called independent (regressor or predictor) variable. The linear algebraic equation used for expressing a dependent variable in terms of independent variable is called linear regression equation.

The term regression was used in 1877 by Sir Francis Galton while studying the relationship between the height of father and sons. He found that though 'tall father has tall sons', the average height of sons of tall father is x above the general height, the average height of sons is $2x/3$ above the general height. Such a fall in the average height was described by Galton as 'regression to mediocrity'. However, the theory of Galton is not universally applicable and the term regression is applied to other types of variables in business and economics. The term regression in the literary sense is also referred as 'moving backward'.

The regression equation can be written as

$$Y = \alpha + \beta X + \epsilon \quad 1$$

Where,

Y = dependent variable or criterion variable

α = the population parameter for the y-intercept of the regression line, or regression coefficient ($r^* = \sigma_y / \sigma_x$)

$\hat{\alpha}$ = population slope of the regression line or regression coefficient ($r^* \sigma_x / \sigma_y$)

ϵ = the error in the equation or residual

The value of α and $\hat{\alpha}$ are not known, since they are values at the level of population. The population level value is called the parameter. It is virtually impossible to calculate parameter. So we have to estimate it. The two parameters estimated are $\hat{\alpha}$ and $\hat{\alpha}$. The estimator of the α is 'a' and the estimator for $\hat{\alpha}$ is 'b'. So at the sample level equation can be written as

$$Y = a + \beta x + e \quad 2$$

Where,

Y = the scores on Y variable

X = scores on X variable

a = the Y-intercept of the regression line for the sample or regression constant in sample

b = the slope of the regression line or regression coefficient in sample

e = error in prediction of the scores on Y variable, or residual

$$\hat{Y} = a + bX$$

3

Where, \hat{Y} = predicted value of Y in sample. This value is not an actual value but the value of Y that is predicted using the equation $\hat{Y} = a + bX$. So, we can write error as by substituting the in the earlier equation.

$$Y - \hat{Y} = e$$

5.7 Types of Regression Models

The primary objective of regression analysis is the development of a regression model to explain the association between two or more variables in the given population. A regression model is the mathematical equation that provides prediction of value of dependent variable based on the known values of one or more independent variables.

The particular form of regression model depends upon the nature of the problem under study and the type of data available. However, each type of association or relationship can be described by an equation relating a dependent variable to one or more independent variables.

a) Simple and Multiple Regression Models

If a regression model characterizes the relationship between a dependent y and only one independent variable x , then such a regression model is called a simple regression model.

But if more than one independent variable is associated with a dependent variable, then such a regression model is called a multiple regression model. For example, sales turnover of a product (a dependent variable) is associated with multiple independent variables such as price of the product, expenditure on advertisement, quality of the product, competitors, and so on. Now if we want to estimate possible sales turnover with respect to only one of these independent variables, then it is an example of a simple regression model, otherwise multiple regression model is applicable.

b) Linear and Nonlinear Regression Models

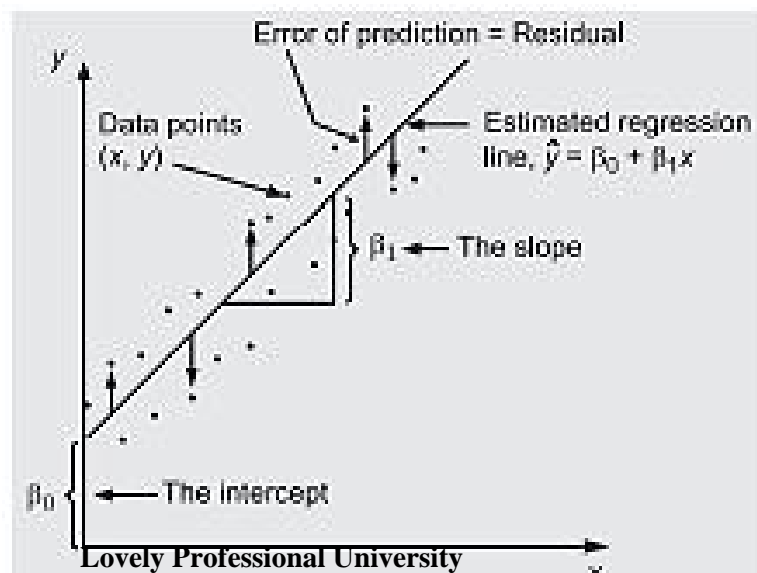
If the value of a dependent (response) variable y in a regression model tends to increase in direct proportion to an increase in the values of independent (predictor) variable x , then such a regression model is called a linear model. Thus, it can be assumed that the mean value of the variable y for a given value of x is related by a straight-line relationship. Such a relationship is called simple linear regression model expressed with respect to the population parameters β_0 and β_1 as:

$$E(y | x) = \beta_0 + \beta_1 x$$

Where,

β_0 = y-intercept that represents mean (or average) value of the dependent variable y when $x = 0$

β_1 = slope of the regression line that represents the expected change in the value of y (either positive or negative) for a unit change in the value of x .



The intercept β_0 and the slope β_1 are unknown regression coefficients. The equation (1) requires to compute the values of β_0 and β_1 to predict average values of y for a given value of x . However, Fig. 14.1 presents a scatter diagram where each pair of values (x_i, y_i) represents a point in a two-dimensional coordinate system. Although the mean or average value of y is a linear function of x , but not all values of y fall exactly on the straight line rather fall around the line.

Since few points do not fall on the regression line, therefore values of y are not exactly equal to the values yielded by the equation: $E(y | x) = \beta_0 + \beta_1 x$, also called line of mean deviations of observed y value from the regression line. This situation is responsible for random error (also called residual variation or residual error) in the prediction of y values for given values of x . In such a situation, it is likely that the variable x does not explain all the variability of the variable y . For instance, sales volume is related to advertising, but if other factors related to sales are ignored, then a regression equation to predict the sales volume (y) by using annual budget of advertising (x) as a predictor will probably involve some error. Thus, for a fixed value of x , the actual value of y is determined by the mean value function plus a random error term as follows:

$Y = \text{Mean value function} + \text{Deviation}$

$$= \beta_0 + \beta_1 x + e = E(y) + e$$

Where e is the observed random error. This equation is also called simple probabilistic linear regression model.

5.8 Ordinary Least Squares (OLS)

In the previous section, we have discussed the simple regression equation with only one regressor variable X and the variable of interest Y . We have also discussed the simple linear regression model with a single regressor variable X . The simple linear regression model has two unknown parameters a and b , which are known as intercept and regression coefficient, respectively. Their values are unknown. Therefore, they must be estimated using sample data. The estimation of the parameters a and b is done by minimizing the error term e .

Let $(Y_1, X_1), (Y_2, X_2), \dots, (Y_n, X_n)$ be n pairs of values in the data. The equation of the simple linear regression model may be written as

$$Y = a + bX + e \quad (4)$$

Where e represents the error term which arises due to the difference of the observed Y and the fitted line $\hat{y} = \hat{a} + \hat{b}x$. We use the method of least squares to minimise the error term e . From equation (4), we may write a simple regression model as

$$Y_i = a + bX_i + e_i \quad i = 1, 2, \dots, n \quad (5)$$

for a sample data of n pairs of values given in terms of

$(Y_i, X_i), (i = 1, 2, \dots, n)$.

We estimate a and b so that the sum of the squares of the differences between the observed values (Y_i) and the points lying on the straight line is minimum, i.e., the sum of squares of the error terms given by

$$E = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2 \quad (6)$$

is minimum. To find the values of a and b for which the sum of squares of the error terms, i.e., E is minimum, we differentiate it with respect to the parameters a and b and equate the results to zero:

$$\frac{\partial E}{\partial a} = -2 \sum_{i=1}^n (Y_i - a - bX_i) = 0 \quad (7)$$

And

$$\frac{\partial E}{\partial b} = -2 \sum_{i=1}^n (Y_i - a - bX_i)X_i = 0 \quad (8)$$

Simplifying equations (7) and (8), we get

$$na + b \sum_{i=1}^n X_i = \sum_{i=1}^n Y_i \quad 9)$$

$$a \sum_{i=1}^n X_i + b \sum_{i=1}^n X_i^2 = \sum_{i=1}^n Y_i X_i \quad 10)$$

Equations (10) and (9) are called the least-squares normal equations. The solution to these normal equations is

$$\hat{Y} - \hat{b}\bar{X} = a \quad 11)$$

Where \bar{Y} and \bar{X} are the averages of Y_i and X_i , respectively. On putting the value of \hat{a} from equation (10) in equation (9), we get

$$\hat{b} = \frac{\sum_{i=1}^n Y_i X_i - (\sum Y_i)(\sum X_i)/n}{\sum_{i=1}^n X_i^2 - (\sum X_i)^2/n} \quad 12)$$

Since the denominator of equation (11) is the corrected sum of squares of X_i , we may rewrite it as

$$SS_x = \sum_{i=1}^n X_i^2 - \frac{(\sum Y_i)^2}{n} = \sum_{i=1}^n (X_i - \bar{X})^2 \quad 13)$$

Similarly, the numerator is the corrected sum of the cross product of X_i and Y_i and may be rewritten as:

$$SS_{XY} = \sum_{i=1}^n X_i Y_i - \frac{\sum Y_i \sum X_i}{n} = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

Therefore, the expression for \hat{b} may be rewritten as

$$\hat{b} = \frac{SS_{XY}}{SS_x}$$

Thus, \hat{a} and \hat{b} are the least squares estimates of the intercept a and slope b , respectively. Therefore, the fitted simple linear regression model is given by

$$\text{Line } \hat{Y} = \hat{a} + \hat{b}x$$

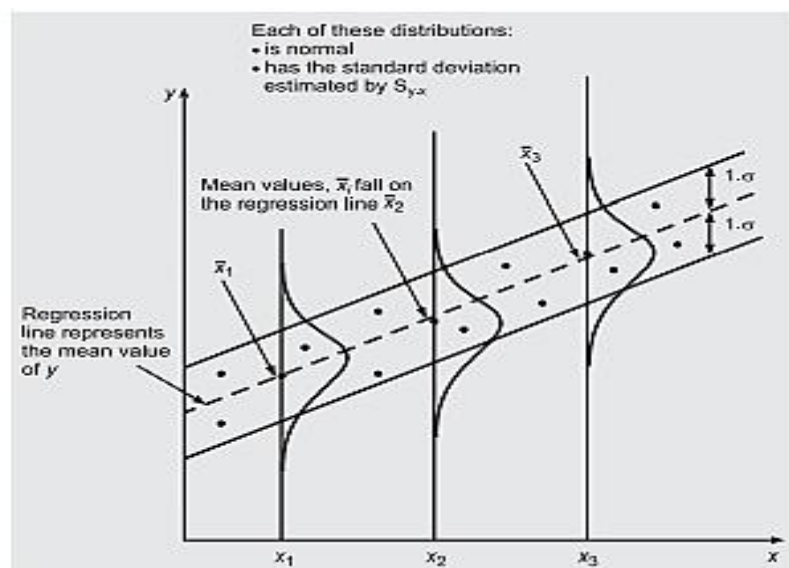
Equation (13) gives a point estimate of the mean of Y for a particular X . The difference between the fitted value \hat{Y}_i and Y_i is known as the residual and is denoted by r_i :

$$r_i = Y_i - \hat{Y}_i, \quad i = 1, 2, \dots, n$$

The role of the residuals and its analysis is very important in regression modeling.

5.9 Assumptions for a Simple Linear Regression Model

To make valid statistical inference using regression analysis, we make certain assumptions about the bivariate population from which a sample of paired observations is drawn and the manner in which observations are generated. These assumptions form the basis for application of simple linear regression models.



Assumptions

1. The relationship between the dependent variable y and independent variable x exists and is linear. The average relationship between x and y can be described by a simple linear regression equation $y = a + bx + e$, where e is the deviation of a particular value of y from its expected value for a given value of independent variable x .
2. For every value of the independent variable x , there is an expected (or mean) value of the dependent variable y and these values are normally distributed. The mean of these normally distributed values fall on the line of regression.
3. The dependent variable y is a continuous random variable, whereas values of the independent variable x are fixed values and are not random.
4. The sampling error associated with the expected value of the dependent variable y is assumed to be an independent random variable distributed normally with mean zero and constant standard deviation. The errors are not related with each other in successive observations.
5. The standard deviation and variance of expected values of the dependent variable y about the regression line are constant for all values of the independent variable x within the range of the sample data.
6. The value of the dependent variable cannot be estimated for a value of an independent variable lying outside the range of values in the sample data.

Regression coefficients

To estimate values of population parameter β_0 and β_1 , under certain assumptions, the fitted or estimated regression equation representing the straight line regression model is written as:

$$\hat{y} = a + bx$$

\hat{y}
= estimated average (mean) value of dependent variable y for a given value of independent variable x .

a or b_0 = y -intercept that represents average value of \hat{y}

b = slope of regression line that represents the expected change in the value of y for unit change in the value of x .

To determine the value of \hat{y} for a given value of x , this equation requires the determination of two unknown constants a (intercept) and b (also called regression coefficient). Once these constants are calculated, the regression line can be used to compute an estimated value of the dependent variable y for a given value of independent variable x .

The regression coefficient ' b ' is also denoted as:

- b_{yx} (regression coefficient of y on x) in the regression line, $y = a + bx$
- b_{xy} (regression coefficient of x on y) in the regression line, $x = c + dy$

Properties of regression coefficients

1. The correlation coefficient is the geometric mean of two regression coefficients, that is, $r = \sqrt{b_{yx} \cdot b_{xy}}$
2. If one regression coefficient is greater than one, then other regression coefficient must be less than one, because the value of correlation coefficient r cannot exceed one. However, both the regression coefficients may be less than one.
3. Both regression coefficients must have the same sign (either positive or negative). This property rules out the case of opposite sign of two regression coefficients.
4. The correlation coefficient will have the same sign (either positive or negative) as that of the two regression coefficients. For example, if $b_{yx} = -0.664$ and $b_{xy} = -0.234$, then $r = -\sqrt{0.664 \cdot 0.234} = -0.394$.
5. The arithmetic mean of regression coefficients b_{xy} and b_{yx} is more than or equal to the correlation coefficient r , that is, $(b_{yx} + b_{xy})/2 \geq r$. For example, if $b_{yx} = -0.664$ and $b_{xy} = -$

0.234, then the arithmetic mean of these two values is $(-0.664 - 0.234)/2 = -0.449$, and this value is more than the value of $r = -0.394$.

6. Regression coefficients are independent of origin but not of scale.

5.10 Least Squared Methods



Example 1: Use least squares regression line to estimate the increase in sales revenue expected from an increase of 7.5 per cent in advertising expenditure.

Table 1

Firm	Annual Percentage Increase in advertising expenditure	Annual Percentage Increase in Sales Revenue
A	1	1
B	3	2
C	4	2
D	6	4
E	8	6
F	9	8
G	11	8
H	14	9

Solution: Assume sales revenue (y) is dependent on advertising expenditure (x). Calculations for regression line using following normal equations are shown in Table 3

$$\Sigma y = na + b\Sigma x \text{ and}$$

$$\Sigma xy = a \Sigma x + b\Sigma x^2$$

Table 2: Calculations of Normal Equations

Sales revenue	Advertising expenditure	Annual Increase in Sales Revenue	Percentage in Sales
1	1	1	1
2	3	9	6
2	4	16	8
4	6	36	24
6	8	64	48
8	9	81	72
8	11	121	88
9	14	196	126
40	56	524	373

Unit 05: Correlation and Linear Bivariate Regression

$$\Sigma y = na + b\Sigma x \text{ or } 40 = 8a + 56b$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \text{ or } 373 = 56a + 524b$$

Solving these equations, we get

$$a = 0.072 \text{ and } b = 0.704$$

Substituting these values in the regression equation $y = a + bx = 0.072 + 0.704x$

For $x = 7.5\%$ or 0.075 increase in advertising expenditure, the estimated increase in sales revenue will be

$$y = 0.072 + 0.704(0.075) = 0.1248 \text{ or } 12.48\%$$



Example 2: The owner of a small garment shop is hopeful that his sales are rising significantly week by week. Treating the sales for the previous six weeks as a typical example of this rising trend, he recorded them in Rs 1000's and analyzed the results in table 3

Table 3

Week	Sales
1	2.69
2	2.62
3	2.80
4	2.70
5	2.75
6	2.81

Fit a linear regression equation to suggest to him the weekly rate at which his sales are rising and use this equation to estimate expected sales for the 7th week.

Solution: Assume sales (y) is dependent on weeks (x). Then the normal equations for regression equation: $y = a + bx$ are written as:

$$\Sigma y = na + b\Sigma x \text{ and}$$

$$\Sigma xy = a \Sigma x + b\Sigma x^2$$

Calculations for sales during various weeks are shown in Table 4

Table 4: Calculations of Normal Equations

Week (x)	Sales(y)	x^2	xy
1	2.69	1	2.69
2	2.62	4	5.24
3	2.80	9	8.40
4	2.70	16	10.80
5	2.75	25	13.75
6	2.81	36	16.86
21	16.37	91	57.74

The gradient 'b' is calculated as:

$$\hat{b} = \frac{SS_{XY}}{SS_{XX}} = 0.445/17.5 = 0.25$$

$$SS_{XY} = \Sigma y - \frac{\Sigma x \Sigma y}{n} = (57.74 - 21 * 16.37)/6 = 0.445$$

$$SS_{XX} = \Sigma x^2 - \frac{\Sigma x^2}{n} = 91 - (21)^2/6 = 17.5$$

The intercept 'a' on the y-axis is calculated as

$$a = \bar{y} - b\bar{x} = \frac{16.37}{6} - 0.025 * \frac{21}{6}$$

$$= 2.728 - 0.025 * 3.5 = 2.64$$

Substituting the values a = 2.64 and b = 0.025 in the regression equation, we have

$$y = a + bx = 2.64 + 0.025x$$

$$\text{For } x = 7, \text{ we have } y = 2.64 + 0.025(7) = 2.815$$

Hence the expected sales during the 7th week are likely to be Rs 2.815 (in Rs 1000's).

5.11 Deviations Method

Calculations to least squares normal equations become lengthy and tedious when values of x and y are large. Thus, the following two methods may be used to reduce the computational time.

a) Deviations Taken from Actual Mean Values of x and y: If deviations of actual values of variables x and y are taken from their mean values \bar{x} and \bar{y} , then the regression equations can be written as:

a.1) Regression equation of y on x

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

Where, b_{yx} = regression coefficient of y on x

The value of b_{yx} can be calculated using the formula

$$b_{yx} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

a.2) Regression equation of x on y

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

Where, b_{xy} = regression coefficient of x on y

The value of b_{xy} can be calculated formula

$$b_{xy} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(y - \bar{y})^2}$$

b) Deviations Taken from Assumed Mean Values for x and y

If mean value of either x or y or both are in fractions, then we must prefer to take deviations of actual values of variables x and y from their assumed means.

b.1) Regression equation of y on x

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

where b_{yx} =

$$\frac{n_1 \Sigma dx dy - (\Sigma dx)(\Sigma dy)}{n \Sigma dx^2 - (\Sigma dx)^2}$$

n = number of observations

dx = x - A; A is assumed mean of x

dy = y - B; B is assumed mean of y

b.2) Regression equation of x on y

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

where $b_{xy} =$

$$\frac{n_1 \sum dx dy - (\sum dx)(\sum dy)}{n \sum d_x^2 - (\sum dy)^2}$$

n = number of observations

$dx = x - A$; A is assumed mean of x

$dy = y - B$; B is assumed mean of y

c) Regression Coefficients in Terms of Correlation Coefficient

If deviations are taken from actual mean values, then the values of regression coefficients can be alternatively calculated as follows:

$$b_{yx} = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$$

$$= \frac{\text{Covariance}(x,y)}{\sigma_x^2} = r \cdot \sigma_y / \sigma_x$$

And

$$b_{xy} = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(y - \bar{y})^2}$$

$$= \frac{\text{Covariance}(x,y)}{\sigma_y^2} = r \cdot \sigma_x / \sigma_y$$



Example 3: The following data relate to the scores obtained by 9 salesmen of a company in an intelligence test and their weekly sales (in Rs 1000's)

Table 5:

Salesman	A	B	C	D	E	F	G	H	I
Test Scores	50	60	50	60	80	50	80	40	70
Weekly Sales	30	60	40	50	60	30	70	50	60

a) Obtain the regression equation of sales on intelligence test scores of the salesmen.

(b) If the intelligence test score of a salesman is 65, what would be his expected weekly sales.

Solution: Assume weekly sales (y) as dependent variable and test scores (x) as independent variable. Calculations for the following regression equation are shown in Table 5

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

Weekly Sales, x	$dx=x-60$	d_x^2	Test score, y	$dy=y-50$	d_y^2	dxd_y
50	-10	100	30	-20	400	200
60	0	0	60	10	100	0
50	-10	100	40	-10	100	100
60	0	0	50	0	0	0

Research Methods and Design

80	20	400	60	10	100	200
50	-10	100	30	-20	400	400
80	20	400	70	20	400	400
40	-20	400	50	0	0	0
70	10	100	60	10	100	100
540	0	1600	450	0	1600	1200

$$a) \bar{x} = \Sigma x/n = 540/9 = 60;$$

$$\bar{y} = \Sigma y/n = 450/9 = 50$$

$$b_{yx} = \frac{\Sigma dx dy - (\Sigma dx)(\Sigma dy)}{\Sigma dx^2 - (\Sigma dx)^2} = 1200/1600 = 0.75$$

Substituting values in the regression equation, we have

$$y - 50 = 0.75 (x - 60) \text{ or } y = 5 + 0.75x$$

For test score $x = 65$ of salesman, we have

$$y = 5 + 0.75 (65) = 53.75$$

Hence we conclude that the weekly sales are expected to be Rs 53.75 (in Rs 1000's) for a test score of 65.



Example 4: The following data give the ages and blood pressure of 10 women

Age	56	42	36	47	49	42	60	72	63	55
Blood Pressure	147	125	118	128	145	140	155	160	149	150

- Find the correlation coefficient between age and blood pressure.
- Determine the least squares regression equation of blood pressure on age.
- Estimate the blood pressure of a woman whose age is 45 years.

Solution: Assume blood pressure (y) as the dependent variable and age (x) as the independent variable. Calculations for regression equation of blood pressure on age are shown in Table 5

Table 5: Calculations for Regression Equation

Age, x	$d_x = x - 49$	d^2_x	Blood, y	$d_y = y - 145$	d^2_y	$d_x d_y$
56	7	49	147	2	4	14
42	-7	49	125	-20	400	140
36	-13	169	118	-27	729	351
47	-2	4	128	-17	289	34
49	0	0	145	0	0	0
42	-7	49	140	-5	25	35

Unit 05: Correlation and Linear Bivariate Regression

60	11	121	155	10	100	110
72	23	529	160	15	225	345
63	14	196	149	4	16	56
55	6	36	150	5	25	30
522	32	1202	1414	-33	1813	1115

a) Coefficient of correlation between age and blood pressure is given by

$$\frac{n\sum dx dy - \sum dx \sum dy}{\sqrt{n\sum dx^2 - (\sum dx)^2} \sqrt{n\sum dy^2 - (\sum dy)^2}}$$

$$= 10(1115) - 32(-33) / \sqrt{10(1202) - (32)^2} \sqrt{10(1813) - (-33)^2}$$

$$= 12206 / 13689 = 0.892$$

We may conclude that there is a high degree of positive correlation between age and blood pressure.

b) The regression equation of blood pressure on age is given by

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$a) \bar{x} = \sum x / n = 522 / 10 = 52.2;$$

$$\bar{y} = \sum y / n = 1417 / 10 = 141.7$$

$$b_{yx} = \frac{\sum dx dy - (\sum dx)(\sum dy)}{\sum dx^2 - (\sum dx)^2} = 10(1115) - 32(-33) / (10(1202) - (32)^2) = 12206 / 10996 = 1.11$$

Substituting these values in the above equation, we have $y - 141.7 = 1.11(x - 52.2)$ or $y = 83.758 + 1.11x$.

This is the required regression equation of y on x.

b) For women, whose age is 45, the estimated average blood pressure will be

$$y = 83.758 + 1.11(45) = 83.758 + 49.95 = 133.708$$

Hence, the likely blood pressure of a woman of 45 years is 134.

Summary

In this unit the concept of correlation or the association between two variables has been discussed. A scatter plot of the variables may suggest that the two variables are related but the value of the Pearson correlation coefficient r quantifies this association. The correlation coefficient r may assume values between -1 and 1. The sign indicates whether the association is direct (+ve) or inverse (-ve). A numerical value of r equal to unity indicates perfect association while a value of zero indicates no association.

Keywords

1. Correlation: Degree of association between two variables.
2. Correlation Coefficient: A number lying between -1 (Perfect negative correlation) and +1 (perfect positive correlation) to quantify the association between two variables.
3. Scatter Diagram: An ungrouped plot of two variables, on the X and Y axes.
4. Rank Correlation: There happen to be many occasions when it may be convenient, economic or even possible to give values to variables. However, various items can be ranked. In such cases, a rank correlation coefficient may be used.

5. **Positive Correlation:** If two variables change in the same direction.
6. **Multiple Correlations:** In multiple correlations, the relationship between more than three variables is considered simultaneously for study.

Self Assessment

1. Correlation analysis is a.....
 - A. Univariate analysis
 - B. Bivariate analysis
 - C. Multivariate analysis
 - D. Both b and c

2. If the ratio of change in one variable is equal to the ratio of change in the other variable, then the correlation is said to be.....
 - A. Linear
 - B. Non-linear
 - C. Curvilinear
 - D. None of these

3. Regression coefficient is independent of.....
 - A. Origin
 - B. Scale
 - C. Both a and b
 - D. Neither origin nor scale

4. Study of correlation among three or more variables simultaneously is called.....
 - A. Partial correlation
 - B. Multiple correlation
 - C. Nonsense correlation
 - D. Simple correlation

5. The unit of Coefficient of correlation is.....
 - A. Percentage
 - B. Ratio
 - C. Same unit of the data
 - D. No unit

6. Correlation analysis between one dependent variable with one independent variable by keeping the other independent variables as constant is called.....
 - A. Partial correlation
 - B. Multiple correlations
 - C. Nonsense correlation
 - D. Simple correlation

Unit 05: Correlation and Linear Bivariate Regression

7. In a correlation analysis, if $r = 0$, then we may say that there is between variables.
- A. No correlation
 - B. Linear correlation
 - C. Perfect correlation
 - D. none of these
8. The coefficient of correlation is independent of
- A. Change of scale only
 - B. Change of origin only
 - C. Both change of origin and scale
 - D. Neither change of origin nor change of scale
9. If the correlation coefficient between the variables X and Y is ρ , the correlation coefficient between X^2 and Y^2 is
- A. ρ
 - B. ρ^2
 - C. 0
 - D. 1
10. Spearman's Rank Correlation Coefficient is usually denoted by.....
- A. k
 - B. r
 - C. S
 - D. R
11. Pearson an correlation coefficient if denoted by the symbol.....
- A. K
 - B. r
 - C. R
 - D. None of these
12. If the dots in a scatter diagram fall on a narrow band, it indicates adegree of correlation.
- A. Zero
 - B. High
 - C. Low
 - D. None of these
13. If all the dots of a scatter diagram lie on a straight line falling from left bottom corner to the right upper corner, the correlation is called.....
- A. Zero correlation
 - B. High degree of positive correlation

Research Methods and Design

- C. Perfect negative correlation
D. Perfect positive correlation
14. Which of following is/ are characteristics of Karl Pearson's coefficient of correlation
A. Indication of degree
B. Indicators of the direction
C. A satisfactory measure
D. All of the above
15. Karl Pearson coefficient of Correlation between two variables is
A. The product of their standard deviation
B. The square root of the product of their regression coefficients
C. The co-variance between the variables
D. None of the above

Answers for Self Assessment

1. D 2. A 3. A 4. B 5. D
6. A 7. A 8. C 9. B 10. D
11. B 12. B 13. D 14. D 15. B

Review Questions

1. The data relating to variable X and Y is given below:

X	72	73	75	76	77	78	79	80	80	81	82	83	84	85	86	88
Y	45	38	41	35	31	40	25	32	36	29	34	38	26	32	28	27

- (a) Sketch a scatter plot.
(b) Compute the correlation coefficient, r.

2. Calculate and analyze the correlation coefficient between the number of study hours and the number of sleeping hours of different students.

Number of Study hours	2	4	6	8	10
Number of sleeping hours	10	9	8	7	6

3. A trainee manager wondered whether the length of time his trainees revised for an examination had any effect on the marks they scored in the examination. Before the exam, he asked a random sample of them to honestly estimate how long, to the nearest hour, they had spent revising. After the examination he investigated the relationship between the two variables.

Trainee	A	B	C	D	E	F	G	H	I	J
Revision time	4	9	10	14	4	7	12	22	1	17

Unit 05: Correlation and Linear Bivariate Regression

Exam mark	31	58	65	73	37	44	60	91	21	84
-----------	----	----	----	----	----	----	----	----	----	----

- (a) Plot the scatter diagram in order to inspect the data.
- (b) Calculate the correlation coefficient
4. What do you understand by the term correlation? Explain how the study of correlation helps in forecasting demand of a product?
5. The coefficient of correlation between two variables x and y is 0.3. The covariance is 9. The variance of x is 16. Find the standard deviation of y series.
6. The correlation between the price of two commodities x and y in a sample of 60 is 0.68. Could the observed value have arisen?
- (a) From an uncorrelated population?
- (b) from a population in which true correlation was 0.8?
7. A small retail business has determined that the correlation coefficient between monthly expenses and profits for the past year, measured at the end of each month, is $r = 0.56$. Assuming that both expenses and profits are approximately normal, test at $\alpha = 0.05$ level of significance the null hypothesis that there is no correlation between them.
8. Define correlation coefficient ' r ' and give its limitations. What interpretation would you give if told that the correlation between the number of truck accidents per year and the age of the driver is $(-)$ 0.60 if only drivers with at least one accident are considered?
9. Does correlation always signify a cause-and-effect relationship between the variables?
10. What is coefficient of rank correlation? Bring out its usefulness. How does this coefficient differ from the coefficient of correlation?
11. The following calculations have been made for prices of twelve stocks (x) at the Calcutta Stock Exchange on a certain day along with the volume of sales in thousands of shares (y). From these calculations find the regression equation of price of stocks on the volume of sales of shares. $\Sigma x = 580$, $\Sigma y = 370$, $\Sigma xy = 11494$, $\Sigma x^2 = 41658$, $\Sigma y^2 = 17206$.
12. The following table gives the aptitude test scores and productivity indices of 10 workers selected at random:
- Aptitude scores (x): 60 62 65 70 72 48 53 73 65 82
- Productivity index (y): 68 60 62 80 85 40 52 62 60 81
- Calculate the two regression equations and estimate (a) the productivity index of a worker whose test score is 92, (b) the test score of a worker whose productivity index is 75



Further Readings

- Nagar, A.L. and R.K Das, 1989: Basic Statistics, Oxford University Press, Delhi.
- Goon, A.M., M.K. Gupta and B.'Dasgupta, 19.87: Basic Statistics, the World Press Pvt. Ltd., and Calcutta.



Web Links

- Edwards, B. 1980. The Readable Maths and Statistics Book, George Allen and Unwin: London.
- Makridakis, S. and S. Wheelwright, 1978. Interactive Forecasting: Univariate and Multivariate Methods, Holden-Day: San Francisco.

Unit 06: Sampling and Sampling Distribution

CONTENTS

Objectives

Introduction

6.1 Sampling: Concept and Significance

6.2 Reasons of Sample Survey

6.3 Concepts in Sampling

6.4 Principles of Sampling

6.5 Types of Sampling

6.6 Non-Probability Sampling

6.7 Choice of Sampling Methods

6.8 Errors In Sampling

6.9 Role of Sampling Theory

Summary

Keywords

Self Assessment

Answers for Self Assessment

Review Questions

Further Readings

Objectives

- Discuss the meaning and importance of sampling
- Describe the steps and criteria involved in selecting a sampling procedure
- Distinguish different types of sampling

Introduction

Sampling has been an age-old practice in everyday life. Whenever we want to buy a huge quantity of a commodity, we decide about the total lot by simply examining a small fraction of it. It has been established that the sample survey if planned properly, can give very precise information. Since in surveys a part of the population is only surveyed and inference is drawn about the whole population, the results likely to be different from the population values. But the advantage with the sample survey is that this type of error can be measured and controlled and it can be eliminated to great extent by employing properly trained persons in surveys. The other advantage of sample surveys is that it is less time consuming and involves less cost. Usually, the population is too large for the researcher to attempt to survey all of its members. A small, but carefully chosen sample can be used to represent the population. The sample reflects the characteristics of the population from which it is drawn.

6.1 Sampling: Concept and Significance

The process of selecting a sample from a population is called sampling. In sampling, a representative sample or portion of elements of a population or process is selected and then analyzed. Based on sample results, called sample statistics, statistical inferences are made about the population characteristic. For instance, a political analyst selects specific or random set of peoples for interviews to estimate the proportion of the votes that each candidate may get from the

population of voters; an auditor selects a sample of vouchers and calculates the sample mean for estimating population average amount; or a doctor examines a few drops of blood to draw conclusions about the nature of disease or blood constitution of the whole body.

According to Levin and Rubin, statisticians use the word, population, to refer not only to people, but, to all items that have been chosen for study. They use the word, sample, to describe a portion chosen from the population.

According to Croach and Housden, a sample is a limited number taken from a large group for testing and analysis, on the assumption that the sample can be taken as representative for the whole group.

According to Boyce, sampling makes an estimate about some of the characteristics of a population. To sample is to make a judgment or a decision about something after experiencing just part of it.

6.2 Reasons of Sample Survey

A census is a count of all the elements in a population. Few examples of census are: population of eligible voters; census of consumer preference to a particular product, buying habits of adult Indians. Some of the reasons to prefer sample survey instead of census are given below.

1. ***Movement of Population Element:*** The population of fish, birds, snakes, mosquitoes, etc. is large and is constantly moving, being born and dying. So instead of attempting to count all elements of such populations, it is desirable to make estimates using techniques such as counting birds at a place picked at random, setting nets at predetermined places, etc.
2. ***Cost and/or Time Required Contacting the Whole Population:*** Time required contacting the whole population. A census involves a complete count of every individual member of the population of interest, such as persons in a state, households in a town, shops in a city, students in a college, and so on. Apart from the cost and the large amount of resources (such as enumerators, clerical assistance, etc.) that are required, the main problem is the time required to process the data. Hence the results are known after a big gap of time.
3. ***Destructive Nature of Certain Tests:*** The census becomes extremely difficult, if not impossible, when the population of interest is either infinite in terms of size (number); constantly changing; in a state of movement; or observation results required destruction. For example, sometimes it is required to test the strength of some manufactured item by applying a stress until the unit breaks. The amount of stress that results in breakage is the value of the observation that is recorded. If this procedure is applied to an entire population, there would be nothing left. This type of testing is called destructive testing and requires that a sample be used in such cases.

6.3 Concepts in Sampling

1. ***Sampling Units and Population:*** a unit may be taken as a well-defined and identifiable element or a group of elements on which observations can be made. The aggregate of these units is termed as population and the population is said to be finite, if the units are countable. The population is sub-divided into suitable small units known as sampling units for the purpose of sampling. Sampling units may consist of one or more elementary units and each elementary unit belongs to one and one sampling unit.
2. ***Sampling Frame:*** a sampling frame is a list of sampling units with identification particulars indicating the location of the sampling units. A sampling frame represents the population under investigation, and it is the base of drawing a sample. As far as possible, it should be up-to-date, i.e., free from omissions and duplications. **Sample:** a fraction of the population is said to constitute a sample. The number of units included in the sample is known as the size of the sample. **Sampling Fraction:** the ratio of the sample size, n , to the population size. N is known as sampling fraction and it is denoted by (n/N) .
3. ***Sampling Procedure/Method:*** this is the method of selecting a sample from a population.

4. **Census:** this denotes all the elements or unit., of a population which are used to explain the features of population. It usually refers to complete enumeration of all persons in the population.
5. **Population Parameter and Sample Estimator:** any function of the values of units in the population, such as population mean or population variance, is termed a population parameter. There can only be one set of values for a population and the population values are treated as constant. However, the function of the values of the units in the sample, such as sample mean and sample variance, is known as a statistic. The value of the mean and variance differ from sample to sample and, therefore, it is a random variable.

6.4 Principles of Sampling

The following are two important principles which determine the possibility of arriving at a valid statistical inference about the features of a population or process:

- i) Principle of statistical regularity
- ii) Principle of inertia of large numbers

Principle of Statistical Regularity

This principle is based on the mathematical theory of probability. According to King, 'The law of statistical regularity lays down that a moderately large number of items chosen at random from a large group are almost sure on the average to possess the characteristic of the large group.' This principle emphasizes on two factors:

- i) **Sample Size should be Large:** As the size of sample increases it becomes more and more representative of parent population and shows its characteristics. However, in actual practice, large samples are more expensive. Thus, a balance has to be maintained between the sample size, degree of accuracy desired and financial resources available.
- ii) **Samples Must be Drawn Randomly:** The random sample is the one in which elements of the population are drawn in a such way that each combination of elements has an equal probability of being selected in the sample. When the term random sample is used without any specification, it usually refers to a simple random sample. The selection of samples based on this principle can reduce the number of efforts required in arriving at a conclusion about the characteristic of a large population. For example, to understand the book buying habit of students in a college, instead of approaching every student, it is easy to talk to a randomly selected group of students to draw the inference about all students in the college.

Principle of Inertia of Large Numbers

This principle is a corollary of the principle of statistical regularity and plays a significant role in the sampling theory. This principle states that, under similar conditions, as the sample size (number of observations in a sample) get large enough, the statistical inference is likely to be more accurate and stable. For example, if a coin is tossed a large number of times, then relative frequency of occurrence of head and tail is expected to be equal.

6.5 Types of Sampling

Sampling methods compared to census provides an attractive means of learning about a population or process in terms of reduced cost, time and greater accuracy. The representation basis and the element selection techniques from the given population, classify several sampling methods into two categories as shown in Table 7.1

Table 7.1 Methods and types of sampling methods:

<i>Element Selection</i>	<i>Representation Basis</i>	
	<i>Probability (Random)</i>	<i>Non-probability (Non-random)</i>
• Unrestricted	Simple random sampling	Convenience sampling
• Restricted	Complex random sampling	Purposive sampling
	• Stratified sampling	• Quota sampling
	• Cluster sampling	• Judgement sampling
	• Systematic sampling	
	• Multi-stage sampling	

Probability Sampling Methods

A probability sample is one in which each element of the population has a known, non-zero chance of being included in the sample.

i) Simple Random Sample:

A process that gives each element in the population an equal chance of being included in the sample is termed as simple random sampling. The elements are selected, using a list of random numbers appended with most textbooks of research and statistics. Before using the table of random numbers, it is first necessary to number all the elements in the population to be studied. Then the table is marked at some point and the cases whose numbers come up as one from this point down the column of numbers are taken into the sample until the desired number of elements is obtained. The selection of any given element places no limits on other element being selected, thus making equally possible the selection of any one of the many possible combinations of elements.

The two popularly used methods in random sampling are

1. **Lottery Method:** It is the easiest way of choosing the sample. Each unit is assigned a particular number and these numbers are then written on a piece of paper and put in a box. Then a neutral person, who is blindfolded, is made to pullout the required number of units for the sample from the box. Here the sample is being chosen by simple chance and there is no favor or partiality involved. It is also important to ensure that the sheets of paper that are used should be of equal size and quality.
2. **Using the Rotating Drum:** This process is similar to the lottery method but with a slight modification. Here the units are itemized into lists and divided into categories from say 0 to 5. Then the same categories 0-5 are printed on pieces of wood or tin etc.(of same size) and placed in a drum. This drum is then rotated and the required number of the pieces is drawn. Now if we draw 5 zeroes 10 fives and 20 twos then we pick 5 units from the zero list 10 units from five list and 20 units from the twos list respectively.
3. **Selection based on a Sequential List:** This process involves maintaining the units in alphabetical, numerical or geographical sequence. In this procedure units are broken up into numerical, alphabetical or geographical Sequence. So, for a numerical selection one can choose units that fall in multiples of 3, for alphabetical selection we can choose all the names that begin with vowels, etc.

For instance, in drawing the random sample of 50 students from a population of 3500 students in a college we make a list of all 3500 students and assign each student an identification number. This gives us a list of 3500 numbers, called frame for experiment. Then we generate by computer or by other means a set of 50 random numbers in the range of values from 1 and 3500. The procedure gives every set of 50 students in the population an equal chance of being included in the sample. Selecting a random sample is analogous to using a gambling device to generate numbers from this list.

This method is suitable for sampling, as many statistical tests assume independence of sample elements. One disadvantage with this method is that all elements of the population have to be available for selection, which many a times is not possible.

ii) Stratified Sampling

This method is useful when the population consists of a number of heterogeneous subpopulations and the elements within a given subpopulation are relatively homogeneous compared to the population as a whole. Thus, population is divided into mutually exclusive groups called strata that are relevant, appropriate and meaningful in the context of the study. A simple random sample, called a sub-sample, is then drawn from each stratum or group, in proportion or a non-proportion to its size. As the name implies, a proportional sampling procedure requires that the number of elements in each stratum be in the same proportion as in the population. In non-proportional procedure, the number of elements in each stratum is disproportionate to the respective numbers in the population. The basis for forming the strata such as location, age, industry type, gross sales, or number of employees, is at the discretion of the investigator. Individual stratum samples are combined into one to obtain an overall sample for analysis.

In Stratified Random Sampling, the target population of N units is first divided into k subpopulations of N_1, N_2, \dots, N_k units. These populations are non-overlapping and together they comprise the whole population. So that $N_1 + N_2 + \dots + N_k = N$

The sub-populations are called strata. The number in each stratum should be known. A sample is drawn from each stratum independently. The sample sizes within 'k' strata are denoted by n_1, n_2, \dots, n_k respectively. If the total sample size 'n' is to be drawn from the target population than $n_1 + n_2 + \dots + n_k = n$

If a simple random sample is drawn in each stratum, the whole procedure is described as stratified random sampling.

Stratified random sampling requires more than making a list of elements (and estimating the number of elements on the list). It also involves ordering that list by sub groups (or strata) and then, to do sampling randomly or systematically within those sub groups. This method of sampling is used for the following reasons.

- It can reduce the errors in the statistical estimates calculated from the sample.
- It allows you to create a sample that is exactly representative of the various sub groups in the population that you find to be of special interest.

For example, the selected village may have households of SC, ST, OBCs, Others, Minority. The village population first may be divided in to smaller sub groups of different sections of population (stratum) and, thus, the village sample may consist of households from each stratum so that sample may contain all the important characteristics of the village population. In the case of SRS, the sample of all strata/ sub groups sometimes may not be included or covered adequately.

- This method helps in conducting and managing a large-scale survey to be conducted in a country like India. The agency conducting the survey may have field offices in different locations; each one can supervise the survey for a part of the population.
- The basic idea is that it sub-divides the heterogeneous population into homogeneous sub-populations. If each stratum is homogenous in itself, a precise estimate of any stratum mean can be obtained from a small sample, thus, saving a lot of time and cost.

There are two types of stratified samples.

A *proportionate stratified sample* selects the number of elements from each stratum so that the stratum sample size (n_1, n_2, \dots, n_k) is proportional to their respective stratum population size (N_1, N_2, \dots, N_k).

Consider the following examples:

- A selected village may have households of SC (10%), ST (5%), OBCs (45%), Others (30%), Minority (10%). A village sample of 100 may constitute the households of various casts in the above proportion/ percentage so that the sample may contain all important characteristics of village population.
- Hospital patients are stratified according to age, dividing the population into those who are aged 50years or above, and, those who are under 50. If there are twice as many people

aged 50 or above admitted to the hospital as those under 50, a proportionate stratified sample will include twice as many people aged 50 or above.

A *disproportionate stratified sample* selects the number of elements from each stratum so that the stratum sample size is not proportional to the stratum population size. The most common reason for selecting this type of sample is when you want to study a relatively rare but important subpopulation, such as younger patients suffering from heart disease. Proportionate stratification may result in too few elements being selected so that little, if any, statistical analysis can be done. Consequently, even if these patients represent only 1% of the population, you might decide to make them 10% of the final sample. However, once we combine values of all strata, the size of the higher selected proportion needs to be readjusted which is called weighted estimate.

iii) Systematic Random Sample

Designing a Systematic Random Sample is sometimes quite difficult and time consuming and therefore, Systematic Random Sample, like Simple Random Sample, also uses a list of all members of the population in its sampling frame. However, instead of using random numbers to select the sample elements, the researcher applies a skip interval to the list to produce a sample of the required size.

$$\text{Skip interval} = \frac{\text{number of elements in the population}}{\text{the required sample size}}$$

$$K = N/n$$

$$K = \text{skip interval}$$

$$N = \text{Universe size}$$

$$n = \text{Sample size}$$

For example, if we have to select a sample of 100 persons from a universe of 1000 population. then the skip is 10. In this case one number between 1 and 10 has to be selected. Suppose 5 is selected, then the first sample would be 5th and the next one 15th, 25th, 35th, 45th and so on. One of the advantages of this method is that it is more convenient than other methods and simple to design. Again, it is used with very large populations.

iv) Cluster Sampling

This method, sometimes known as area sampling method, has been devised to meet the problem of costs or inadequate sampling frames (a complete listing of all elements in the population so that each member can be identified by a distinct number). The entire population to be analyzed is divided into smaller groups or chunks of elements and a sample of the desired number of areas selected by a simple random sampling method. Such groups are termed as clusters. The elements of a cluster are called elementary units. These clusters do not have much heterogeneity among the elements. A household where individuals live together is an example of a cluster.

Briefly, the procedure for selecting a cluster sample is given below.

- The population is divided into N groups, called clusters.
- The researcher randomly selects n clusters to include in the sample.
- The number of observations within each cluster is known: $M = M_1 + M_2 + M_3 + \dots + M_N$
- Each element of the population can be assigned to one, and only one, cluster.

Cluster sampling should be used only when it is economically justified - when reduced costs can be used to overcome losses in precision. This is most likely to occur in the following situations.

- Constructing a complete list of population elements is difficult, costly, or impossible. For example, it may not be possible to list all elementary units of the populations, for example all households in village, block, etc. However, it would be possible to randomly select a subset of villages, blocks (stage 1 of cluster sampling) and, then, inter-view the head of family in a house of the selected cluster (stage 2).

- The population is concentrated in natural clusters (city blocks, schools, hospitals, etc.). For example, to conduct personal interviews of operating room nurses, it might make sense to randomly select a sample of hospitals (stage 1 of cluster sampling) and then interview all of the operating room nurses at that hospital. Using cluster sampling, the interviewer could conduct many interviews in a single day at a single hospital. Simple random sampling, in contrast, might require the interviewer to spend all day travelling to conduct a single interview at a single hospital.

As discussed above, in the cluster sampling method, the primary selecting unit is not a household, rather a natural cluster of households, viz., hamlets in villages, or, created clusters, viz., schools, Malls, etc., may be decided. The first list of clusters may be selected using the SRS or the PPS sampling techniques. Then, from each selected cluster, all units, or, some of the units, may be selected as per the required sample size using Stratified Random Sampling or the Systematic Random Sampling techniques.

This sampling technique is quite popular in evaluation surveys in health - it is also called the 30 Cluster Sampling Techniques. This is also a rapid method of data collection as the researcher can collect more data in less time due to the decrease in transportation time as compared with other sampling techniques.

v) *Multistage Sampling:*

This method of sampling is useful when the population is very widely spread and random sampling is not possible. The researcher might stratify the population in different regions of the country, then stratify by urban and rural and then choose a random sample of communities within these strata. These communities are then divided into city areas as clusters and randomly consider some of these for study. Each element in the selected cluster may be contacted for desired information.

For example, for the purpose of a national pre-election opinion poll, the first stage would be to choose as a sample a specific state (region). The size of the sample that is the number of interviews, from each region would be determined by the relative populations in each region. In the second stage, a limited number of towns/cities in each of the regions would be selected, and then in the third stage, within the selected towns/cities, a sample of respondents could be drawn from the electoral roll of the town/city selected at the second stage.

The essence of this type of sampling is that a subsample is taken from successive groups or strata. The selection of the sampling units at each stage may be achieved with or without stratification. For example, at the second stage when the sample of towns/ cities is being drawn, it is customary to classify all the urban areas in the region in such a way that the elements (towns/cities) of the population in those areas are given equal chances of inclusion.

6.6 Non-Probability Sampling

A non-probability sample is one in which a case in a sample is chosen in such a manner that it gives you information for the sample itself and makes it possible to generalize the findings for the population with certain degree of precision. Such a sample is also called a purposive sample. This kind of sampling is primarily used to collect information on market surveys to know the attitude, opinion, behavior, reactions of individuals. There are many types of non-probability samples, including snowball sampling, convenience, purposive/ judgment, quota sampling, etc.

1. *Convenience Sampling*

In this procedure, units to be included in the sample are selected at the convenience of the investigator rather than by any prespecified or known probabilities of being selected. For example, a student for his project on 'food habits among adults' may use his own friends in the college to constitute a sample simply because they are readily available and will participate for little or no cost. Other examples are, public opinion surveys conducted by any TV channel near the railway station; bus stop, or in a market.

Convenience samples are easy for collecting data on a particular issue. However, it is not possible to evaluate its representativeness of the population and hence precautions should be taken in interpreting the results of convenient samples that are used to make inferences about a population.

2. Purposive Sampling

Instead of obtaining information from those who are most conveniently available, it sometimes becomes necessary to obtain information from specific targets—respondents who will be able to provide the desired information either because they are the only ones who can give the desired information or because they satisfy to some criteria set by researcher.

3. Judgment Sampling

Judgment sampling involves the selection of respondents who are in the best position to provide the desired information. The judgment sampling is used when a limited number of respondents have the information that is needed. In such cases, any type of probability sampling across a cross section of respondents is purposeless and not useful. This sampling method may curtail the generalizability of the findings due to the fact that we are using a sample of respondents who are conveniently available to us. It is the only viable sampling method for obtaining the type of information that is required from very specific section of respondents who possess the knowledge and can give the desired information. However, the validity of the sample results depends on the proper judgment of the investigator in choosing the sample. Great precaution is needed in drawing conclusions based on judgment samples to make inferences about a population.

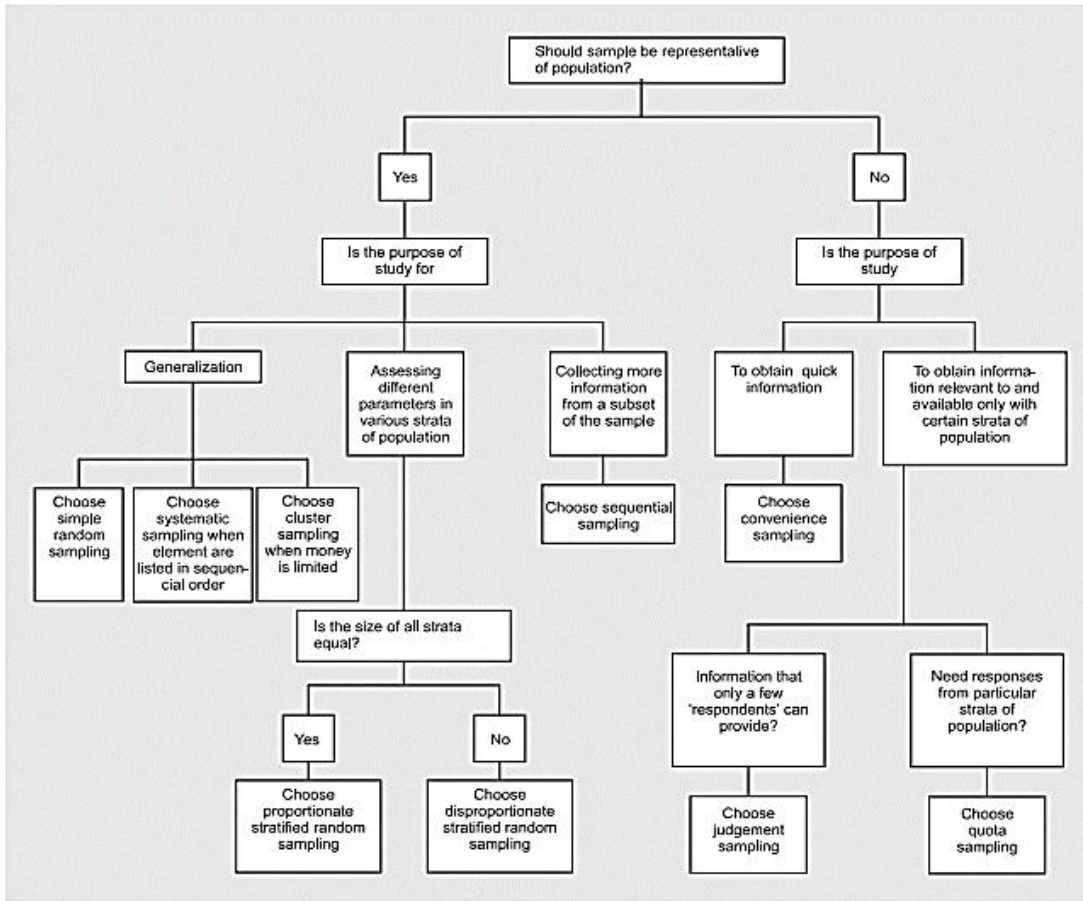
4. Quota Sampling

Quota Sampling is a form of proportionate stratified sampling in which a predetermined proportion of elements are sampled from different groups in the population, but on convenience basis. In other words, in quota sampling the selection of respondents lies with the investigator, although in making such selection he/she must ensure that each respondent satisfies certain criteria which are essential for the study. For example, the investigator may choose to interview ten men and ten women in such a way that two of them have annual income of more than two lakh rupees five of them have annual income between one and two lakh rupees and thirteen whose annual income is below one lakh rupees. Furthermore, some of them should be between 25 and 35 years of age, others between 36 and 45 years of age, and the balance over 45 years. This means that the investigator's choice of respondent is partly dictated by these 'controls. Quota sampling has been criticized because it does not satisfy the fundamental requirement of a sample, that is, it should be random. Consequently, it is not possible to achieve precision of results on any valid basis.

6.7 Choice of Sampling Methods

The choice of particular sampling method (procedure) must be decided according to various factors such as: nature of study, size of the population, size of the sample, availability of resources, degree of precision desired, etc. A choice plan is shown in Fig.7.2

Fig. 7.2: Guidelines to Choose Sample



Judging the Reliability of a Sample

The reliability of a sample can be determined in the following ways to ensure dependable results:

- i. A number of samples may be taken from the same population and the results of various samples compared. If there is not much variation in the results of the different samples, it is a measure of its reliability.
- ii. Sub-sample may be taken from the main sample and studied. If the results of the sub-samples are similar to those given by the main sample it gives a measure of its reliability.
- iii. If some mathematical properties are found in the distribution under study, the sample result can be compared with expected values obtained on the basis of mathematical relationship and if the difference between them is not significant, the sample has given dependable results.

In probability distributions where binomial, normal, Poisson or any other theoretical probability distribution is applicable, sample results can be compared with the expected values to get an idea about the reliability of the sample.

6.8 Errors in Sampling

Many mistakes and errors in social science research happen because of misleading and biased sampling. A sample which does not represent the population is called a biased sample. According to Yule and Kendal, "Bias may be due to imperfect instruments, the personal qualities of the observer, defective techniques and other cases. Like experimental error, it is difficult to eliminate entirely, but usually may be reduced to relatively small dimensions by taking proper care." There are two types of errors such as sampling errors and non-sampling errors. These are discussed below:

1. Sampling Error

Any statistical inference based on sample results (statistics) may not always be correct, because sample results are either based on partial or incomplete analysis of the population features (or characteristics). This error is referred to as the sampling error because each sample taken may produce a different estimate of the population characteristic compared to those results that would have been obtained by a complete enumeration of the population. It is, therefore, necessary to measure these errors so as to have an exact idea about the reliability of sample-based estimates of population features. The likelihood that a sampling error exceeds any specified magnitude must always be specified in terms of a probability value, say 5%. This acceptable margin of error is then used to produce a confidence in the decision maker to arrive at certain conclusions with the limited data at his disposal. In general, in the business context, decision-makers wish to be 95 per cent or more confident that the range of values of sample results reflects the true characteristic of the population or process of interest.

Five Common Types of Sampling Errors

1. Population Specification Error – this error occurs when the researcher does not understand who they should survey. For example, imagine a survey about breakfast cereal consumption. Who to survey? It might be the entire family, the mother, or the children. The mother might make the purchase decision, but the children influence her choice.
2. Sample Frame Error – A frame error occurs when the wrong sub-population is used to select a sample. A classic frame error occurred in the 1936 presidential election between Roosevelt and Landon. The sample frame was from car registrations and telephone directories. In 1936, many Americans did not own cars or telephones, and those who did were largely Republicans. The results wrongly predicted a Republican victory.
3. Selection Error – This occurs when respondents self-select their participation in the study – only those that are interested respond. Selection error can be controlled by going extra lengths to get participation. A typical survey process includes initiating pre-survey contact requesting cooperation, actual surveying, and post-survey follow-up. If a response is not received, a second survey request follows, and perhaps interviews using alternate modes such as telephone or person-to-person.
4. Non-Response – Non-response errors occur when respondents are different than those who do not respond. This may occur because either the potential respondent was not contacted or they refused to respond. The extent of this non-response error can be checked through follow-up surveys using alternate modes.
5. Sampling Errors – These errors occur because of variation in the number or representativeness of the sample that responds. Sampling errors can be controlled by (1) careful sample designs, (2) large samples, and (3) multiple contacts to assure representative response.

2. Non-Sampling Error

Non-sampling error refers to an error that arises from the result of data collection, which causes the data to differ from the true values. It is different from sampling error, which is any difference between the sample values and the universal values that may result from a limited sampling size. Non-sampling errors can come in various forms, including non-response error, measurement error, interviewer error, adjustment error, and processing error.

Types of Non-Sampling Error:

1. *Non-response error*: A non-response error is caused by the differences between the people who choose to participate compared to the people who do not participate in a given survey. In other words, it exists when people are given the option to participate but choose not to; therefore, their survey results are not incorporated into the data.

2. **Measurement error:** A measurement error refers to all errors relating to the measurement of each sampling unit, as opposed to errors relating to how they were selected. The error often arises when there are confusing questions, low-quality data due to sampling fatigue (i.e., someone is tired of taking a survey), and low-quality measurement tools.
3. **Interviewer error:** Interviewer error occurs when the interviewer (or administrator) makes an error when recording a response. In qualitative research, an interviewer may lead a respondent to answer a certain way. In quantitative research, an interviewer may ask the question differently, which leads to a different result.
4. **Adjustment error:** An adjustment error describes a situation where the analysis of the data adjusts it so that it is not entirely accurate. Forms of adjustment error include errors with weighting the data, data cleaning, and imputation.
5. **Processing error:** A processing error arises when there is a problem with processing the data that causes an error of some kind. An example will be if the data were entered incorrectly or if the data file is corrupt.

6.9 Role of Sampling Theory

Everyone who has ever worked on a research project knows that resources are limited; time, money and people never come in an unlimited supply. For that reason, most research projects aim to gather data from a sample of people, rather than from the entire population (the census being one of the few exceptions). This is because sampling allows researchers to:

1. Save Time

Contacting everyone in a population takes time. And, invariably, some people will not respond to the first effort at contacting them, meaning researchers have to invest more time for follow-up. Random sampling is much faster than surveying everyone in a population, and obtaining a non-random sample is almost always faster than random sampling. Thus, sampling saves researchers lots of time.

2. Save Money

The number of people a researcher contacts is directly related to the cost of a study. Sampling saves money by allowing researchers to gather the same answers from a sample that they would receive from the population.

Non-random sampling is significantly cheaper than random sampling, because it lowers the cost associated with finding people and collecting data from them. Because all research is conducted on a budget, saving money is important.

3. Collect Richer Data

Sometimes, the goal of research is to collect a little bit of data from a lot of people (e.g., an opinion poll). At other times, the goal is to collect a lot of information from just a few people (e.g., a user study or ethnographic interview). Either way, sampling allows researchers to ask participants more questions and to gather richer data than does contacting everyone in a population.

Summary

The concept of sampling may be defined as a process which allows us to study a small group of people from the large group to derive inferences that are likely to be applicable to all the people of the large group. The rationale for selecting a small group of people (sample) for study is if we can get almost same results by studying a carefully selected small group of people why we should study the large group at all.

A single unit of study is referred to as an element of population. The aggregate of all the elements that conform to some defined set of definitions is called population.

A representative sampling procedure ensures that the sample statistics will be correct within certain limits. In other words, a representative sampling plan ensures that the selected sample is sufficiently representative of the population to justify our running the risk of taking it as representative.

There are two methods of sampling, namely, probability and non-probability. The essential characteristic of probability sampling is that one can specify for each element of the population the chance of being included in the sample. Major types of probability sampling are: simple random sampling, stratified random sampling and cluster sampling.

In non-probability sampling, there is no way of estimating the probability that each element has of being included in the sample and no assurance that every element has some chance of being included. The four important types of non-probability sampling are accidental sampling, quota sampling, snowball sampling and purposive sampling.

If sampling is carried out in a series of stages, it is possible to combine probability and non-probability sampling in one design. That is, one or more of the stages can be carried out according to probability sampling principles and the balance by non-probability method of sampling. The investigators may select clusters by probability cluster sampling techniques, but, at the final stage, select the elements as a quota sample.

To determine the size of a sample is really a difficult task, particularly in social research where we consider a combination of variables. However, it is possible to determine the size of sample by using a representative sampling procedure.

Keywords

1. Sample: A sample is simply a subset of a larger aggregation, i.e., typically a population and it contains all the characteristics of a population,
2. Sampling: The process of selection of subjects/study elements to create a sample for collecting information about a population.
3. Sampling Error: While collecting information from a sample, there is Sampling a chance that the sampling statistics may not be equal to the same values in the population. The error is that the sample does not contain complete information about the population.
4. Probability Sampling: probability of an element to be included in a sample.
5. Non-Probability Sampling: no assurance that every element has some chance of being included in the sample.
6. Cluster Sampling: the whole research area divided into such area is known as clusters.
7. Non- Probability Sampling: It is the sampling technique that does not depend on randomization. It banks upon the ability of the researcher choose the elements of a sample.
8. Probability Sampling: It is a sampling technique in which each element of the population has an equal probability of selection and this is because of randomization and hence it is also known as random sampling.

Self Assessment

1. Interviewing all members of a given population is called:
 - A. A sample.
 - B. A Gallup poll.
 - C. A census.
 - D. A Nielsen audit.
2. A sample consists of
 - A. All units of the population

-
- B. 5% units of the population
 - C. 10% units of the population
 - D. Any fraction of the population
3. Sampling is used in the situations
- A. Blood test of the patients
 - B. Cooking rice in an utensil
 - C. Purchase of food commodity from shopkeeper
 - D. All the above
4. The number of possible samples of size n out of N population size in SRSWR is equal to
- A. N_{cn}
 - B. N^n
 - C. $(N-n)/N$
 - D. n/N
5. A function of sample observations is known as
- A. Statistic
 - B. Estimator
 - C. Both (a)&(b)
 - D. None
6. Of the following sampling methods, which is a probability method?
- A. Judgment
 - B. Quota
 - C. Simple random
 - D. Convenience
7. Which of the following is not a type of non-probability sampling?
- A. Quota sampling
 - B. Convenience sampling
 - C. Snowball sampling
 - D. Stratified random sampling
8. Which of the following is the non-random method of selecting samples from a population?
- A. Multistage sampling
 - B. Cluster sampling
 - C. Quota sampling
 - D. All of the above
9. People who are available, volunteer or can be easily recruited are used in the sampling method called.....
- A. Connivance Sampling

- B. Simple random Sampling
 - C. Cluster Sampling
 - D. Systematic Sampling
10. In which of the following non random sampling techniques does the researcher ask the research participants to identify other potential research participants?
- A. Quota
 - B. Purposive
 - C. Convenience
 - D. Snowball
11. Increasing the sample size has the following effect upon the sampling error?
- A. It increases the sampling error
 - B. It reduces the sampling error
 - C. It has no effect on the sampling error
 - D. All of the above
12. The difference between a statistic and the parameter is called:
- A. Non-random
 - B. Probability
 - C. Sampling error
 - D. Random
13. The error in a survey other than sampling error is known as
- A. Sampling error
 - B. Non-sampling error
 - C. Formula error
 - D. None
14. If the sample sizes are large from the population, then which error will contribute less errors
- A. Sampling error
 - B. Non-sampling error
 - C. Both (a)&(b)
 - D. None
15. A sampling frame is:
- A. A summary of the various stages involved in designing a survey
 - B. An outline view of all the main clusters of units in a sample
 - C. A list of all the units in the population from which a sample will be selected
 - D. A wooden frame used to display tables of random numbers

Answers for SelfAssessment

1. A 2. D 3. A 4. B 5. C
6. C 7. D 8. C 9. C 10. D
11. B 12. C 13. B 14. A 15. C

Review Questions

1. Briefly explain (a) The fundamental reason for sampling (b) Some of the reasons why a sample is chosen instead of testing the entire population
2. What is sampling? Explain the importance in solving business problems. Critically examine the well-known methods of probability sampling and non-probability sampling.
3. List some of the situations where (a) sampling is more appropriate than census and (b) census is more appropriate than sampling.
4. Discuss the method of cluster sampling. What is the difference between cluster sampling and stratified random sampling?
5. Discuss the sources of sampling and non-sampling errors.
6. What are the main steps involved in a sample survey? Discuss different sources of error in such surveys and point out how these errors can be controlled.
7. If only one sample is selected in a sampling problem, how is it possible to have an entire distribution of the sample mean?
8. Explain the concept of standard error. Discuss the role of standard error in large sample theory.

**Further Readings**

- Gupta, C.B., & Vijay Gupta, An Introduction to Statistical Methods, Vikas Publishing House Pvt. Ltd., New Delhi.
- Kothari, C.R.(2004) Research Methodology Methods and Techniques, New Age International (P) Ltd., New Delhi.
- Levin, R.I. and D.S. Rubin. (1999) Statistics for Management, Prentice-Hall of India, New Delhi
- Mustafi, C.K.(1981) Statistical Methods in Managerial Decisions, Macmillan, New Delhi

Unit 07: Design of Experiments

CONTENTS

Objectives

Introduction

7.1 Design of Experiments

7.2 Informal Experiment Designs

7.3 Formal Experimental Designs

Summary

Keywords

Self Assessment

Answers for Self Assessment

Review Questions

Further Reading

Objectives

- enlist different design of experiments,
- elaborate different design of experiments,
- select relevant design of experiment for a particular problem,
- apply design of experiments appropriately.

Introduction

The seventh unit endeavors to make detailed discussion on design of experiments. In this unit we must understand the concept and different types/categories of design of experiment or experimental design. A proper understanding of the various design of experiment is essential for a researcher for selected a correct design of experiment for his/her research. It can be done through the detail description of the various design of experiment. Let's start with the design of experiment or experimental design.

7.1 Design of Experiments

Experimental design refers to the framework or structure of an experiment. Experimental designs can be classified into two broad categories - formal and informal experimental designs.

Formal experimental designs are divided into four types which are - Completely randomized design (C.R. Design), Factorial designs, Latin square design (L.S. Design), and Randomized block design (R.B. Design).

Informal experimental designs are divided into three types which are - Before-and-after without control design, After-only with control design, and Before-and-after with control design.

The diagrammatical representation of design of experiments or experimental design is as follows:

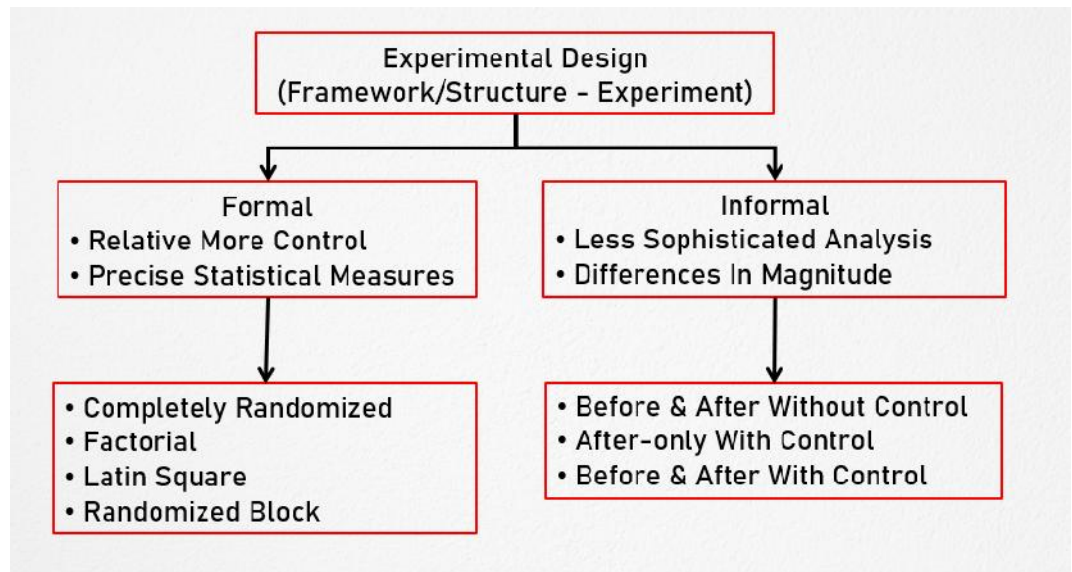


Figure 1.1

7.2 Informal Experiment Designs

Informal experimental designs normally use a less sophisticated form of analysis based on differences in magnitudes. Informal experimental designs are further divided into three types which are given as below:

- (i) Before-and-after without control design.
- (ii) After-only with control design.
- (iii) Before-and-after with control design.

Let us first discuss all types of informal experimental designs one by one as follows:

Before-and-after without control design

In this design, a single test group or area is selected, and dependent variable is measured before the introduction of the treatment. The treatment is then introduced, and the dependent variable is measured again after the treatment has been introduced. The effect of the treatment would be equal to the level of the phenomenon after the treatment minus the level of the phenomenon before the treatment. The main difficulty of the design is that with the passage of time considerable extraneous variations may be there in its treatment effect. The diagrammatical representation of before-and-after without control design is as follows:

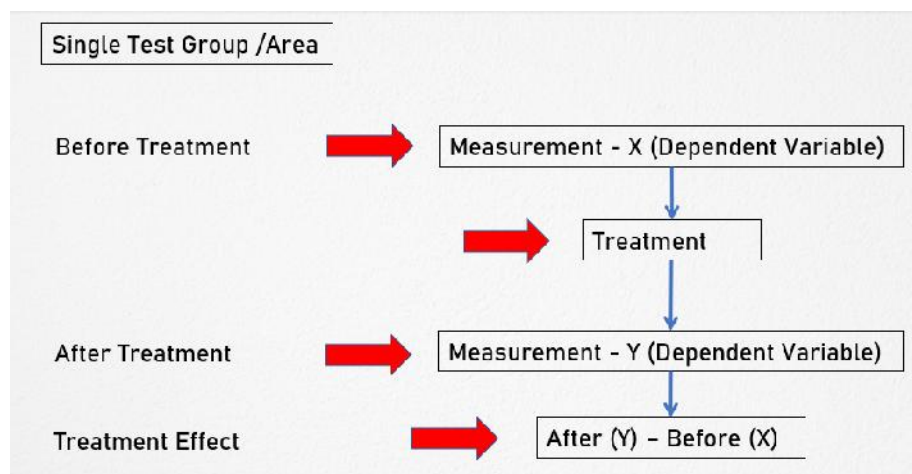


Figure 1.2

The main limitation of this design is that there may be an effect of considerable extraneous variations on its treatment effect with the passage of time

After-only with control design

Two groups (control and test group) are selected, and the treatment is introduced into the test group only. The dependent variable is then measured in both the groups at the same time. Treatment impact is assessed by subtracting the value of the dependent variable in the control group (before) from its value in the test group (after). The diagrammatical representation of after-only with control design is as follows:

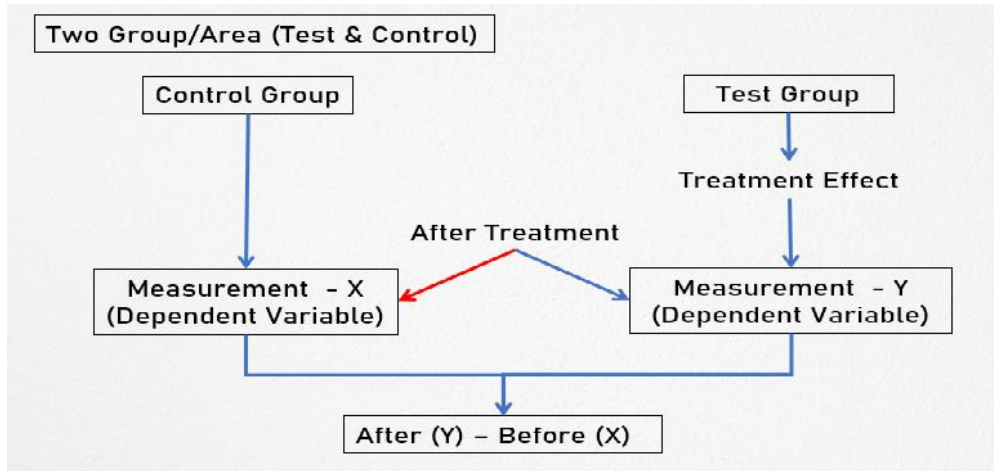


Figure 1.3

The basic assumption in this design is that the two areas are identical with respect to their behaviour towards the phenomenon considered. If this assumption is not true, there is the possibility of extraneous variation entering into the treatment effect. However, data can be collected in such a design without the introduction of problems with the passage of time. In this respect the design is superior to before-and-after without control design.

Before-and-after with control design

In this design, two groups are selected, and the dependent variable is measured in both the groups for an identical time-period before the treatment. The treatment is then introduced into the test group only, and the dependent variable is measured in both for an identical time-period after the introduction of the treatment. The treatment effect is determined by subtracting the change in the dependent variable in the control group from the change in the dependent variable in test group.

This design is superior because it avoids extraneous variation resulting both from the passage of time and from non-comparability of the test and control groups. But due to lack of historical data, time or a comparable control group, a researcher should prefer to select one out of Before-and-after without control design and After-only with control design. The diagrammatical representation of before-and-after with control design is as follows:

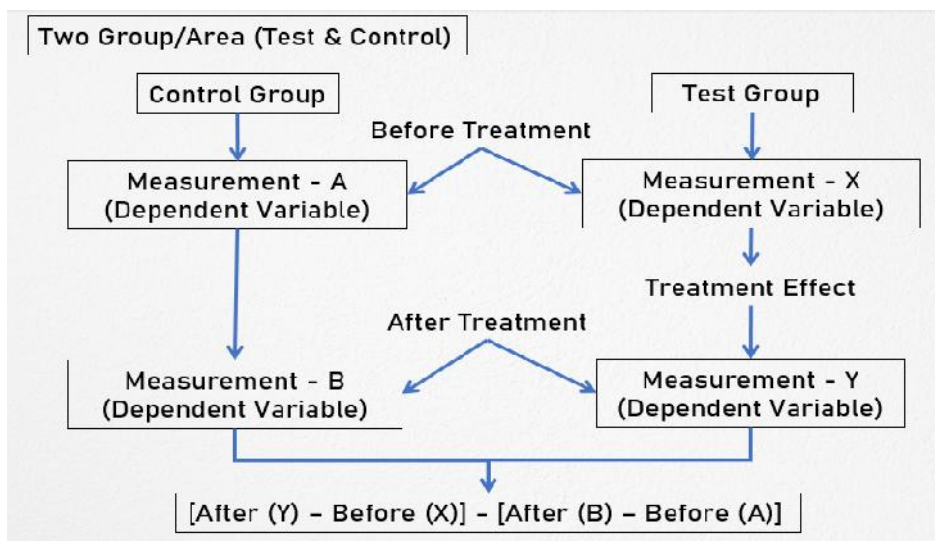


Figure 1.4

Let us now discuss all types of formal experimental designs one by one as follows:

7.3 Formal Experimental Designs

Formal experimental designs offer relatively more control and use precise statistical procedures for analysis. Formal experimental designs are further divided into four types which are given as below:

- (i) Completely randomized design (C.R. Design)
- (ii) Factorial designs
- (iii) Latin square design (L.S. Design)
- (iv) Randomized block design (R.B. Design)

Completely randomized design(C.R. design)

This design involves only two principles viz., the principle of replication and the principle of randomization of experimental designs. It is the simplest possible design, and its procedure of analysis is also easier. The essential characteristic of the design is that subjects are randomly assigned to experimental treatments (or vice-versa).

For example, if we have twenty subjects and if we wish to test ten under treatment X and ten under treatment Y, randomization process gives every possible group of ten subjects selected from a set of twenty an equal opportunity of being assigned to treatment X and treatment Y.

One-way analysis of variance (one-way ANOVA) is used to analyze such a design. Even unequal replications can also work in this design. It provides maximum number of degrees of freedom to the error.

This design is generally used when experimental areas happen to be homogeneous. Technically, when all the variations due to uncontrolled extraneous factors are included under the heading of chance variation, we refer to the design of experiment as C.R. design.

A brief description of the two forms of C.R. design is given below:

Two-group simple randomized design

In a two-group simple randomized design, first, the population is defined and then from the population a sample is selected randomly. Items, after being selected randomly from the population, be randomly assigned to the experimental and control groups. This random assignment of items to two groups is technically described as principle of randomization. Thus, this design yields two groups as representatives of the population.

Since in the sample randomized design the elements constituting the sample are randomly drawn from the same population and randomly assigned to the experimental and control groups, it becomes possible to draw conclusions based on samples applicable for the population.

The experimental and control groups are given different treatments of the independent variable. The treatment effect is determined by subtracting the change in the dependent variable in the control group from the change in the dependent variable in experiment group.

The diagrammatical representation of two-group simple randomized design is as follows:

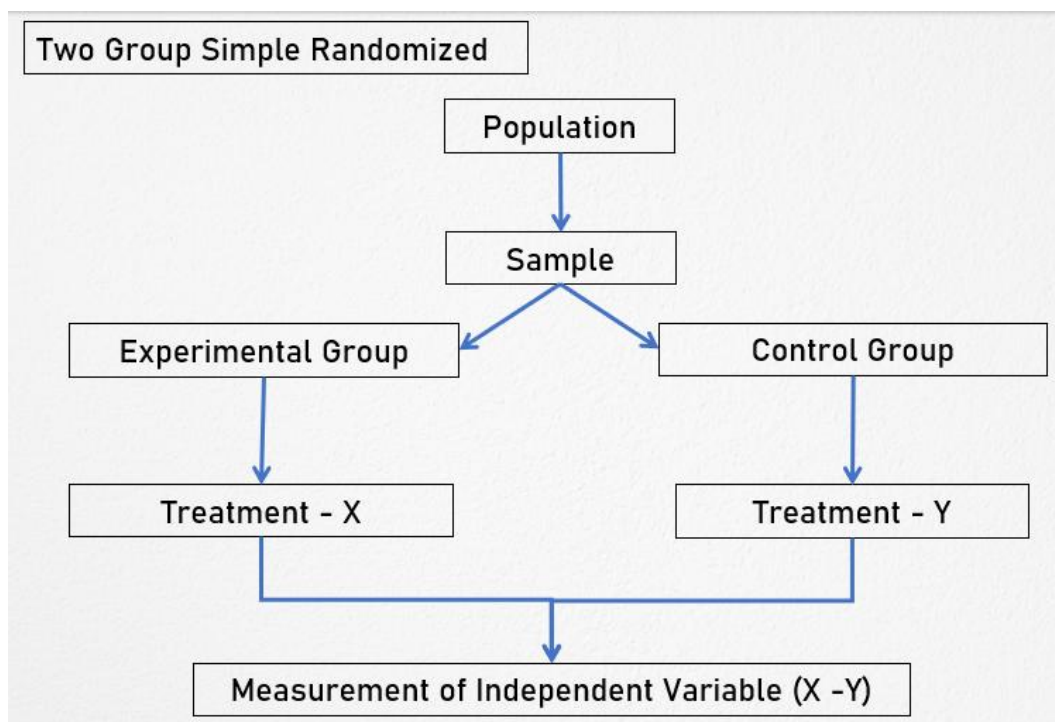


Figure 1.5

The strength of the two-group simple randomized design is that it is simple and randomizes the differences among the sample items.

The limitation of the two-group simple randomized design is that the individual differences among those conducting the treatments are not eliminated, i.e., the two-group simple randomized design does not control the extraneous variable and as such the result of the experiment may not depict a correct picture.

For example, suppose the researcher wants to compare two groups of students who have been randomly selected and randomly assigned. Two different treatments viz., the usual training and the specialized training are being given to the two groups. The researcher hypothesizes greater gains for the group receiving specialized training. To determine this, he tests each group before and after the training, and then compares the amount of gain for the two groups to accept or reject his hypothesis. But this does not control the differential effects of the extraneous independent variables which were individual differences among those conducting the training programme.

Random replications design

The limitation of the two-group randomized design is usually eliminated within the random replications design. The teacher differences on the dependent variable were ignored, i.e., the extraneous variable was not controlled. But in a random replications design, the effect of such differences is minimized/reduced by providing several repetitions for each treatment.

Each repetition is technically called a 'replication'. Random replication design serves two purposes viz., it provides controls for the differential effects of the extraneous independent variables and secondly, it randomizes any individual differences among those conducting the treatments.

The diagrammatical representation of two-group simple randomized design is as follows:

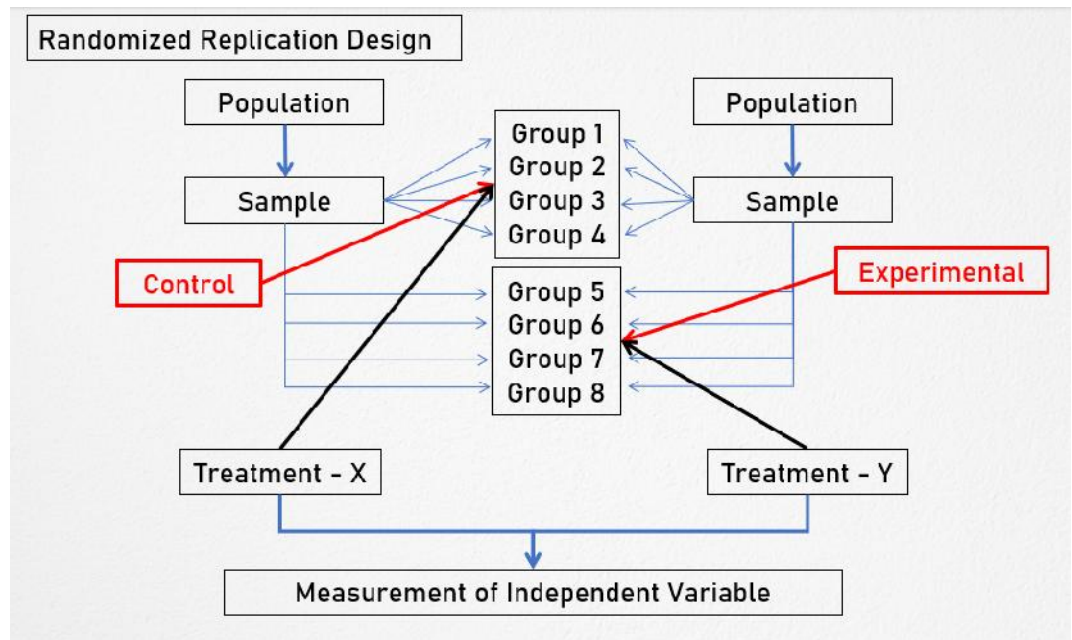


Figure 1.6

There are two populations in the replication design. The sample is taken randomly from the population available for study and is randomly assigned, suppose, four experimental and four control groups. Similarly, sample is taken randomly from population available to conduct experiments (because of the eight groups eight such individuals be selected) and the eight individuals so selected should be randomly assigned to the eight groups.

Generally, equal number of items are put in each group so that the size of the group is not likely to affect the result of the study. Variables relating to both population characteristics are assumed to be randomly distributed among the two groups. Thus, this random replication design is, in fact, an extension of the two-group simple randomized design.

Factorial designs

Factorial designs are used in experiments where the effects of varying more than one factor are to be determined. They are important in several economic and social phenomena where usually many factors affect a particular problem.

Factorial designs can be of two types:

- (i) Simple factorial designs
- (ii) Complex factorial designs

Now, Simple and Complex factorial designs are discussed as follows:

Simple factorial designs

In simple factorial designs, we consider the effects of varying two factors on the dependent variable. Simple factorial design is also termed as a 'two-factor-factorial design', Simple factorial design may either be a 2×2 , 3×4 or 5×3 simple factorial design

In this design the extraneous variable to be controlled by homogeneity is called the control variable and the independent variable, which is manipulated, is called the experimental variable.

The diagrammatical representation of simple factorial designs is as follows:

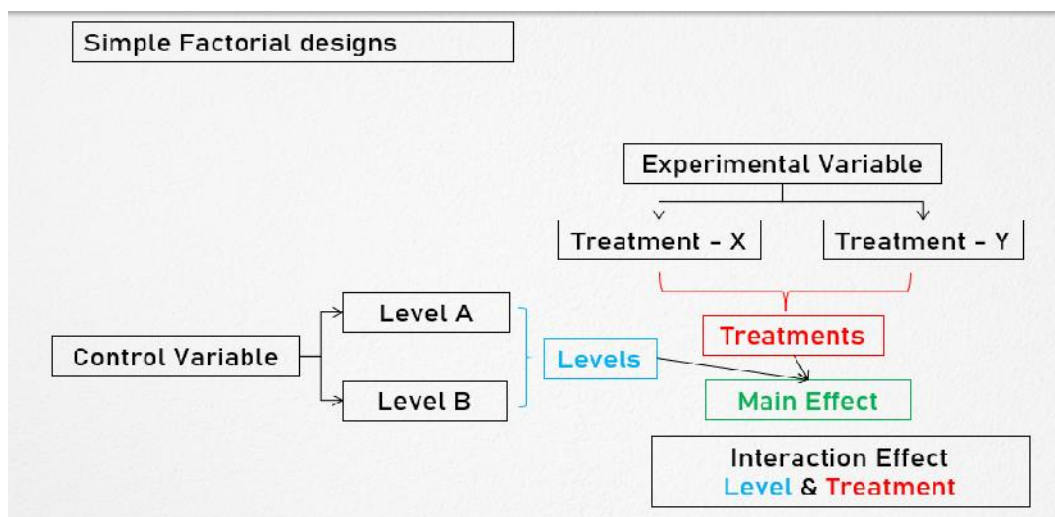


Figure 1.7

Then there are two treatments of the experimental variable and two levels of the control variable. As such there are four cells into which the sample is divided. Each of the four combinations would provide one treatment or experimental condition.

Subjects are assigned at random to each treatment in the same manner as in a randomized group design. The means for different cells may be obtained along with the means for different rows and columns.

Means of different cells represent the mean scores for the dependent variable and the column means in the given design are termed the main effect for treatments without considering any differential effect that is due to the level of the control variable. Similarly, the row means in the said design are termed the main effects for levels without regard to treatment. Thus, through this design we can study the main effects of treatments as well as the main effects of levels. An additional merit of this design is that one can examine the interaction between treatments and levels, through which one may say whether the treatment and levels are independent of each other, or they are not so.

Complex factorial designs

Experiments with more than two factors at a time involve the use of complex factorial designs. A design which considers three or more independent variables simultaneously is called a complex factorial design. Complex factorial design is known as 'multifactor-factorial design.'

The diagrammatical representation of complex factorial designs is as follows:

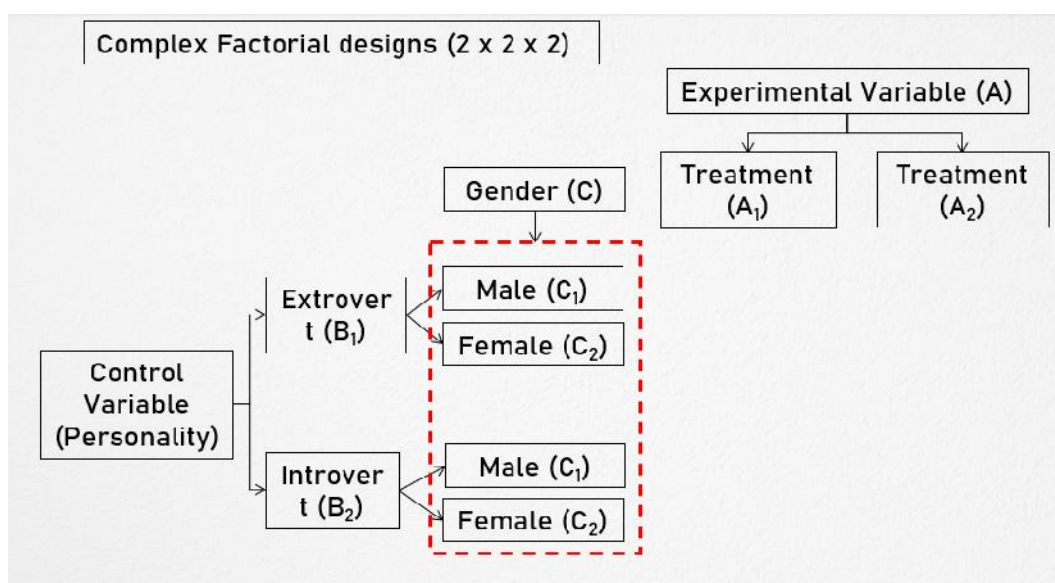


Figure 1.8

In case of three factors with one experimental variable having two treatments and two control variables, each one of which having two levels, the design used will be termed $2 \times 2 \times 2$ complex factorial design which will contain a total of eight cells.

Latin square design (L.S. design)

It is an experimental design very frequently used in agricultural research. The conditions under which agricultural investigations are carried out are different from those in other studies for nature plays an important role in agriculture.

For instance, an experiment must be made through which the effects of five different varieties of fertilizers on the yield of a certain crop, say wheat, it to be judged. In such a case the varying fertility of the soil in different blocks in which the experiment must be performed must be taken into consideration; otherwise, the results obtained may not be very dependable because the output happens to be the effect not only of fertilizers, but it may also be the effect of fertility of soil.

Similarly, there may be impact of varying seeds on the yield. To overcome such difficulties, the L.S. design is used when there are two major extraneous factors such as the varying soil fertility and varying seeds.

The diagrammatical representation of Latin Square design is as follows:

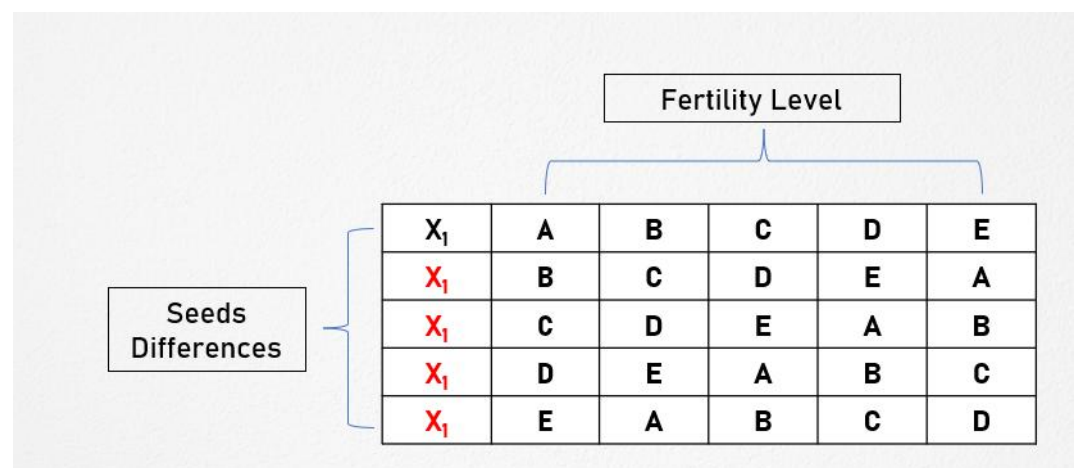


Figure 1.9

The Latin-square design is one wherein each fertilizer appears five times but is used only once in each row and in each column of the design. The treatments in a L.S. design is so allocated among the plots that no treatment occurs more than once in any one row or any one column.

The two blocking factors may be represented through rows and columns (one through rows and the other through columns).

In a L.S. design the field is divided into as many blocks as there are varieties of fertilizers and then each block is again divided into as many parts as there are varieties of fertilizers in such a way that each of the fertilizer variety is used in each of the block (whether column-wise or row-wise) only once. The analysis of the L.S. design is very similar to the two-way ANOVA technique.

The merit of this design is that it enables differences in fertility gradients in the field to be eliminated in comparison to the effects of different varieties of fertilizers on the yield of the crop.

The limitation of this design is that although each row and each column represents equally all fertilizer varieties, there may be considerable difference in the row and column means both up and across the field. This means that in L.S. design we must assume that there is no interaction between treatments and blocking factors. This defect can, however, be removed by taking the means of rows and columns equal to the field mean by adjusting the results.

It requires number of rows, columns, and treatments to be equal. This reduces the utility of this design. In case of (2×2) L.S. design, there are no degrees of freedom available for the mean square error and hence the design cannot be used.

If treatments are 10 or more, then each row and each column will be larger in size so that rows and columns may not be homogeneous. This may make the application of the principle of local control ineffective. Therefore, L.S. design of orders (5×5) to (9×9) are generally used.

Randomized block design

It is an improvement over Completely randomized design. Here, Principle of local control can be applied along with the other two principles of experimental designs. Subjects are first divided into groups, known as blocks, such that within each group the subjects are relatively homogeneous in respect to some selected variable. The variable selected (for grouping the subjects) will be related to the measures to be obtained (in respect of the dependent variable).

The number of subjects in each given block = number of treatments.

One subject in each block would be randomly assigned to each treatment. Blocks are the levels at which we hold the extraneous factor fixed to measure extraneous factor contribution to the total variability of data.

In Randomized block design each treatment appears the same number of times in each block. It is analyzed by the two-way analysis of variance (two-way ANOVA) technique.

For example, suppose four different forms of a standardized test in statistics were given to each of five students (selected one from each of the five I.Q. blocks) and following are the scores which they obtained. If each student separately randomized the order in which he or she took the four tests (by using random numbers), we refer to the design of this experiment as a R.B. design. The purpose of this randomization is to take care of such possible extraneous factors (say as fatigue) or perhaps the experience gained from repeatedly taking the test.

The diagrammatical representation of randomized block design is as follows:

		Level of IQ				
		Very Low	Low	Average	High	Very High
Student	→	A	B	C	D	E
Form	1	B	C	D	E	A
	2	C	D	E	A	B
	3	D	E	A	B	C
	4	E	A	B	C	D

Figure 1.9

Summary

Experimental design refers to the framework or structure of an experiment. Experimental designs can be classified into two broad categories - formal and informal experimental designs.

Formal experimental designs are divided into four types which are - Completely randomized design (C.R. Design), Factorial designs, Latin square design (L.S. Design), and Randomized block design (R.B. Design).

Informal experimental designs are divided into three types which are - Before-and-after without control design, After-only with control design, and Before-and-after with control design.

The summary of design of experiments is given with the help of flow chart as follows:

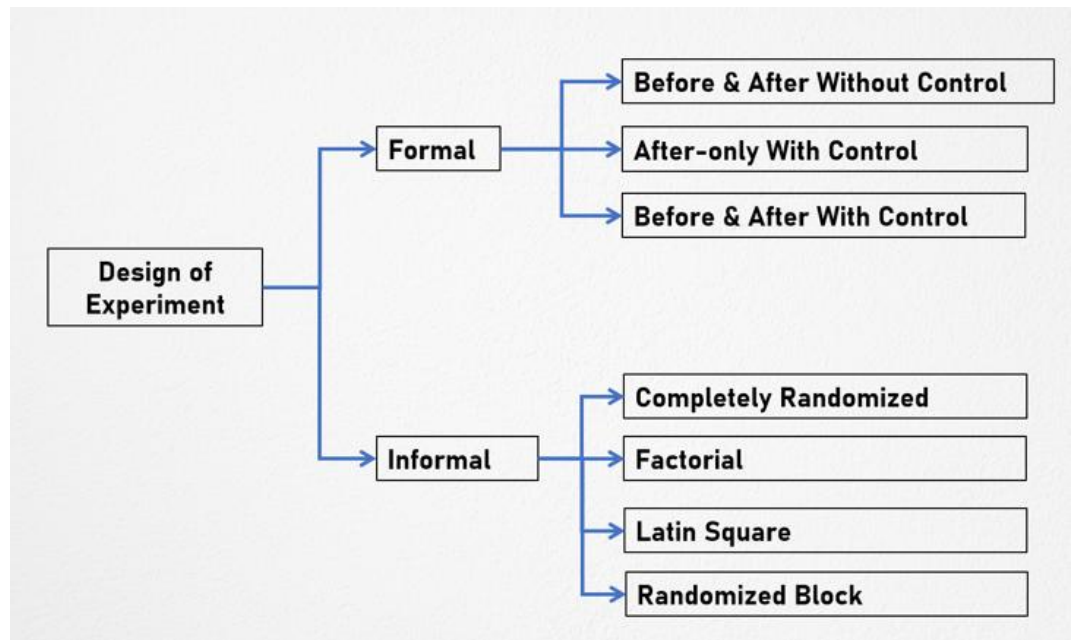


Figure 1.10

Keywords

Experimental design refers to the framework or structure of an experiment.

Informal experimental designs normally use a less sophisticated form of analysis based on differences in magnitudes.

Formal experimental designs offer relatively more control and use precise statistical procedures for analysis.

Self Assessment

1. Which of the following is the informal experimental design?
 - A. Completely randomized design
 - B. Factorial designs
 - C. Before-and-after without control design
 - D. Latin square design

2. Which of the following is the formal experimental design?
 - A. Randomized block design
 - B. Before with control design
 - C. After-only with control design
 - D. Before-and-after with control design

3. Which of the following is a superior design?
 - A. After-only with control design
 - B. Before-and-after without control design
 - C. Before-and-after with control design
 - D. After-only without control design

4. _____ involves the principle of replication and the principle of randomization.
 - A. Before-and-after with control design

- B. Completely randomized design
 - C. Factorial designs
 - D. Randomized block design
5. Random replications design is a type of
- A. Factorial designs
 - B. Randomized block design
 - C. Latin square design
 - D. Completely randomized design
6. One-way analysis of variance is used to analyse_____.
- A. Before-and-after with control design
 - B. Completely randomized design
 - C. Factorial designs
 - D. Randomized block design
7. Items after being selected randomly from the population, be randomly assigned to the experimental and control groups. This random assignment of items to two groups is technically described as
- A. Principle of randomization
 - B. Principle of local control
 - C. Principle of replication
 - D. Principle of replication and order
8. _____ are used in experiments where the effects of varying more than one factor are to be determined.
- A. Before-and-after with control design
 - B. Completely randomized design
 - C. Factorial designs
 - D. Randomized block design
9. The effects of varying two factors on the dependent variable is considered in
- A. Complex factorial design
 - B. Three factorial designs
 - C. Multiple factorial designs
 - D. Simple factorial designs
10. The extraneous variable which is supposed to be controlled by homogeneity in simple factorial design is called the
- A. Experimental Variable
 - B. Control Variable
 - C. Simple Variable
 - D. Complex Variable

11. The independent variable, which is manipulated, is called the
 - A. Experimental Variable
 - B. Control Variable
 - C. Simple Variable
 - D. Complex Variable

12. Experiments with more than two factors at a time involve the use of
 - A. Three factorial designs
 - B. Multiple factorial designs
 - C. Complex factorial design
 - D. Simple factorial designs

13. Complex factorial design is known as
 - A. Two-factor-factorial design
 - B. Three-factorial design
 - C. Four-factorial design
 - D. Multifactor-factorial design

14. Which of the following design is very frequently used in agricultural research?
 - A. Factorial designs
 - B. Randomized block design
 - C. Latin square design
 - D. Completely randomized design

15. In _____ each treatment appears the same number of times in each block.
 - A. Before-and-after with control design
 - B. Randomized block design
 - C. Completely randomized block design
 - D. Factorial block designs

Answers for Self Assessment

- | | | | | |
|-------|-------|-------|-------|-------|
| 1. C | 2. A | 3. C | 4. B | 5. D |
| 6. B | 7. A | 8. C | 9. D | 10. B |
| 11. A | 12. C | 13. D | 14. C | 15. B |

Review Questions

1. How formal experimental designs are different from informal experimental designs.
2. Discuss Completely Randomized design. Explain random replications design with the help of an example.

3. Explain two-group simple randomized design with the help of an example.
4. Illustrate Latin Square design with the help of an appropriate example.
5. Describe Simple Factorial design in detail with the help of suitable example.
6. Explain Complex Factorial design with the help of an example
7. What is Randomized block design. Discuss in detail with the help of an illustration.
8. Illustrate before-and-after without control design with the help of an appropriate example.
9. Explain after-only with control design,
10. Discuss before-and-after with control design in detail with the help of an example.



Further Reading

- Research Methodology- Methods and Techniques by Gaurav Garg and C. R. Kothari, New Age International (P) Limited.
- Methodology of Educational Research by Lokesh Koul, Vikas Publishing House Pvt. Ltd.
- Tests, Measurements and Research Methods in Behavioural Sciences by A.K. Singh, Bharti Bhawan Publishers and Distributors.
- Essentials of Scientific Behavioural Research by R.A. Sharma, R. Lall Book Depot.



Web Links

- <http://home.iitk.ac.in/~shalab/anova/chapter4-anova-experimental-design-analysis.pdf>
- http://www.ru.ac.bd/stat/wp-content/uploads/sites/25/2019/03/502_06_Montgomery-Design-and-analysis-of-experiments-2012.pdf

Unit 08: Probability

CONTENTS

Objectives

Introduction

8.1 History of Probability

8.2 Probability

8.3 Definition Of Probability

8.4 Basic Concepts of Probability

8.5 Probability Rules

8.6 Rules of Addition

8.7 Rules of Multiplication

Summary

Keywords

Self Assessment

Answers for Self Assessment

Review Questions

Further Readings

Objectives

- Help yourself to understand the amount of uncertainty that is involved before making important decisions.
- Understand fundamentals of probability and various probability rules that help to you measure uncertainty involving uncertainty.

Introduction

Probability theory plays a very important role in many areas of physical, social, biological, engineering and management sciences. It lays the foundations for a systematic study of mathematical statistics. Games of chance, number of accidents, birth and death rates, system reliability and expected gain in a business venture are some examples where the probability concepts are used.

In probability theory, we are usually interested in the occurrence or non-occurrence of some events. There are several ways of defining the probability of an event. Various definitions of probability are, in general, consistent with one another. In this unit, you will learn about these definitions. You will also learn about the ways, of calculating the probability of events in simple cases and methods of building some simple probability models.

8.1 History of Probability

The history of probability theory dates back to the 17th century. During that period, the classical theory of probability was propounded by several distinguished scientists. Among others, Pascal, Fermat, Huygens and Jakob Bernoulli applied it to games of chance and obtained numerical values of probability of various events by using the classical theory.

The relative frequency definition of probability of events which show statistical regularity in repeated experiments is due to Richard von Mises. He developed this theory around the year 1921.

At that time, the classical theory was prevalent and the relative frequency approach provided a new dimension. This definition is quite popular among engineers and experimental scientists as it gives a physical interpretation to the occurrence of an event.

Both classical and relative frequency definitions give a method of assigning a numerical value to the probability of an event. This can also be done simply by interpreting probability as the degree of belief according to the subjective opinion of a person. For example, the event that a particular team will win a tournament may be assigned a probability 0.65 by an expert of the game. Such probabilities when assigned are called subjective probabilities.

The modern theory of probability owes a great deal to the work of several Russian mathematicians, notably A. Kolmogorov, who gave the axiomatic definition of probability in the year 1933. According to this, the probability of an event satisfies three axioms without any concern for its physical interpretation. As you will find out shortly it provides a solid foundation for a deeper study of probability and statistics.

8.2 Probability

Initial studies in probability theory originated from calculation of gambling odds. Slowly over a period of years, it found its applications in other areas where the outcome of an individual random experiment cannot be predicted with certainty. In many cases some events occur more often than others and it is desirable to attach a quantitative measure to various events of a random experiment. For example, if you toss five coins, you may like to assign a quantitative measure to the following two events:

1. The total number of heads observed is more than the total number of tails, and
2. Either two or three heads are observed.

Similarly, before the birth of a baby, a doctor may like to assign a quantitative measure to the event that the weight of baby is less than 1.5 kgms. Probability is such a quantitative measure. It is measured in a unit of "unity", although sometimes in everyday life we also express it as a percentage after multiplying it by 100. Thus, you may say that the chances that a team will win a particular tournament are 35%. The probability "p" of an event also reflects the degree of belief, which you have in the occurrence of that event. A high value of "p" indicates that you are almost certain that the event will occur, whereas a low value of "p" indicates that the event is almost impossible. Using appropriate analysis, if you find that the probability that a given dam will develop a major structural defect in the next 50 years is 0.001, then you are almost certain that this event will not occur and that the design is a safe one. You have thus seen that the probability of an event is a quantitative measure showing the degree of belief which one has in the occurrence or non-occurrence of the event under consideration.

8.3 Definition Of Probability

A general definition of probability states that probability is a numerical measure (between 0 and 1 inclusively) of the likelihood or chance of occurrence of an uncertain event. However, it does not tell us how to compute the probability. In this section, we shall discuss different conceptual approaches of calculating the probability of an event.

1. Classical Approach

This approach of defining the probability is based on the assumption that all the possible outcomes (finite in number) of an experiment are mutually exclusive and equally likely. It states that, during a random experiment, if there is 'a' possible outcome where the favorable event A occurs and 'b' possible outcomes where the event A does not occur, and all these possible outcomes are mutually exclusive, exhaustive, and equiprobable, then the probability that event A will occur is defined as

$$P(A) = \frac{a}{a+b} = \frac{\text{Number of Favourable Outcomes } c(A)}{\text{Total number of outcomes } c(S)}$$

For example, if a fair die is rolled, then on any trial each event (face or number) is equally likely to occur since there are six equally likely exhaustive events, each will occur 1/6 of the time, and therefore the probability of any one event occurring is 1/6. Similarly for the process of selecting a card at random, each event or card is mutually exclusive, exhaustive, and equiprobable. The probability of selecting any one card on a trial is equal to 1/52, since there are 52 cards. Hence, in

general, for random experiment with mutually exclusive, exhaustive, equiprobable events, the probability of any of the events is equal to $1/n$.

Since the probability of occurrence of an event is based on prior knowledge of the process involved, therefore this approach is often called a priori classical probability approach. This means, we do not have to perform random experiments to find the probability of occurrence of an event. This also implies that no experimental data are required for computation of probability. Since the assumption of equally likely simple events can rarely be verified with certainty, therefore this approach is not used often other than in games of chance.

The assumption that all possible outcomes are equally likely may lead to a wrong calculation of probability in case some outcomes are more or less frequent in occurrence. For example, if we classify two children in a family according to their sex, then the possible outcomes in terms of number of boys in the family are 0, 1, 2. Thus according to the classical approach, the probability for each of the outcomes should be $1/3$. However, it has been calculated that the probabilities are approximately $1/4$, $1/2$, and $1/4$ for 0, 1, 2 boys respectively. Similarly, we cannot apply this approach to find the probability of a defective unit being produced by a stable manufacturing process as there are only two possible outcomes, defective or non-defective.

2. Relative Frequency Approach

In situations where the outcomes of a random experiment are not all equally likely or when it is not known whether outcomes are equally likely, application of the classical approach is not desirable to quantify the possible occurrence of a random event. For example, it is not possible to state in advance, without repetitive trials of the experiment, the probabilities in cases like (i) whether a number greater than 3 will appear when die is rolled or (ii) if a lot of 100 items will include 10 defective items.

This approach of computing probability is based on the assumption that a random experiment can be repeated a large number of times under identical conditions where trials are independent to each other. While conducting a random experiment, we may or may not observe the desired event. But as the experiment is repeated many times, that event may occur some proportion of time. Thus, the approach calculates the proportion of the time (i.e. the relative frequency) with which the event occurs over an infinite number of repetitions of the experiment under identical conditions. Since no experiment can be repeated an infinite number of times, therefore a probability can never be exactly determined. However, we can approximate the probability of an event by recording the relative frequency with which the event has occurred over a finite number of repetitions of the experiment under identical conditions. For example, if a die is tossed n times and s denotes the number of times the event A (i.e., number 4, 5, or 6) occurs, then the ratio $P(A) = c(s)/n$ gives the proportions of times the event A occurs in n trials, and are also called relative frequencies of the event in n trials. Although our estimate about $P(A)$ may change after every trial, yet we will find that the proportion $c(s)/n$ tends to cluster around a unique central value as the number of trials n becomes even larger. This unique central value (also called probability of event A) is defined as:

$$p(A) = \lim_{n \rightarrow \infty} \left\{ \frac{c(s)}{n} \right\}$$

Where $c(s)$ represents the number of times that an event s occurs in n trials of an experiment.

Since the probability of an event is determined objectively by repetitive empirical observations of experimental outcomes, it is also known as empirical probability. Few situations to which this approach can be applied are follows:

- i. Buying lottery tickets regularly and observing how often you win
- ii. Commuting to work daily and observing whether or not a certain traffic signal is red when cross it.
- iii. Observing births and noting how often the baby is a female
- iv. Surveying many adults and determine what proportion smokes.

3. Subjective Approach

The subjective approach of calculating probability is always based on the degree of beliefs, convictions, and experience concerning the likelihood of occurrence of a random event. It is thus a way to quantify an individual's beliefs, assessment, and judgment about a random phenomenon. Probability assigned for the occurrence of an event may be based on just guess or on having some idea about the relative frequency of past occurrences of the event. This approach must be used

when either sufficient data are not available or sources of information giving different results are not known.

8.4 Basic Concepts of Probability

Probability, in common parlance, connotes the chance of occurrence of an event or happening. In order that we are able to measure it, a more formal definition is required. This is achieved through the study of certain basic concepts in probability theory, like experiment, sample space and event. In this section we explore these concepts.

Experiment

The term experiment is used in probability theory in a much broader sense than in physics or chemistry. Any action, whether it is the tossing of a coin, or measurement of a product's dimension to ascertain quality, or the launching of a new product in the market, constitute an experiment in the probability theory terminology.

These experiments have three things in common:

- 1 There is two or more outcomes of each experiment.
- 2 It is possible to specify the outcomes in advance.
- 3 There is uncertainty about the outcomes.



For example, a coin tossing may result in two outcomes, in head or tail, which we know in advance, and we are not sure whether a head or a tail will come up when we toss the coin. Similarly, the product we are measuring may turn out to be undersize or right size or oversize, and we are not certain which way it will be when we measure it. Also, launching a new product involves uncertain outcome of meeting with a success or failure in the market.

Sample Space

The set of all possible outcomes of an experiment is defined as the sample space. Each outcome is thus visualized as a sample point in the sample space. Thus, the set (head, tail) defines the sample space of a coin tossing experiment. Similarly, (success, failure) defines the sample space for the launching experiment. You may note here, that given any experiment, the sample space is fully determined by listing down all the possible outcomes of the experiment.

Event

An event, in probability theory, constitutes one or more possible outcomes of an experiment. Thus, an event can be defined as a subset of the sample space. Unlike the common usage of the term, where an event refers to a particular happening or incident, here, we use an event to refer to a single outcome or a combination of outcomes. Suppose, as a result of a market study experiment of a product, we find that the demand for the product for the next month is uncertain, and may take values from 100, 101, 102... 150. We can obtain different event like:

The event that demands is exactly 100.

The event that demands lies between 101 to 120.

The event that demands is 101 or 102.

In the first case, out of the 51 sample points that constitute the sample space, only one sample point or outcome defines the event, whereas the number of outcomes used in the second and third case is 20 and 2 respectively.

With this background on the above concepts, we are now in a position to formalize the definition of probability of an event. In the next section, we will look at the different approaches to probability that have been developed, and present the axioms for the definition of probability.



Example 1: Suppose we are interested in the following Event A in the above experiment: The number of defectives is exactly two. How many sample' points does this event correspond to?

Solution:

We can see from the sample space that there are three outcomes where D occurs twice, viz, DDG, DGD and GDD, thus the Event A corresponds to 3 sample point.

Exhaustive Cases

The total number of possible outcomes in a random experiment is called the exhaustive cases. In other words, the number of elements in the sample space is known as number of exhaustive cases, e.g.

- i. If we toss a coin, then the number of exhaustive cases is 2 and the sample space in this case is {H, T}.
- ii. If we throw a die then number of exhaustive cases is 6 and the sample space in this case is {1, 2, 3, 4, 5, and 6}.

Favorable Cases

The cases which favor to the happening of an event are called favorable cases. e.g.

- i. For the event of drawing a card of spade from a pack of 52 cards, the number of favorable cases is 13.
- ii. For the event of getting an even number in throwing a die, the number of favorable cases is 3 and the event in this case is {2, 4, 6}.

Mutually Exclusive Cases

Cases are said to be mutually exclusive if the happening of any one of them prevents the happening of all others in a single experiment, e.g.

- i. In a coin tossing experiment head and tail are mutually exclusive as there cannot be simultaneous occurrence of head and tail.

Equally Likely Cases

Cases are said to be equally likely if we do not have any reason to expect one in preference to others. If there is some reason to expect one in preference to others, then the cases will not be equally likely, For example,

- i. Head and tail are equally likely in an experiment of tossing an unbiased coin. This is because if someone is expecting say head, he/she does not have any reason as to why he/she is expecting it.
- ii. All the six faces in an experiment of throwing an unbiased die are equally likely.

You will become more familiar with the concept of “equally likely cases” from the following examples, where the non-equally likely cases have been taken into consideration.

- i. Cases of “passing” and “not passing” a candidate in a test are not equally likely. This is because a candidate has some reason(s) to expect “passing” or “not passing” the test. If he/she prepares well for the test, he/she will pass the test and if he/she does not prepare for the test, he/she will not pass. So, here the cases are not equally likely.
- ii. Cases of “falling a ceiling fan” and “not falling” are not equally likely. This is because; we can give some reason(s) for not falling if the bolts and other parts are in good condition.

8.5 Probability Rules

In probability we use set theory notations to simplify the presentation of ideas. As discussed earlier in this chapter, the probability of the occurrence of an event A is expressed as:

$P(A)$ =probability of event A occurrence

Such single probabilities are called marginal (or unconditional) probabilities because it is the probability of a single event occurring. In the coin tossing example, the marginal probability of a tail or head in a toss can be stated as $P(T)$ or $P(H)$.

8.6 Rules of Addition

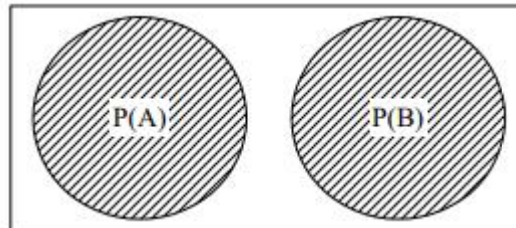
The addition rules are helpful when we have two events and are interested in knowing the probability that at least one of the events occurs.

1. Addition Rule for Mutually Exclusive Events

If two events, A and B, are mutually exclusive, then the probability of occurrence of either A or B is given by the following formula:

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

If A and B are mutually exclusive events, this rule is depicted in Figure 8.1, below



The essential requirement for any two events to be mutually exclusive is that there are no outcomes common to the occurrence of both. This condition is satisfied when sample space does not contain any outcome favorable to the occurrence of both A and B means $A \cap B = \varnothing$



Example 2: In a game of cards, where a pack contains 52 cards, 4 categories exist namely spade, club, diamond, and heart. If you are asked to draw a card from this pack, what is the probability that the card drawn belongs to either spade or club category

Solution: Here, $P(\text{Spade or club}) = 13/52 = 1/4$, $P(\text{Club}) = 13/52 = 1/4$

$$P(\text{Spades}) + P(\text{Club}) = 1/4 + 1/4 = 1/2$$

There is an important special case for any event E, either E happens or it does not. So, the events E and not E are exhaustive and exclusive.

$$\text{So, } P(E) + P(\text{not } E) = 1$$

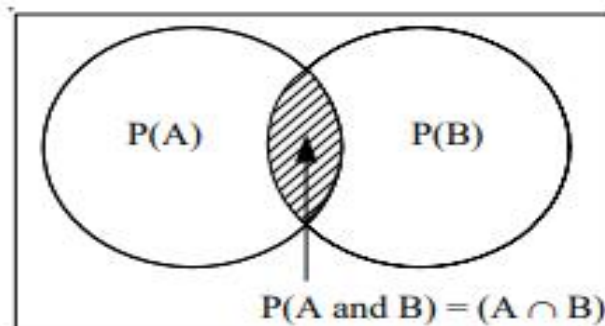
$$\text{Or, } P(E) = 1 - P(\text{not } E)$$

Sometimes $P(\text{not } E)$ is also written as either $P(E) = 1 - P(\bar{E})$

$$\text{So, } P(E) = 1 - P(\bar{E}) = 1 - P(\bar{E})$$

2. Addition Rule for Non-Mutually Exclusive Events

Non-mutually exclusive (overlapping) events present another significant variant of the additive rule. Two events (A and B) are not mutually exclusive if they have some outcomes common to the occurrence of both, then the above rule has to be modified in order to account for the overlapping areas, as it is clear from Figure 8.2. Below.



In this situation, the probability of occurrence of event A or event B is given by the formula

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \text{ and } B)$$

Where $P(A \text{ and } B)$ is the joint probability of events A and B , i.e., both occurring together and is usually written as $P(A \cap B)$.

Thus, it is clear that the probability of outcomes that are common to both the events is to be subtracted from the sum of their simple probability.



Example 3: The event of drawing either a Jack or a spade from a well shuffled deck of playing cards. Find the probability.

Solution: These events are not mutually exclusive, so the required probability of drawing a Jack or a spade is given by:

$$P(\text{Jack or Spade}) = P(\text{Jack}) + P(\text{Spade}) - P(\text{Jack and Spade})$$

$$= 4/52 + 13/52 - 1/52 = 16/52 = 4/13$$

8.7 Rules of Multiplication

Statistically Independent Event:

When the occurrence of an event does not affect and is not affected by the probability of occurrence of any other event, the event is said to be a statistically independent event. There are three types of probabilities under statistical independence: marginal, joint, and conditional.

Probability under Statistical Independence

The two or more events are termed as statistically independent events, if the occurrence of any one event does not have any effect on the occurrence of any other event. For example, if a fair coin is tossed once and supposes head comes, then this event has no effect in any way on the outcome of second toss of that same coin. Similarly, the results obtained by drawing hearts from a pack have no effect in any way on the results obtained by throwing a dice. These events thus are being termed as statistically independent events. There are three types of probability under statistically independent case.

- a) Marginal Probability;
- b) Joint Probability;
- c) Conditional Probability

a. Marginal Probability under Statistical Independence

A Marginal/Simple/Unconditional probability is the probability of the occurrence of an event. For example, in a fair coin toss, probability of having a head is:

$$P(H) = \frac{1}{2} = 0.5$$

Therefore, the marginal probability of an event (i.e. having a head) is 0.5. Since, the subsequent tosses are independent of each other; therefore, it is a case of statistical independence.

Another example can be given in a throw of a fair die, the marginal probability of the face bearing number 3, is:

$$P(3) = \frac{1}{6} = 0.166$$

Since, the tosses of the die are independent of each other; this is a case of statistical independence.

b. Joint Probability under Statistical Independence

This is also termed as "Multiplication Rule of Probability". In many situations we are interested in finding out the probability of two or more events either occurring together or in quick succession to each other, for this purpose the concept of joint probability is used.

This joint probability of two or more statistically independent events occurring together is determined by the product of their marginal probability. The corresponding formula may be expressed as:

$$P(A \text{ and } B) = P(A) \times P(B)$$

Similarly, it can be extended to more than two events also as:

$$P(A \text{ and } B \text{ and } C) = P(A) \times P(B) \times P(C) \text{ and so on.}$$

$$\text{i.e. } P(A \text{ and } B \text{ and } C \text{ and } \dots) = P(A) \times P(B) \times P(C) \times \dots$$

For instance, when a fair coin is tossed twice in quick succession, the probability of head occurring in both the tosses is:

$$P(H_1 \text{ and } H_2) = P(H_1) \times P(H_2)$$

$$= 0.5 \times 0.5 = 0.25$$

Where, H_1 is the occurrence of head in 1st toss, and H_2 is the occurrence of head in 2nd toss

Take another example: When a fair die is thrown twice in quick succession, then to find the probability of having 2 in the 1st throw and 4 in second throw is, given as:

$$P(2 \text{ in } 1\text{st} \text{ throw and } 4 \text{ in } 2\text{nd} \text{ throw})$$

$$= P(2 \text{ in the } 1\text{st} \text{ throw}) \times P(4 \text{ in the } 2\text{nd} \text{ throw})$$

$$= 1/6 \times 1/6 = 1/36 = 0.028$$

c. Conditional Probability under the Condition of Statistical Independence

The third type of probability under the condition of statistical independence is the Conditional Probability. It is symbolically written as $P(A/B)$, i.e., the conditional probability of occurrence of event A, on the condition that event B has already occurred.

In case of statistical independence, the conditional probability of any event is akin to its marginal probability, when both the events are independent of each other.

Therefore, $P(A/B) = P(A)$, and

$$P(B/A) = P(B).$$



For example, if we want to find out what is the probability of heads coming up in the second toss of a fair coin, given that the first toss has already resulted in head. Symbolically, we can write it as:

$$P(H_2/H_1)$$

As, the two tosses are statistically independent of each other

$$\text{so, } P(H_2/H_1) = P(H_2)$$

The following table 8.1 summarizes these three types of probabilities, their symbols and their mathematical formulae under statistical independence.

Probability's type	Symbol	Formula
Marginal	$P(A)$	$P(A)$
Joint	$P(AB)$	$P(A) \times P(B)$
Conditional	$P(B/A)$	$P(B)$

Probability under Statistical Dependence

Two or more events are said to be statistically dependent, if the occurrence of any one event affects the probability of occurrence of the other event.

There are three types of probability under statistical dependence case. They are:

- a) Conditional Probability;
- b) Joint Probability;
- c) Marginal Probability

a) Conditional Probability under Condition of Statistical Dependence

The conditional probability of event A, given that the event B has already occurred, can be calculated as follows:

$$P(A / B) = \frac{P(AB)}{P(B)}$$

Where, P (AB) is the joint probability of events A and B.



Example 4: (i) A box containing 10 balls which have the following distribution on the basis of color and pattern.

- a) 3 are colored and dotted.
- b) 1 is colored and stripped.
- a) Suppose someone draws a colored ball from the box. Find what is the probability that it is (i) dotted and (ii) it is stripped?

Solution: The problem can be expressed as P (D/C) i.e., the conditional probability that the ball drawn is dotted given that it is colored.

Now from the information given in the question.

- i) P (CD) = 3/10 = Joint Probability of drawn ball becoming a colored as well as a dotted one.
Similarly, P (CS) = 1/10, P (GD) = 2/10, and P (GS) = 4/10

$$P(D / C) = \frac{P(DC)}{P(C)}$$

Where, P(C) = Probability of drawing a colored ball from the box = 4/10 (4 colored balls out of 10 balls).

$$P(D/C) = \frac{\frac{3}{10}}{\frac{4}{10}} = 0.75$$

- ii) Similarly, P(S/C) = Conditional probability of drawing a stripped ball on the condition of knowing that it is a colored one.

$$P(D/C) = \frac{\frac{1}{10}}{\frac{4}{10}} = 0.25$$

Thus, the probability of colored and dotted ball is 0.75. Similarly, the probability of colored and stripped ball is 0.25.

- iii) Continuing the same illustration, if we wish to find the probability of (i) P (D/G) and (ii) P (S/G)

Solution:

$$P(D / G) = \frac{P(DG)}{P(G)} = 2/10/6/10 = 1/3 = 0.33$$

where, P(G) = Total probability of grey balls, i.e., 6/10 and

$$\text{II) } P(S/G) = \frac{P(SG)}{P(G)} = (4/10)/(6/10) = 2/3 = 0.66$$

b) Joint Probability under the Condition of Statistical Dependence

This is an extension of the multiplication rule of probability involving two or more events, which have been discussed in the previous section 13.6, for calculating joint probability of two or more events under the statistical independence condition.

The formula for calculating joint probability of two events under the condition of statistical independence is derived from the formula of Bayes' Theorem.

Therefore, the joint probability of two statistically dependent events A and B is given by the following formula:

$$P (AB) = P (A/B) \times P (B)$$

$$\text{Or } P(BA) = P(B/A) \times P(A)$$

Depending upon whether order of occurrence of two events is B, A or A, B.

Since, $P(A/B) = P(B/A)$, so the product on the RHS of the formula must also be equal to each other.

$$\therefore P(A/B) \times P(B) = P(B/A) \times P(A)$$

Notice that this formula is not the same under conditions of statistical independence, i.e., $P(BA) = P(B) \times P(A)$. Continuing with our previous illustration 4, of a box containing 10 balls, the value of different joint probabilities can be calculated as follows:

Converting the above general formula i.e., $P(AB) = P(A/B) \times P(B)$ into our illustration and to the terms colored, dotted, stripped, and grey, we would have calculated the joint probabilities of $P(CD)$, $P(GS)$, $P(GD)$, and $P(CS)$ as follows:

$$\text{i) } P(CD) = P(C/D) \times P(D) = 0.6 \times 0.5 = 0.3$$

$$\text{ii) } P(GS) = P(G/S) \times P(S) = 0.8 \times 0.5 = 0.4$$

c) Marginal Probability under the Condition of Statistical Dependence

Finally, we discuss the concept of marginal probability under the condition of statistical dependence. It can be computed by summing up all the probabilities of those joint events in which that event occurs whose marginal probability we want to calculate.



Example 5: Consider the previous illustration 4, to compute the marginal probability under statistical dependence of the event: i) dotted balls occurred, ii) colored balls occurred, iii) grey balls occurred, and iv) stripped balls occurred.

Solution: We can obtain the marginal probability of the event dotted balls by adding the probabilities of all the joint events in which dotted balls occurred.

$$P(D) = P(CD) + P(GD) = 3/10 + 2/10 = 0.5$$

In the same manner, we can compute the joint probabilities of the remaining events as follows:

$$\text{i) } P(C) = P(CD) + P(CS) = 3/10 + 1/10 = 0.4$$

$$\text{ii) } P(G) = P(GD) + P(GS) = 2/10 + 4/10 = 0.6$$

The following table 8.2 summarizes three types of probabilities, their symbols and their mathematical formulae under statistical dependence.

Table 8.2: Probabilities under Statistical Dependence

Probability's type	Symbol	Formula
Marginal	$P(A)$	Sum of the probabilities of joint events in which 'A' occurs
Joint	$P(AB)$ or $P(BA)$	$P(A/B) \times P(B)$ OR $P(B/A) \times P(A)$
Conditional	$P(B/A)$ or $P(A/B)$	$P(B/A)/P(A)$ OR $P(A/B)/P(B)$

Summary

At the beginning of this unit the historical evolution and the meaning of probability has been discussed. Contribution of leading mathematicians has been highlighted. Fundamental concepts and approaches to determining probability have been explained. The three approaches namely; the classical, the relative frequency, and the subjective approaches are used to determine the probability in case of risky and uncertain situation have been discussed. Probability rules for calculating probabilities of different types of events have been explained. Further the condition of statistical independence and statistical dependence has been defined. Three types of probabilities

namely: marginal, joint and conditional under statistical independence and statistical dependence have been explained. Finally, the Bayesian approach to the revision of a priori probability in the light of additional information has been undertaken.

Keywords

- Classical/Logical Approach: An objective way of assessing probabilistic value based on logic.
- Collectively Exclusive Event: This is the collection of all possible outcomes of an experiment.
- Conditional Probability: The probability of the happening of an event on the condition that another event has already occurred.
- Dependent Event: This is the situation in which the occurrence of one event affects the happening of another event.
- Independent Event: This is the situation in which the occurrence of an event has no effect on the probability of the occurrence of any other event.
- Joint Probability: The probability of occurring of events together or in quick succession
- Marginal/Simple Probability: As the name suggests, it is the simple probability of occurrence of an event.
- Mutually Exclusive Events: A situation in which only one event can occur on any given trial/experiment. It means events that cannot occur together.

Self Assessment

1. _____ is a mechanism that produces a definite outcome that _____.
 - A. Dependent experiment; can be predicted with certainty.
 - B. Simple experiment; cannot be predicted with certainty.
 - C. Random experiment; can be predicted with certainty.
 - D. Random experiment; cannot be predicted with certainty.
2. Consider the experiment of tossing three coins simultaneously. The sample space is given by
 - A. $S = \{HHH, HHT, HTH, HHH, HTT, THT, TTH, TTT\}$
 - B. $S = \{HHH, HHT, HTH, THH, TTT, THT, TTH, TTT\}$
 - C. $S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$
 - D. $S = \{HHH, HHT, HTH, THH, THT, THT, TTH, TTT\}$
3. A set of events $\{A_1, A_2, \dots, A_n\}$ is collectively exhaustive if
 - A. Union of A_1, A_2, \dots, A_n is identical with the sample space, $S = \{A_1 \cup A_2 \cup \dots \cup A_n\}$.
 - B. Union of A_1, A_2, \dots, A_n is not identical with the sample space, $S = \{A_1 \cup A_2 \cup \dots \cup A_n\}$.
 - C. Union of A_1, A_2, \dots, A_n is equal to one.
 - D. Union of A_1, A_2, \dots, A_n lie between zero and one.
4. Compound events
 - A. Are dependent only.
 - B. Are independent only

- C. May be dependent or independent.
 D. Neither dependent nor independent.
5. _____ is a way to quantify an individual's belief, assessment, and judgement about a random experiment.
 A. Relative frequency approach
 B. Subjective approach
 C. Classical approach
 D. Fundamental approach
6. If A and B are any two events then the probability of happening of at least one of the events is defined as
 A. $P(A \cup B) = P(A)/P(B)$
 B. $P(A \cup B) = P(A) - P(B)$
 C. $P(A \cup B) = P(A) + P(B)$
 D. $P(A \cup B) = P(A) * P(B)$
7. Partially overlapping events are those events which
 A. Are mutually exclusive.
 B. Are not mutually exclusive.
 C. Have no sample points in common.
 D. Are mutually exclusive and have no point in common.
8. Which of the following is correct?
 A. $P(A \cup B \cup C) = P(A) + P(B) - P(A \cap B) + P(C) - [P(A \cap C) + P(B \cap C) - P((A \cap C) \cap (B \cap C))]$
 B. $P(A \cup B \cup C) = P(A) + P(B) + P(A \cap B) + P(C) - [P(A \cap C) + P(B \cap C) - P((A \cap C) \cap (B \cap C))]$
 C. $P(A \cup B \cup C) = P(A) + P(B) - P(A \cap B) - P(C) - [P(A \cap C) + P(B \cap C) - P((A \cap C) \cap (B \cap C))]$
 D. $P(A \cup B \cup C) = P(A) + P(B) - P(A \cap B) + P(C) - [P(A \cap C) + P(B \cap C) + P((A \cap C) \cap (B \cap C))]$
9. Joint Probability (two or more independent events)
 A. = Product of marginal probabilities of only two independent events
 B. = Product of marginal probabilities of two or more independent events
 C. = Product of marginal probabilities of two or more dependent events
 D. = Product of marginal probabilities of only two dependent events
10. The conditional probability of B, given that event A has already occurred, is given by
 A. $P(B | A) = P(A)/P(A \cap B)$
 B. $P(A | B) = P(A \cap B)/P(B)$
 C. $P(B | A) = P(A \cap B)/P(B)$
 D. $P(B | A) = P(A \cap B)/P(A)$
11. If two events (both with probability greater than 0) are mutually exclusive, then:
 A. They also must be independent.

- B. They also could be independent.
 C. They cannot be independent.
 D. None of the above
12. Suppose that the probability of event A is 0.2 and the probability of event B is 0.4. Also, suppose that the two events are independent. Then $P(A | B)$ is:
 A. $P(A) = 0.2$
 B. $P(A)/P(B) = 0.2/0.4 = \frac{1}{2}$
 C. $P(A) \times P(B) = (0.2)(0.4) = 0.08$
 D. None of the above.
13. The range of probability is
 A. any value greater than zero
 B. any value less than one
 C. zero to one
 D. any value between -1 to 1
14. Two events A and B are statistically independent when
 A. $P(A \cap B) = P(A) \times P(B)$
 B. $P(A | B) = P(A)$
 C. $P(A \cup B) = P(A) + P(B)$
 D. Both (a) and (b)
15. Which of the following pairs of events are mutually exclusive?
 A. A contractor loses a major contract and he increases his work force by 50 per cent.
 B. A man is older than his uncle and he is younger than his cousins.
 C. A football team loses its last game of the year, and it wins the world cup.
 D. none of these

Answers for Self Assessment

1. D 2. C 3. A 4. C 5. B
 6. C 7. B 8. A 9. B 10. D
 11. C 12. C 13. C 14. D 15. C

Review Questions

1. Explain whether or not each of the following claims could be correct:
 a) A businessman claims the probability that he will get contract A is 0.15 and that he will get contract B is 0.20. Furthermore, he claims that the probability of getting A or B is 0.50.
 b) A market analyst claims that the probability of selling ten million rupees of plastic A or five million rupees of plastic B is 0.60. He also claims that the probability of selling ten million rupees of A and five million rupees of B is 0.45.

2. Explain what you understand by the term probability. Discuss its importance in business decision-making.
3. Define independent and mutually exclusive events. Can two events be mutually exclusive and independent simultaneously? Support your answer with an example.
4. Explain the meaning of each of the following terms:
(a) Random phenomenon (b) Statistical experiment (c) Random event (d) Sample space
5. Distinguish between the two concepts in each of the following pairs: (a) Elementary event and compound events (b) Mutually exclusive events and overlapping events (c) Sample space and sample point
6. Suppose an entire shipment of 1000 items is inspected and 50 items are found to be defective. Assume the defective items are not removed from the shipment before being sent to retail outlet for sale. If you purchase one item from this shipment, what is the probability that it will be one of the defective items?
7. Life insurance premiums are higher for older people, but auto insurance premiums are generally higher for younger people. What does this suggest about the risks and probabilities associated with these two areas of insurance business?
8. A problem in business statistics is given to five students, A, B, C, D, and E. Their chances of solving it are $1/2$, $1/3$, $1/4$, $1/5$, and $1/6$ respectively. What is the probability that the problem will be solved?
9. There are three brands, say X, Y, and Z of an item available in the market. A consumer chooses exactly one of them for his use. He never buys two or more brands simultaneously. The probabilities that he buys brands X, Y, and Z are 0.20, 0.16, and 0.45.
(a) What is the probability that he does not buy any of the brands? (b) Given that a customer buys some brand, what is the probability that he buys brand X?
10. Two sets of candidates are competing for positions on the board of directors of a company. The probability that the first and second sets will win are 0.6 and 0.4 respectively. If the first set wins, the probability of introducing a new product is 0.8 and the corresponding probability if the second set wins are 0.3. (a) What is the probability that the new product will be introduced? (b) If the new product was introduced, what is the probability that the first set won as directors?



Further Readings

- Levin, R.I. and Rubin, D.S., 1991, *Statistics for Management*, PHI, : New Delhi.
- Feller, W., 1957, *An Introduction to Probability Theory and Its Applications*, John Wiley & Sons Inc. : New York.
- Hooda, R.P. 2001, *Statistics for Business and Economics*. MacMillan India Limited, Delhi.
- Gupta S.C., and V.K. Kapoor, 2005, *Fundamentals of Mathematical Statistics*, Sultan Chand & Sons, Delhi.

Unit 09: Probability Distribution

CONTENTS

Objectives

Introduction

9.1 Concepts Of Probability Distributions

9.2 Probability Distribution Function (pdf)

9.3 Discrete Probability Distributions

9.4 Continuous Probability Distributions

9.5 Normal Distribution/ Normal Probability Curve

Summary

Keywords

Self Assessment

Answers for Self Assessment

Review Questions

Further Readings

Objectives

- understand the random variables and how they are inseparable to probability distributions.
- solve the problems of probability, which fit into binomial and normal distributions.

Introduction

In our study of Probability Theory, we have so far been interested in specific outcomes of an experiment and the chances of occurrence of these outcomes. In the last unit, we have explored different ways of computing the probability of an outcome. For example, we know how to calculate the probability of getting all heads in a toss of three coins. We recognize that this information on probability is helpful in our decisions. In this case, a mere 0.125 chance of all heads may dissuade you from betting on the event of "all heads". It is easy to see that it would have been further helpful, if all the possible outcomes of the experiment together with their chances of occurrence were made available. Thus, given your interest in betting on head's, you find that a toss of three coins may result in zero, one, two or three heads with the respective probabilities of $1/8$, $3/8$, $3/8$, and $1/8$. The wealth of information, presented in this way, helps you in drawing many different inferences. Looking at this information, you may be more ready to bet on the event that either one or two heads occur in a toss of three coins. This representation of all possible outcomes and their probabilities is known as a probability distribution. Thus, we refer to this as the probability distribution of "number of heads" in the experiment of tossing of three coins. While we see that our previous knowledge on computation of probabilities helps us in arriving at such representations, we recognize that the calculations may be quite tedious. This is apparent, if you try to calculate the probabilities of different number of heads in a tossing of twelve coins. Developments in Probability Theory help us in specifying the probability distribution in such cases with relative ease. The theory also gives certain standard probability distributions and provides the conditions under which they can be applied. We will study the probability distributions and their applications in this and the subsequent unit. The objective of this unit is to look into a type of probability distribution, viz., a discrete probability distribution. Accordingly, after the initial presentation on the basic concepts and definitions, we will discuss as to how discrete probability distributions can be used in decision-making.

9.1 Concepts Of Probability Distributions

Before we attempt a formal definition of probability distribution, the concept of 'random variable' which is central to the theme, needs to be elaborated.

In the example given in the Introduction, we have seen that the outcomes of the experiment of a toss of three coins were expressed in terms of the "number of heads". Denoting this "number of heads" by the letter H, we find that in the example, H can assume values of 0, 1, 2 and 3 and corresponding to each value, a probability is associated. This uncertain real variable H, which assumes different numerical values depending on the outcomes of an experiment, and to each of whose values a probability assignment can be made, is known as a random variable. The resulting representation of all the values with their probabilities is termed as the probability distribution of H. It is customary to present the distribution as follows:

H	P(H)
0	0.125
1	0.375
2	0.375
3	0.125

In this case, as we find that H takes only discrete values, the variable H is called a discrete random variable and the resulting distribution is a discrete probability distribution.

In the above situation, we have seen that the random variable takes a limited number of values. There are certain situations where the variable of interest may take infinitely many values. Consider for example that you are interested in ascertaining the probability distribution of the weight of the one kilogram tea pack, that is produced by your company. You have reasons to believe that the packing process is such that the machine produces a certain percentage of the packs slightly below one kilogram and some above one kilogram. It is easy to see that there is essentially to chance that the pack will weigh exactly 1.000000 kg., and there are infinite number of values that the random variable "weight" can take. In such cases, it makes sense to talk of the probability that the weight will be between two values, rather than the probability of the weight will be between two values, rather than the probability of the weight taking any specific value. These types of random variables which can take an infinitely large number of values are called continuous random variables, and the resulting distribution is called a continuous probability distribution. Sometimes, for the sake of convenience, a discrete situation with a large number of outcomes is approximated by a continuous distribution: Thus, if we find that the demand of a product is a random variable taking values of 1, 2, 3... to 1000, it may be worthwhile to treat it as a continuous variable. Obviously, the representation of the probability distribution for a continuous random variable is quite different from the discrete case that we have seen. We will be discussing this in a later unit when we take up continuous probability distributions.

Coming back to our example on the tossing of three coins, you must have noted the presence of another random variable in the experiment, namely, the number of tails (say T). T has got the same distribution as H. In fact, in the same experiment, it is possible to have some more random variables, with a slight extension of the experiment. Supposing a friend comes and tells you that he will toss 3 coins, and will pay you Rs. 100 for each head and Rs. 200 for each tail that turns up. However, he will allow you this privilege only if you pay him Rs. 500 to start with.

You may like to know whether it is worthwhile to pay him Rs. 500. In this situation, over and above the random variables H and T, we find that the money that you may get is also a random variable. Thus,

if H = number of heads in any outcome, then $3 - H$ = number of tails in any outcome (as the total number of heads and tails that can occur in a toss of three coins is 3) The money you get in any outcome = $100H + 200(3 - H) = 600 - 100H = x$ (say)

Unit 09: Probability Distribution

We find that x which is a function of the random variable H , is also a random variable.

We can see that the different values x will take in any outcome are

$$(600 - 100 \times 0) = 600$$

$$(600 - 100 \times 1) = 500$$

$$(600 - 100 \times 2) = 400$$

$$(600 - 100 \times 3) = 300$$

Hence the distribution of x is:

X	P(X)
600	1/8
500	3/8
400	3/8
300	1/8

The above gives you the probability of your getting different sums of money. This may help you in deciding whether you should utilise this opportunity by paying Rs. 500.

9.2 Probability Distribution Function (pdf)

Probability distribution functions can be classified into two categories:

1. Discrete probability distributions
2. Continuous probability distributions
 - a. A discrete probability distribution assumes that the outcomes of a random variable under study can take on only integer values, such as:
 - A book shop has only 0, 1, 2, ... copies of a particular title of a book
 - A consumer can buy 0, 1, 2, ... shirts, pants, etc.

If the random variable x is discrete, its probability distribution called probability mass function (pmf) must satisfy following two conditions:

- (i) The probability of a any specific outcome for a discrete random variable must be between 0 and 1. Stated mathematically, $0 \leq f(x=k) \leq 1$, for all value of k
- (ii) The sum of the probabilities over all possible values of a discrete random variable must equal 1. Stated mathematically, all $\sum_{all\ k} f(x = k) = 1$
- b. A continuous probability distribution assumes that the outcomes of a random variable can take on only value in an interval such as:

If the random variable x is continuous, then its probability density function must satisfy following two conditions:

- i) $P(x) \geq 0$; $-\infty < x < \infty$ (non-negativity condition)
- ii) $\int_{-\infty}^{\infty} (P \times dx) = 1$ (Area under the continuous curve must total 1)

9.3 Discrete Probability Distributions

Binomial Probability Function

Binomial distribution which was discovered by J. Bernoulli (1654-1705) and was first published eight years after his death i.e. in 1713 and is also known as "Bernoulli distribution for n trials".

Research Methods and Design

Binomial distribution is applicable for a random experiment comprising a finite number (n) of independent Bernoulli trials having the constant probability of success for each trial.

Before defining binomial distribution, let us consider the following example: Suppose a man fires 3 times independently to hit a target. Let p be the probability of hitting the target (success) for each trial and $q (= 1 - p)$ be the probability of his failure.

Let S denote the success and F the failure. Let X be the number of successes in 3 trials,

$P[X = 0]$ = Probability that target is not hit at all in any trial

= P [Failure in each of the three trials]

= $P(F \cap F \cap F)$

= $P(F) \cdot P(F) \cdot P(F)$ [trials are independent]

= $q \cdot q \cdot q$

= q^3

This can be written as

$$P[X=3] = {}^3C_3 p^3 q^{3-3} \quad [b: {}^3C_3 = 1, q^{3-3} = 1]$$

From the above four rectangle results, we can write

$$P[X=r] = {}^3C_r p^r q^{3-r}$$

which is the probability of r successes in 3 trials. 3C_r , here, is the number of ways in which r successes can happen in 3 trials.

The result can be generalized for n trials in the similar fashion and is given as $P[X=r] = {}^nC_r p^r q^{n-r}$

$r=0, 1, 2, \dots, n$.

1)

This distribution is called the binomial probability distribution. The reason behind giving the name binomial probability distribution for this probability distribution is that the probabilities for $x = 0, 1, 2, \dots, n$ are the respective probabilities ${}^nC_0 p^0 q^{n-0}, {}^nC_1 p^1 q^{n-1}, {}^nC_n p^n q^{n-n}$ which are the successive terms of the binomial expansion $(q + p)^n$.



Example 1: An unbiased coin is tossed six times. Find the probability of obtaining

- (i) exactly 3 heads
- (ii) less than 3 heads
- (iii) more than 3 heads

Solution: Let p be the probability of getting head (success) in a toss of the coin and n be the number of trials.

$\therefore n = 6, p = 1/2$ and hence $q = 1 - p = 1 - 1/2 = 1/2$.

Let X be the number of successes in n trials,

\therefore by binomial distribution, we have

$$P[X=r] = {}^nC_r p^r q^{n-r}, \quad r=0, 1, 2, \dots, n.$$

$$= {}^6C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{6-x}; \quad x = 1, 2, \dots, 6$$

$$= {}^6C_x \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^{6-0}; \quad x = 1, 2, \dots, 6$$

$$= 1/64 \cdot {}^6C_x, \quad x = 1, 2, \dots, 6$$

Therefore,

$$(i) \quad P[\text{exactly 3 heads}] = P[X = 3]$$

$$= 1/64 ({}^6C_3)$$

$$= 1/64 (6 \cdot 5 \cdot 4 / 3 \cdot 2) = 5/16$$

$$(ii) \quad P[\text{less than 3 heads}] = P[X < 3]$$

$$\begin{aligned}
&= P[X=2 \text{ or } X=1 \text{ or } X=0] \\
&= P[X=2] + P[X=1] + P[X=0] \\
&= 1/64 [6C_2 + 6C_1 + 6C_0] \\
&= 1/64 [(6 \cdot 5)/2 + 6 + 1] \\
&= 22/64 = 11/32
\end{aligned}$$

$$\begin{aligned}
\text{iii) } P[\text{more than 3 heads}] &= P[X > 3] \\
&= P[X=4 \text{ or } X=5 \text{ or } X=6] \\
&= P[X=4] + P[X=5] + P[X=6] \\
&= 1/64 [6C_4 + 6C_5 + 6C_6] \\
&= 1/64 [(6 \cdot 5)/2 + 6 + 1] \\
&= 22/64 = 11/32
\end{aligned}$$



Example 2: The chances of catching cold by workers working in an ice factory during winter are 25%. What is the probability that out of 5 workers 4 or more will catch cold?

Solution: Let catching cold be the success and p be the probability of success for each worker.

\therefore Here, $n = 5$, $p = 0.25$, $q = 0.75$ and by binomial distribution

$$P[X=r] = nC_r p^r q^{n-r}, \quad r=0, 1, 2, \dots, n.$$

$$= 5C_r 0.25^r 0.75^{5-r}, \quad r=0, 1, 2, \dots, n.$$

Therefore, the required probability = $P[X \geq 4]$

$$= P[X=4 \text{ or } X=5]$$

$$= P[X=4] + P[X=5]$$

$$= 5C_4 0.25^4 0.75^1 + 5C_5 0.25^5 0.75^0$$

$$= (5)(0.00293) + 1(0.000977)$$

$$= 0.014650 + 0.000977$$

$$= 0.015627$$

Characteristics of the Binomial Distribution

The expression (1) is known as binomial distribution with parameters n and p . Different values of n and p identify different binomial distributions which lead to different probabilities of r -values. The mean and standard deviation of a binomial distribution are computed in a shortcut manner as follows

Mean, $\mu = np$,

Standard deviation, $\sigma = \sqrt{npq}$

Knowing the values of first two central moments $\mu_0 = 1$ and $\mu_1 = 1$, other central moments are given by

Second moment, $\mu_2 = npq$

Third moment, $\mu_3 = npq(q-p)$

Fourth moment, $\mu_4 = 3n^2p^2q^2 + npq(1-6pq)$

so that $y_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{qp}{\sqrt{npq}}$, where $\beta_1 = \frac{n^2p^2q^2(9-p)^2}{n^2p^2q^2}$

and $y_2 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{1-6pq}{npq}$, where $\beta_2 = \frac{3n^2p^2q^2 + npq(1-6pq)^2}{n^2p^2q^2}$

For a binomial distribution, variance < mean. This distribution is unimodal when np is a whole number, and mean = mode = np .

A binomial distribution satisfies both the conditions of pdf, because

$P(x = r) \geq 0$ for all $r = 0, 1, 2, \dots, n$

$$\sum_{r=0}^n p(x = r) = \sum_{r=0}^n [{}^n C_r p^r q^{n-r}] = (p + q)^n = 1$$

Fitting a Binomial Distribution

A binomial distribution can be fitted to the observed values in the data set as follows:

- (i) Find the value of p and q . If one of these is known, the other can be obtained by using the relationship $p + q = 1$.
- (ii) Expand $(p + q)^n = p^n + {}^n C_1 p^{n-1} q + {}^n C_2 p^{n-2} q^2 + \dots + {}^n C_r p^{n-r} q^r + \dots + {}^n C_n q^n$ using the concept of binomial theorem.
- (iii) Multiply each term in the expansion by the total number of frequencies, N , to obtain the expected frequency for each of the random variable value.

The following recurrence relation can be used for fitting of a binomial distribution:

$$f(r) = {}^n C_r p^r q^{n-r}$$

$$f(r+1) = {}^n C_{r+1} p^{r+1} q^{n-r-1}$$

Therefore, $\frac{f(r+1)}{f(r)} = \frac{p}{q} \frac{n-r}{r+1}$ or $f(r+1) = \frac{p}{q} \frac{n-r}{r+1} * f(r)$



Example 3: A brokerage survey reports that 30 per cent of individual investors have used a discount broker, i.e. one which does not charge the full commission. In a random sample of 9 individuals, what is the probability that

- (a) exactly two of the sampled individuals have used a discount broker?
- (b) not more than three have used a discount broker
- (c) at least three of them have used a discount broker

Solution: The probability that individual investors have used a discount broker is, $p = 0.30$, and therefore $q = 1 - p = 0.70$

- (a) Probability that exactly 2 of the 9 individual have used a discount broker is given by

$$P(x = 2) = {}^9 C_2 (0.30)^2 (0.70)^7 = \frac{9!}{(9-2)!2!} (0.30)^2 (0.70)^7$$

$$= (9 \times 8) / 2 * 0.09 * 0.082 = 0.2656$$

- b) Probability that out of 9 randomly selected individuals not more than three have used a discount broker is given by

$$P(x \leq 3) = P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3)$$

$$= {}^9 C_0 (0.30)^0 (0.70)^9 + {}^9 C_1 (0.30) (0.70)^8 + {}^9 C_2 (0.30)^2 (0.70)^7 + {}^9 C_3 (0.30)^3 (0.70)^6$$

$$= 0.040 + 9 \times 0.30 \times 0.058 + 36 \times 0.09 \times 0.082 + 84 \times 0.027 \times 0.118$$

$$= 0.040 + 0.157 + 0.266 + 0.268 = 0.731$$

- c) Probability that out of 9 randomly selected individuals at least three have a discount broker is given by

$$P(x \geq 3) = 1 - P(x < 3) = 1 - [P(x = 0) + P(x = 1) + P(x = 2)]$$

$$= 1 - [0.040 + 0.157 + 0.266] = 0.537$$



Example 4: Mr Gupta applies for a personal loan of Rs 1,50,000 from a nationalized bank to repair his house. The loan offer informed him that over the years bank has received about 2920 loan applications per year and that the probability of approval was, on average, above 0.85.

- (a) Mr Gupta wants to know the average and standard deviation of the number of loans approved per year.

(b) Suppose bank actually received 2654 loan applications per year with an approval probability of 0.82. What are the mean and standard deviation now?

Solution: (a) Assuming that approvals are independent from loan to loan, and that all loans have the same 0.85 probability of approval. Then

$$\text{Mean, } \mu = np = 2920 \times 0.85 = 2482$$

$$\text{Standard deviation, } \sigma = \sqrt{npq} = \sqrt{2920 \times 0.85 \times 0.15} = 19.295$$

$$\text{(b) Mean, } \mu = np = 2654 \times 0.82 = 2176.28$$

$$\text{Standard deviation, } \sigma = \sqrt{npq} = \sqrt{2654 \times 0.82 \times 0.18} = 19.792$$



Example 5: The incidence of occupational disease in an industry is such that the workers have 20 per cent chance of suffering from it. What is the probability that out of six workers 4 or more will come in contact of the disease?

Solution: The probability of a worker suffering from the disease is, $p = 20/100 = 1/5$. Therefore $q = 1 - p = 1 - (1/5) = 4/5$.

The probability of 4 or more, that is, 4, 5, or 6 coming in contact of the disease is given by

$$P(x \geq 4) = P(x = 4) + P(x = 5) + P(x = 6)$$

$$= {}^6C_4 \left(\frac{1}{5}\right)^4 \left(\frac{4}{5}\right)^2 + {}^6C_5 \left(\frac{1}{5}\right)^5 \left(\frac{4}{5}\right) + {}^6C_6 \left(\frac{1}{5}\right)^6$$

$$= \frac{{}^{15} \times 16}{15625} + \frac{{}^{6 \times 4}}{15625} + \frac{1}{15625} = \frac{1}{15625} (240 + 24 + 1)$$

$$= \frac{265}{15625}$$

$$= 0.01695$$

9.4 Continuous Probability Distributions

If a random variable is discrete, then it is possible to assign a specific probability to each of its value and get the probability distribution for it. The sum of all the probabilities associated with the different values of the random variable is 1. However, not all experiments result in random variables that are discrete. Continuous random variables such as height, time, weight, monetary values, length of life of a particular product, etc. can take large number of observable values corresponding to points on a line interval much like the infinite number of gains of sand on a beach. The sum of probability to each of these infinitely large values is no longer sum to 1.

Unlike discrete random variables, continuous random variables do not have probability distribution functions specifying the exact probabilities of their specified values. Instead, probability distribution is created by distributing one unit of probability along the real line, much like distributing a handful of sand along a line. The probability of measurements (e.g. gains of sand) piles up in certain places resulting into a probability distribution called probability density function. Such distribution is used to find probabilities that the random variable falls into a specified interval of values. The depth or density of the probability that varies with the random variable (x) may be described by a mathematical formula.

The probability density function for a continuous random variable x is a curve such that the area under the curve over an interval equals the probability that x falls into that interval, i.e. the probability that x is in that interval can be found by summing the probabilities in that interval. Certain characteristics of probability density function for the continuous random variable, x are follows:

- (i) The area under a continuous probability distribution is equal to 1.
- (ii) The probability $P(a \leq x \leq b)$ that random variable x value will fall into a particular interval from a to b is equal to the area under the density curve between the points (values) a and b .

Nature seems to follow a predictable pattern for many kinds of measurements. Most numerical values of a random variable are spread around the center, and greater the distance a numerical

value has from the center, the fewer numerical values have that specific value. A frequency distribution of values of random variable observed in nature which follows this pattern is approximately bell shaped. A special case of distribution of measurements is called a normal curve (or distribution).

If a population of numerical values follows a normal curve and x is the randomly selected numerical value from the population, then x is said to be normal random variable, which has a normal probability distribution.

The normal distribution also known as Gaussian distribution is due to the work of German mathematician Karl Friedrich Gauss during the early part of the 19th century. Normal distribution provides an adequate representation of a continuous phenomenon or process such as daily changes in the stock market index, frequency of arrivals of customers at a bank, frequency of telephone calls into a switch board, customer servicing times, and so on.

9.5 Normal Distribution/ Normal Probability Curve

Carefully look at the following hypothetical frequency distribution, which a teacher has obtained after examining 150 students of class IX on a Mathematics achievement test.

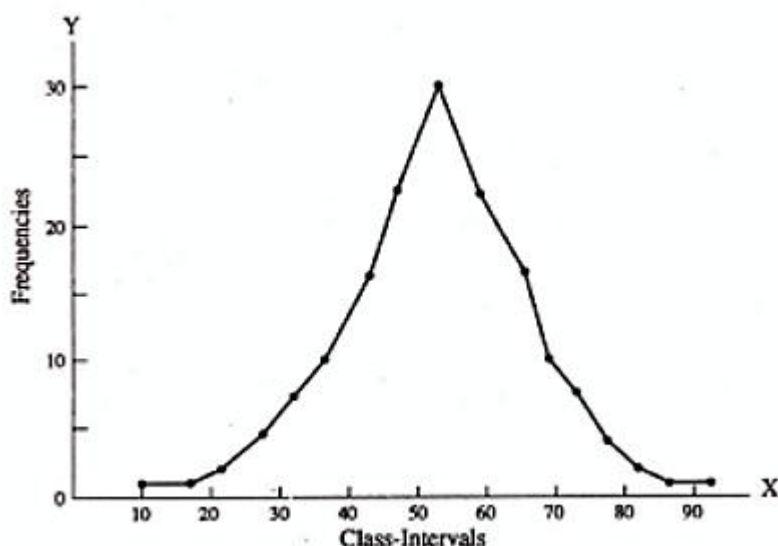
Table 9.1 Frequency distribution of the Mathematics achievement test scores

Class Intervals	Tallies	Frequency
85 – 89	I	1
80 – 84	II	2
75 – 79	IIII	4
70 – 74	IIII II	7
65 – 69	IIII IIII	10
60 – 64	IIII IIII II I	16
55 – 59	IIII IIII IIII IIII	20
50 – 54	IIII IIII IIII IIII IIII IIII	30
45 – 49	IIII IIII IIII IIII	20
40 – 44	IIII IIII IIII I	16
35 – 39	IIII IIII	10
30 – 34	IIII II	7
25 – 29	IIII	4
20 – 24	II	2
15 – 19	I	1
	Total	150

Are you able to find some special trend in the frequencies shown in the column 3 of the above table? Probably yes! The concentration of maximum frequencies ($f = 30$) lies near a central value of distribution and frequencies gradually taper off symmetrically on both the sides of this value

Concept of Normal Curve

Now, suppose if we draw a frequency polygone with the help of above distribution, we will have a curve as shown in the fig. 9.1



The shape of the curve in Fig. 9.1 is just like a 'Bell' and is symmetrical on both the sides.

If you compute the values of Mean, Median and Mode, you will find that these three are approximately the same ($M = 52$; $Md = 52$ and $Mo = 52$).

This Bell-shaped curve technically known as Normal Probability Curve or simply Normal Curve and the corresponding frequency distribution of scores, having just the same values of all three measures of central tendency (Mean, Median and Mode) is known as Normal Distribution.

Many variables in the physical (e.g. height, weight, temperature etc.) biological (e.g. age, longevity, blood sugar level and behavioural (e.g. Intelligence; Achievement; Adjustment; Anxiety; Socio-Economic-Status etc.) sciences are normally distributed in the nature. This normal curve has a great significance in mental measurement. Hence to measure such behavioural aspects, the Normal Probability Curve in simple terms Normal Curve worked as reference curve and the unit of measurement is described as σ (Sigma).

Theoretical Base of the Normal Probability Curve

The normal probability curve is based upon the law of Probability (the various games of chance) discovered by French Mathematician Abraham Demoiver (1667-1754). In the eighteenth century, he developed its mathematical equation and graphical representation also.

The law of probability and the normal curve that illust-rates it are based upon the law of chance or the probable occurrence of certain events. When any body of observations conforms to this mathematical form, it can be represented by a bell shaped curve with definite characteristics.

Characteristics of A Normal Curve

The following are the characteristics of the normal curve.

1. Normal curves are of symmetrical distribution. It means that the left half of the normal curve is a mirror image of the right half. If we were to fold the curve at its highest point at the center, we would create two equal halves.
2. The first and third quartiles of a normal distribution are equidistance from the median.
3. For the curve the mean median and mode all have the same value.
4. In skewed distribution mean median and mode fall at different points.
5. The normal curve is unimodal, having only one peak or point of maximum frequency that point in the middle of the curve.
6. The curve is a asymptotic. It means starting at the centre of the curve and working outward, the height of the curve descends gradually at first then faster and finally slower. An important situation exists at the extreme of the curve. Although the curve descends promptly toward the horizontal axis it never actually touches it. It is therefore said to be asymptotic curve.

7. In the normal curve the highest ordinate is at the centre. All ordinate on both sides of the distribution are smaller than the highest ordinate.
8. A large number of scores fall relatively close to the mean on either side. As the distance from the mean increases, the scores become fewer.
9. The normal curve involves a continuous distribution.

Normal Probability Distribution Function

The formula that generates normal probability distribution is as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{(-1/2)[(x-\mu)/\sigma]^2}$$

Where,

π = constant 3.1416

e = constant 2.7183

μ = mean of the normal distribution

σ = standard of normal distribution

The $f(x)$ values represent the relative frequencies (height of the curve) within which values of random variable x occur. The graph of a normal probability distribution with mean μ and standard deviation σ is shown in Fig. 9.2. The distribution is symmetric about its mean μ that locates at the centre.

Since the total area under the normal probability distribution is equal to 1, the symmetry implies that the area on either side of μ is 50 per cent or 0.5. The shape of the distribution is determined by μ and σ values.

In symbols, if a random variable x follows normal probability distribution with mean μ and standard deviation σ , then it is also expressed as: $x \sim N(\mu, \sigma)$.

Standard Normal Probability Distribution:

To deal with problems where the normal probability distribution is applicable more simply, it is necessary that a random variable x is standardized by expressing its value as the number of standard deviations (σ) it lies to the left or right of its mean (μ). The standardized normal random variable, z (also called z -statistic, z -score or normal variate) is defined as:

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

or equivalently $x = \mu + z\sigma$

A z -score measures the number of standard deviations that a value of the random variable x fall from the mean. From formula (1) we may conclude that

- (i) When x is less than the mean (μ), the value of z is negative
- (ii) When x is more than the mean (μ), the value of z is positive
- (iii) When $x = \mu$, the value of $z = 0$.

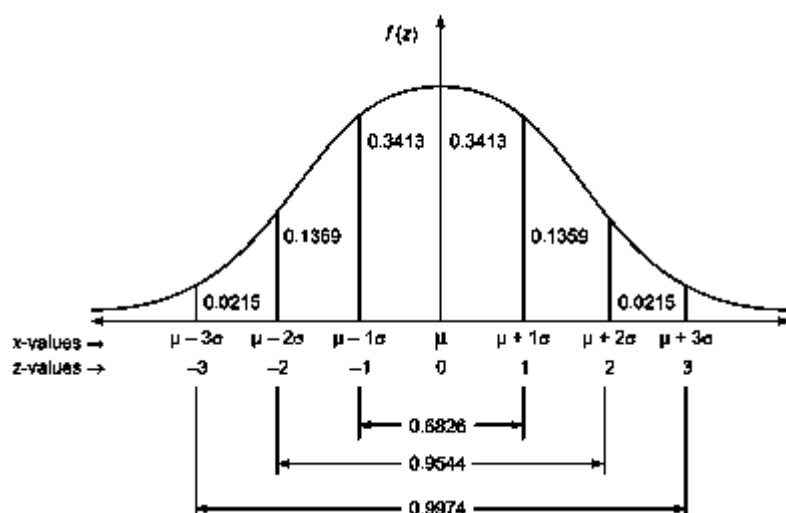


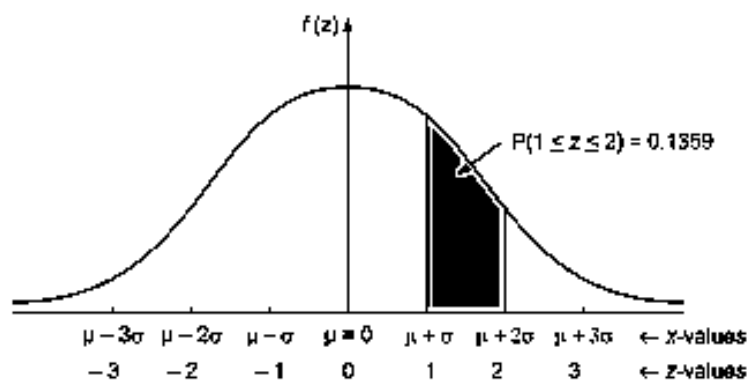
Fig. 9.2 Standard Normal Distribution

Any normal probability distribution with a set of μ and σ value with random variable can be converted into a distribution called standard normal probability distribution z , as shown in Fig. 9.2, with mean $\mu_z = 0$ and standard deviation $\sigma_z = 1$ with the help of the formula (1).

A z -value measures the distance between a particular value of random variable x and the mean (μ) in units of the standard deviation (σ). With the value of z obtained by using the formula (1), we can find the area or probability of a random variable under the normal curve by referring to the standard distribution in Appendix. For example, $z = \pm 2$ implies that the value of x is 2 standard deviations above or below the mean (μ).

Area Under the Normal Curve

Since the range of normal distribution is infinite in both the directions away from μ , the pdf function $f(x)$ is never equal to zero. As x moves away from μ , $f(x)$ approaches x -axis but never actually touches it.

Fig. 9.3: Diagram for Finding $P(1 < z < 2)$ 

The area under the standard normal distribution between the mean $z = 0$ and a specified positive value of z , say z_0 is the probability $P(0 \leq z \leq z_0)$ and can be read off directly from standard normal (z) tables. For example, area between $1 \leq z \leq 2$ is the proportion of the area under the curve which lies between the vertical lines erected at two points along the x -axis. For example, as shown in Fig. 9.3, if x is σ away from μ , that is, the distance between x and μ is one standard deviation or $(x - \mu)/\sigma$

= 1, then 34.134 per cent of the distribution lies between x and μ . Similarly, if x is at 2σ away from μ , that is, $(x - \mu)/\sigma = 2$, then the area will include 47.725 per cent of the distribution, and so on, as shown in Table 9.2.

Table 9.2: Area Under the Normal Curve

$z = x - \mu/\sigma$	Area Under Normal Curve Between x and μ
1.0	0.34134
2.0	0.47725
3.0	0.49875
4.0	0.49997

Since the normal distribution is symmetrical, Table 9.2 indicates that about 68.26 per cent of the normal distribution lies within the range $\mu - \sigma$ to $\mu + \sigma$. The other relationships derived from Table 9.2 are shown in Table 9.3

Table 9.3: Percentage of the Area of the Normal Distribution Lying within the Given Range

<i>Number of Standard Deviations from Mean</i>	<i>Approximate Percentage of Area under Normal Curve</i>
$x \pm \sigma$	68.26
$x \pm 2\sigma$	95.45
$x \pm 3\sigma$	99.75

The standard normal distribution is a symmetrical distribution and therefore

$$P(0 \leq z \leq a) = P(-a \leq z \leq 0) \text{ for any value } a.$$



For example, $P(1 \leq z \leq 2) = P(z \leq 2) - P(z \leq 1)$
 $= 0.9772 - 0.8413 = 0.1359$



Example 5: 1000 light bulbs with a mean life of 120 days are installed in a new factory and their length of life is normally distributed with standard deviation of 20 days.

- (a) How many bulbs will expire in less than 90 days?
 (b) If it is decided to replace all the bulbs together, what interval should be allowed between replacements if not more than 10% should expire before replacement?

Solution: (a) Given, $\mu = 120$, $\sigma = 20$, and $x = 90$. Then

$$\begin{aligned} z &= x - \mu/\sigma \\ &= (90-120)/20 \\ &= -1.5 \end{aligned}$$

The area under the normal curve between $z = 0$ and $z = -1.5$ is 0.4332. Therefore, area to the left of -1.5 is $0.5 - 0.4332 = 0.0668$. Thus, the expected number of bulbs to expire in less than 90 days will be $0.0668 \times 1000 = 67$ (approx.).

(b) The value of z corresponding to an area 0.4 (0.5 - 0.10). Under the normal curve is 1.28. Therefore

$$z = x - \mu/\sigma$$

$$=(x-120)/20$$

$$= 94$$

Hence, the bulbs will have to be replaced after 94 days.



Example 6: The lifetimes of certain kinds of electronic devices have a mean of 300 hours and standard deviation of 25 hours. Assuming that the distribution of these lifetimes, which are measured to the nearest hour, can be approximated closely with a normal curve.

(a) Find the probability that any one of these electronic devices will have a lifetime of more than 350 hours.

(b) What percentage will have lifetimes of 300 hours or less?

(c) What percentage will have lifetimes from 220 or 260 hours?

Solution: (a) Given, $\mu = 300$, $\sigma = 25$, and $x = 350$. Then

$$z = x - \mu/\sigma$$

$$=(350-300)/25$$

$$=2$$

The area under the normal curve between $z = 0$ and $z = 2$ is 0.9772. Thus the required probability is, $1 - 0.9772 = 0.0228$.

$$b.z = x - \mu/\sigma$$

$$=(300-300)/25$$

$$=0$$

Therefore, the required percentage is, $0.5000 \times 100 = 50\%$.

c. Given, $x_1 = 220$, $x_2 = 260$, $\mu = 300$ and $\sigma = 25$. Thus

$$Z_1 = x - \mu/\sigma$$

$$=(220-300)/25$$

$$=-3.2$$

$$Z_2 = x - \mu/\sigma$$

$$=(260-300)/25$$

$$=-1.6$$

From the normal table, we have

$$P(z = -1.6) = 0.4452 \text{ and } P(z = -3.2) = 0.4903$$

Thus, the required probability is

$$P(z = -3.2) - P(z = -1.6) = 0.4903 - 0.4452 = 0.0541$$

Hence the required percentage = $0.0541 \times 100 = 5.41$ per cent.

Summary

In this unit, we have discussed the meaning of frequency distribution and probability distribution, and the concepts of random variables and probability distribution. In any uncertain situation, we are often interested in the behaviour of certain quantities that take different values in different outcomes of experiments. These quantities are called random variables and a representation that specifies the possible values a random variable can take, together with the associated probabilities, is called a probability distribution. The distribution of a discrete variable is called a discrete probability distribution and the function that specifies a discrete distribution is termed as a probability mass function (p.m.f.). In the discrete distribution we have considered the binomial and poisson distributions and discussed how these distributions are helpful in decision-making. We have shown the fitting of such distributions to a given observed data. In the final section, we have examined situations involving continuous random variables and the resulting probability distributions. The random variable which can take an infinite number of values is called a continuous random variable and the probability distribution of such a variable is called a

continuous probability distribution. The function that specifies such distribution is called the probability density function (p.d.f.). One such important distribution, viz., the normal distribution has been presented and we have seen how probability calculations can be done for this distribution.

Keywords

1. Binomial Distribution: It is a type of discrete probability distribution function that includes an event that has only two outcomes (success or failure) and all the trials are mutually independent.
2. Continuous Probability Distribution: In this distribution the variable under consideration can take any value within a given range.
3. Discrete Probability Distribution: A probability distribution in which the variable is allowed to take on only a limited number of values.
4. Normal Distribution: It is a type of continuous probability distribution with a single peaked, bell-shaped curve. The curve is symmetrical around a vertical line erected at the mean. It is also known as Gaussian distribution.
5. Probability Distribution: A curve that shows all the values that the random variable can take and the likelihood that each will occur.

Self Assessment

1. Which of the following is correct for a binomial distribution?
 - A. variance < mean
 - B. variance = mean
 - C. variance > mean
 - D. variance \geq mean
2. 2. When the value of p is _____, then the distribution is skewed to the left.
 - A. less than 0.3
 - B. less than 0.5
 - C. more than 0.3
 - D. more than 0.5
3. 3. If $q = 0.13$ and $n = 50$ then the approximate values of mean and standard deviation are
 - A. 43; 2.38
 - B. 43; 2.36
 - C. 44; 2.38
 - D. 44; 2.36
4. 4. If the P (success) of an event is 0.4, then the P (success in at least one trial) out of five trials is
 - A. 0.6
 - B. 0.07
 - C. 0.26
 - D. 0.92
5. 5. If $p = 1/5$, $n = 5$ then the value of $P(x = 3)$ is

- A. $10 (4)^3(1/5)^2$
B. $10 (4)^2 (1/5)^5$
C. $10 (4/5)^5$
D. $10 (4/5)^3$
6. Which of the following statement is correct?
A. A large value of standard deviation reduces the height of the curve.
B. A large value of standard deviation decreases the spread of the curve.
C. A large value of standard deviation increases the height of the curve.
D. A large value of standard deviation increases the height and reduces the spread of the curve.
7. Which of the following statement is correct?
A. The maximum value of the ordinate in a mesokurtic curve is 0.2639.
B. The maximum value of the ordinate in a mesokurtic curve is 0.3989.
C. The maximum value of the ordinate in a mesokurtic curve is 0.2989.
D. The maximum value of the ordinate in a mesokurtic curve is 0.3689.
8. Which of the following condition represent positive skewness in a frequency distribution?
A. Mean < Median < Mode
B. Mean > Median > Mode
C. Mean = Median > Mode
D. Mean > Median = Mode
9. A frequency distribution is said to be platykurtic if
A. $-0.263 < Ku < 0.263$
B. $Ku < 0.263$
C. $Ku = 0.263$
D. $Ku > 0.263$
10. Normal curve has significance in the _____.
1. mental measurement
2. educational evaluation
A. only option 1
B. only option 2
C. both options 1 and 2
D. neither option 1 nor 2
11. If $N = 125$, $\mu = 10$, $X = 24$ and $\sigma = 8$, the z-value will be
A. -1.75
B. 1.70
C. 1.75
D. 1.74
12. If the total number of cases lie between 0 and +1 are 3413 and total cases lie between mean (0) and +0.5 = 1915, then the percentage of cases between +0.5 and +1 are
A. 53.28%

- B. 14.98%
- C. 13.28%
- D. 14.89%

13. If there are an odd number of categories then in the normal probability curve the middle category will be

- A. immediately on the right side of the mean
- B. immediately on the left side of the mean
- C. half on the immediate left and half on the immediate right side of the mean
- D. coincides with the mean

14. If $M = 20$, $\sigma = 4$, $Z = +0.53 \sigma$ then the value of raw score will be

- A. 2.12
- B. 22.12
- C. 20.12
- D. 21.22

15. Which of the following is not the application of normal distribution?

- A. Determine the percentile rank of a student in his group.
- B. Determine the percentile value of a student's percentile rank.
- C. Compare the two distributions in terms of overlapping.
- D. All are the application of normal distribution.

Answers for Self Assessment

- | | | | | |
|-------|-------|-------|-------|-------|
| 1. A | 2. D | 3. C | 4. D | 5. B |
| 6. A | 7. B | 8. B | 9. D | 10. C |
| 11. C | 12. B | 13. C | 14. B | 15. C |

Review Questions

1. Define binomial distribution stating its parameters, mean, and standard deviation, and give two examples where such a distribution is ideally suited.
2. What information is provided by the mean, standard deviation, and central moments of the binomial distribution?
3. The normal rate of infection of a certain disease in animals is known to be 25 per cent. In an experiment with 6 animals injected with a new vaccine it was observed that none of the animals caught the infection. Calculate the probability of the observed result.

4. The incidence of a certain disease is such that on an average 20 per cent of workers suffer from it. If 10 workers are selected at random, find the probability that (i) exactly 2 workers suffer from the disease, (ii) not more than 2 workers suffer from the disease.
Calculate the probability upto fourth decimal place.
5. A supposed coffee connoisseur claims that he can distinguish between a cup of instant coffee and a cup of percolator coffee 75 per cent of the time. It is agreed that his claim will be accepted if he correctly identifies at least 5 out of 6 cups. Find (a) his chance of having the claim accepted if he is in fact only guessing, and (b) his chance of having the claim rejected when he does have the ability he claims.
6. Normal distribution is symmetric with a single peak. Does this mean that all symmetric distributions are normal? Explain.
7. Briefly describe the characteristics of the normal probability distribution. Why does it occupy such a prominent place in statistics?
8. The income of a group of 10,000 persons was found to be normally distributed with mean = Rs 750 p.m. and standard deviation = Rs 50. Show that in this group about 95 per cent had income exceeding Rs. 668 and only 5 per cent had income exceeding Rs 832. What was the lowest income among the richest 100?
9. A aptitude test for selecting officers in a bank was conducted on 1000 candidates. The average score is 42 and the standard deviation of scores is 24. Assuming normal distribution for the scores, find:
 - (a) the number of candidates whose scores exceeds 58.
 - (b) the number of candidates whose scores lie between 30 and 66.
10. A workshop produces 2000 units of an item per day. The average weight of units is 130 kg with a standard deviation of 10 kg. Assuming normal distribution, how many units are expected to weigh less than 142 kg?



Further Readings

- Hoel, P (1962), Introduction to Mathematical Statistics, Wiley John & Sons, New York.
- Hoel, Paul G. (1971), Introduction to Probability Theory, Universal Book Stall, New Delhi.
- Olkin, I., L.J. Gleser, and C. Derman (1980), Probability Models and Applications, Macmillan Publishing, New York.

Unit 10: Estimation

CONTENTS

Objectives

Introduction

10.1 Estimation

10.2 Hypothesis Testing

10.3 Types of Estimates

10.4 Estimation of Parameter Value

10.5 Point Estimation

10.6 Properties of Point Estimator

10.7 Limitation/Drawback of Point Estimates

10.8 Interval Estimate

10.9 Interval Estimation of Population Mean

10.10 Interval Estimation for Difference of Two Means

10.11 Interval Estimation of Population Mean (σ unknown)

10.12 Interval Estimation for Population Proportion

Summary

Keywords

Self Assessment

Answers for Self Assessment

Review Questions

Further Reading

Objectives

- Differentiate between hypothesis test and estimation,
- Analyze essentials of estimation of parametric value,
- Examine properties of a point estimator,
- Apply estimation and point estimation to decide statistical inferences.
- Understand the concept of interval estimation,
- Compute interval estimation for population mean and proportion,
- Apply interval estimation in solving practical problems.

Introduction

The tenth unit endeavors to make detailed discussion on the topic estimation which includes the concept of estimation, hypothesis testing, types of estimates, estimation parameter value, point estimation, properties of point estimator, limitations of point estimates, interval estimate, interval estimation of population mean, interval estimation for difference of two means, interval estimation of population mean (σ unknown), interval estimation for population proportion.

In many situations, due to the lack of enough information, it is not easy to calculate an exact value of a population parameter (like σ , μ , p). In such situations, we make the best estimate of value from corresponding sample statistics (\bar{x} , s , and \bar{p}).

A decision-maker needs to examine the two concepts i.e., estimation and hypothesis testing, that are useful for drawing statistical inference about an unknown value of population or process parameters based upon random samples.

10.1 Estimation

It is a method to estimate the value of a population parameter from the value of the corresponding sample statistic.

In general terms, estimation uses a sample statistic as the basis for estimating the value of the corresponding population parameter.

Estimation is needed for finding the degree of the influence/effect of the treatment or how much effect the treatment has?

10.2 Hypothesis Testing

Hypothesis Testing is a claim or belief about an unknown parameter value.

A hypothesis test addresses academic question concerning the *existence* of a treatment effect.

Testing of hypothesis, in general, is needed to determine, whether or not, a treatment has an effect.

Estimation and hypothesis testing are similar in many respects, and they are complementary inferential processes.

10.3 Types of Estimates

There are two types of estimates i.e., point and confidence interval estimate, for the value of population parameter. Let us understand point and confidence interval estimate.

Point Estimate

It is the value of sample statistic, that is, used to estimate the most likely value of the unknown population.

Confidence Interval Estimate

It is the range of values that is likely to have a population parameter value with a specified level of confidence.

10.4 Estimation of Parameter Value

For estimating a parameter value, it is very much essential/important to know the following:

- A point estimate,
- The amount of possible error in the point estimate or an interval likely to contain the parameter value, and
- The statement/degree of confidence that the interval contains the parameter value. The knowledge of such information is called confidence interval or interval estimation.

10.5 Point Estimation

A sample statistic, which is, calculated using sample data to estimate the most likely value of the corresponding unknown population parameter, is termed as point estimator. The point estimate/estimation is the numerical value of the estimator.

Important Issues - Statistical Inferences

As we know that the sampling distribution of the estimator provides information about best estimator for a statistical point estimate. Therefore, before drawing any statistical inference, it is essential to resolve the following important issues:

- Select an appropriate statistic to serve as the best estimator of a population parameter, and
- The nature of the sampling distribution of the selected statistic.

As the sample statistic value varies from sample to sample, the accuracy of the given estimator also varies from sample to sample. Therefore, there is no certainty of the accuracy achieved for the sample. In practice, only one sample is selected at any given time. The average value of the estimator, over all possible samples of equal size, act as the base for judging the accuracy of an estimator. Therefore, it is better to select/choose the estimator whose average accuracy is close to the value of the population parameter that is supposed to be estimated.

10.6 Properties of Point Estimator

Three important properties of a good point estimator are -

- Unbiasedness
- Consistency
- Efficiency

Bias/Unbiasedness, Consistency, and Efficiency are also considered as the criteria for selecting an estimator.

As we know that the different sample statistics can be used as point estimators of different population parameters, therefore, the general notations are

θ = Population parameter (such as μ , σ or p) of interest being estimated

$\bar{\theta}$ = Sample Statistic (such as \bar{x} , s , or \bar{p}) or point estimator of θ

Here θ is the Greek letter and $\bar{\theta}$ is read as theta hat

Bias of Point Estimator/Unbiasedness

It is defined as the difference between the expected value of the estimator and the value of the parameter being estimated. When the estimated value of the parameter and the value of the parameter being estimated are equal, the estimator is considered unbiased.

Also, the closer the expected value of a parameter is to the value of the parameter being measured, the lesser the bias is.

Generally, it is found that the value of a statistic measured from a given sample is above or below the actual value of population parameter due to sampling error.

It is desirable that the mean of the sampling distribution of sample means taken from a population is equal to the population mean. If they are equal, then sample statistics is said to be an unbiased estimator of the population parameter.

Therefore, the sample statistic $\bar{\theta}$ is said to be an unbiased estimator of the population parameter, provided expected value or mean of the sample statistic $\bar{\theta}$ represented by $E(\bar{\theta}) = \theta$

If they are not equal, then $\bar{\theta}$ is said to be biased estimator.



Did you know?

- Error of estimation - The distance between the true and estimate value of parameter is called Error of estimation.
- Margin of error - For the unbiased estimators, difference between the point estimator and true value of parameter will be less than 1.96 standard deviation/error. It is called margin

of error. It is also defined as the value added or subtracted from a point estimate to develop an interval estimate of a population parameter.

Margin of error = $1.96 \times$ standard error (SE) of estimator

$$= 1.96 \times SD(\sigma) / \text{square root of } n$$

If σ is unknown and sample size greater than or equals to 30 the sample standard deviation s can be used to approximate $SD(\sigma)$

Consistency

A point estimator is said to be consistent if its value $\bar{\theta}$ tends to become closer to the population parameter θ as sample size increases.

For example, the standard error of sampling distribution of the mean, $\sigma_{\bar{x}} = \sigma / \sqrt{n}$, tends to become smaller as sample size n increases. Thus, the sample mean \bar{x} is a consistent estimator of the population mean μ .

Similarly, the sample proportion \bar{p} is a consistent estimator of the population proportion p becomes $\sigma_{\bar{p}} = \sigma / \sqrt{n}$.

Efficiency

It deals with the spread of the sampling distribution. It is the desirable characteristic of an unbiased estimator. If the spread (as measured by the variance) of the sampling distribution of an unbiased estimator is as small as possible then it is said to be efficient. In such situation, an individual estimate will fall close to the curve value of population parameter with high probability. The reason for the same is that there is less variation in the sampling distribution of the statistic.

For example, for a simple random sample of size n , if $\bar{\theta}_1$ and $\bar{\theta}_2$ are two unbiased point estimators of the population parameter θ , then relative efficiency of $\bar{\theta}_2$ to $\bar{\theta}_1$ is given by

$$\text{Relative efficiency} = \sigma(\bar{\theta}_1) / \sigma(\bar{\theta}_2)$$

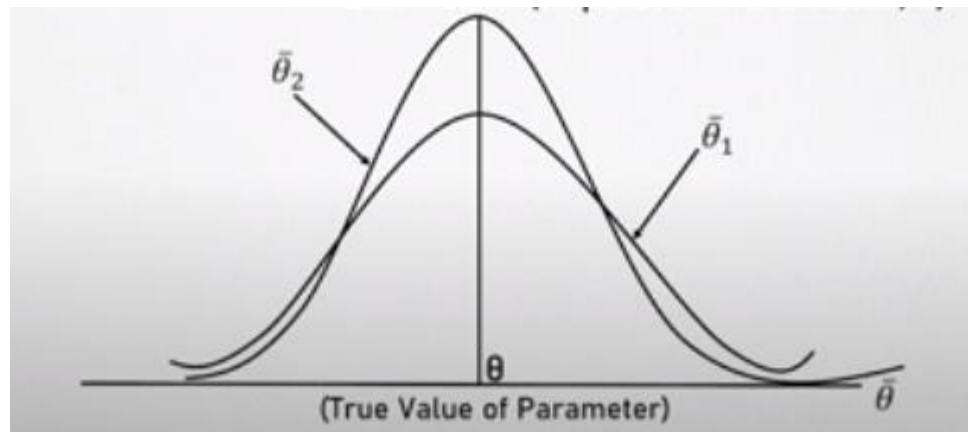


Figure 10.1 Sampling distribution of Two Unbiased Statistics of Population Parameter

Source: *Fundamentals of Business Statistics* by J. K. Sharma, Pearson Education (2014).

The figure 10.1 shows the sampling distribution of two unbiased estimators $\bar{\theta}_1$ and $\bar{\theta}_2$ for estimation of the population parameter θ .

Since standard deviation/error of statistic $\bar{\theta}_2 < \bar{\theta}_1$, the statistic $\bar{\theta}_1$ tends to produce a larger estimator error whereas the statistic $\bar{\theta}_2$ tends to produce a less estimator error as compared to statistic $\bar{\theta}_1$ both above and below the parameter θ . Therefore, value of $\bar{\theta}_2$ is more likely to provide an estimate which is more consistently closer to the true value of the parameter θ for a given sample as compared $\bar{\theta}_1$.

10.7 Limitation/Drawback of Point Estimates

The limitations or drawbacks of point estimates are as follows:

- Non-availability of the information related to the reliability of a point estimator or how close it is to its true population parameter is its limitation.
- The chance/probability of a single sample statistic originally equals the population parameter is extremely small.
- This is the reason that point estimates are rarely used alone to estimate population parameters.
- The better option is to offer range of values within which population parameters are expected to fall.
- By doing this, reliability (probability) of the estimate can be measured.
- Interval estimation is used for this purpose.

10.8 Interval Estimate

As we know that a point estimate does not provide information about how close is estimate to the population parameter. That is why, a decision maker prefers to use an interval estimate (the range of value defined around a sample statistic).

Interval estimate

The interval calculated from a sample expected to include the corresponding population parameter. An interval estimate is a rule for calculating two numerical values that contains the required population parameter. It is generally referred as a confidence coefficient $(1-\alpha)$. It is also important to state-how confident you are that the interval estimate contain parameter value.

Confidence Interval Estimate

It is the range of values that is likely to have a population parameter value with a specified level of confidence.

Hence an interval estimate of a population parameter is a confidence interval with a statement of confidence that the interval contains the parameter value.

In short, a confidence interval estimation is an interval of values computed from sample data that is likely to contain the true population parameter value.

The confidence interval estimate of a population parameter is obtained by applying the following formula:

$$\text{Point estimate} \pm \text{Margin of error}$$

The Margin of error can be calculated by using the following formula:

$$\text{Margin of error} = Z_c \times \text{Standard error of a particular statistic}$$

Where Z_c = critical value of standard normal variable that represents confidence level (like

0.90, 0.95, etc.)



Did you know?

- Confidence Interval - The interval within which the population parameter is expected to lie.

10.9 Interval Estimation of Population Mean

If true population standard deviation (σ) is known and population mean (μ) is unknown, then for a large sample size ($n \geq 30$), sample mean (\bar{x}) is the best point estimator for the Population mean (μ). Since the sampling distribution is approximately normal, it can be used to compute the confidence interval of the population mean (μ). The procedure to compute the confidence interval of the population mean (μ) is given as follows:

$$\begin{aligned} & \bar{x} \pm Z_{\alpha/2} \sigma_{\bar{x}} \\ & \bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ & \bar{x} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \end{aligned}$$

where $Z_{\alpha/2}$ = z-value representing an area in the right tail of the standard normal probability distribution

$1-\alpha$ = level of confidence

Alternate Approach

We know that $z = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}}$ with standard normal distribution.

If $Z_{\alpha/2}$ is the z-value with an area $\alpha/2$ in the right tail of normal curve, then

$$\begin{aligned} P \left[(-Z_{\alpha/2} < \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}} < +Z_{\alpha/2}) \right] &= 1-\alpha \\ -Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{x} - \mu < +Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ -\bar{x} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{x} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ \bar{x} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} > +\mu > \bar{x} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ \bar{x} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ P \left[\bar{x} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right] &= 1-\alpha \end{aligned}$$

$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ will contain population mean μ with probability $1-\alpha$.

The value of z that has tail area $\alpha/2$ to its right and left is called its critical value (represented by $+Z_{\alpha/2}$ and $-Z_{\alpha/2}$).

For example, if 99 per cent level of confidence is desired to estimate the mean then 99 percent of the area under the normal curve would be divided equally leaving an area = 49.5% between each limit and population mean.



Did you know?

- Confidence limits - The boundaries i.e., both upper and lower, of a confidence interval.
- Confidence level - The confidence associated with an interval estimate with an interval estimate. It is expressed in terms of probability that the true population parameter is included in the confidence interval.
- Sampling error - The difference between the value of an unbiased point estimator \bar{x} or \bar{p} and the value of the population parameter μ or p .



Example 1

If \bar{x} is the sample mean, $n = 100$, and $\sigma = 25$, then compute sampling error, interval estimate, margin of error, confidence coefficient, and confidence interval. Also, interpret the result.

Solution:

Given Sample mean = \bar{x} , $n = 100$, and $\sigma = 25$

$$\begin{aligned} \text{We know that } \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \\ \sigma_{\bar{x}} &= \frac{25}{\sqrt{100}} \\ \sigma_{\bar{x}} &= \frac{25}{10} \\ \sigma_{\bar{x}} &= 2.5 \end{aligned}$$

For 95%, the table value for $Z_{\alpha/2} = \pm 1.96$

$$Z_{\alpha/2}\sigma_{\bar{x}} = \pm 1.96 \times 2.5 = \pm 4.90 \text{ range}$$

95% of the sample means will be within ± 4.90 of the population mean (μ).

$$\begin{aligned} \text{Sampling Error} &= | \bar{x} - \mu | \\ &= 4.90 \text{ or less} \end{aligned}$$

$$\begin{aligned} \text{Interval estimate} &= \bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ &= \bar{x} \pm 4.90 \end{aligned}$$

Margin of error is the value '4.90' which provides an upper limit on sampling error.

The value '0.95' is termed as confidence coefficient.

95 % confidence Interval is the interval estimate given by $(\bar{x} \pm 4.90)$.

95% confidence interval estimate implies that if all possible samples of the same size were drawn then it would contain the true population mean (μ) in the interval $(\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}})$ and 5% area under the curve would not contain value of mean (μ).



Example 2

The average monthly electricity consumption for a sample of 100 families is 1250 units. Assuming the standard deviation of electric consumption of all families is 150 units, construct a 95 percent confidence interval estimate of the actual mean electric consumption.

Solution:

Given $\bar{x} = 1250$, $\sigma = 150$, $n = 100$, confidence level $(1-\alpha) = 95\%$, and $Z_{\alpha/2} = 1.96$

$$\begin{aligned} \text{We know that } \text{Confidence limits} &= \bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ \text{Confidence limits} &= 1250 \pm 1.96 \times \frac{150}{\sqrt{100}} \\ \text{Confidence limits} &= 1250 \pm 29.40 \text{ units} \\ \text{Population mean } (\mu) &= 1220.60 (=1250 - 29.40) \text{ to } 1279.4 (=1250 + 29.40) \text{ units} \end{aligned}$$

10.10 Interval Estimation for Difference of Two Means

If all possible samples of large size n_1 and n_2 are drawn from two different populations, then, the sampling distribution of $(\bar{x}_1 - \bar{x}_2)$ is approx. normal with

$$\text{Mean} = (\mu_1 - \mu_2)$$

$$\text{and Standard deviation} = \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Confidence interval limits for population mean $(\mu_1 - \mu_2)$ corresponding to desired confidence level are given by

$$= (\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha/2} \sigma_{\bar{x}_1 - \bar{x}_2}$$



Example 3

The strength of the wire produced by company A has a mean of 4500 kg and standard deviation of 200 kg. Company B has a mean of 4000 kg and a standard deviation of 300 kg. A sample of 50 wires of company A and 100 wires of company B are selected at random for testing the strength. Find 99 % confidence limits on the difference in the average strength of the population of wires produced by the two companies.

Solution

Given, For Company A, $\bar{x}_1 = 4500$, $\sigma_1 = 200$ and $n_1 = 50$

For Company B, $\bar{x}_2 = 4000$, $\sigma_2 = 300$ and $n_2 = 100$

$$\begin{aligned} \text{We Know that Standard deviation} &= \sigma_{\bar{x}_1 - \bar{x}_2} \\ &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\ &= \sqrt{\frac{(200)^2}{50} + \frac{(300)^2}{100}} \\ &= \sqrt{\frac{40000}{50} + \frac{90000}{100}} \\ &= \sqrt{\frac{80000 + 90000}{100}} \\ &= \sqrt{\frac{170000}{100}} \\ &= \sqrt{1700} \end{aligned}$$

$$\text{Standard deviation} = 41.23$$

$$\begin{aligned} \text{Now, } \bar{x}_1 - \bar{x}_2 &= 4500 - 4000 \\ &= 500, \end{aligned}$$

$$\text{and } Z_{\alpha/2} = 2.58$$

$$\text{Confidence interval limits (at 99\%)} = (\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha/2} \sigma_{\bar{x}_1 - \bar{x}_2}$$

$$\begin{aligned} \text{Confidence interval limits (at 99\%)} &= 500 \pm 2.58(41.23) \\ &= 500 \pm 106.20 \end{aligned}$$

Average strength of wire produced by two company is $500 - 106.20 \leq \mu \leq 500 + 106.20$

$$393.80 \leq \mu \leq 606.20$$

10.11 Interval Estimation of Population Mean (σ unknown)

If sample size $n \geq 30$ (large) and σ is unknown, then interval estimation of population mean can be approximated by the sample deviation (s).

Interval estimator of a population mean (μ) with confidence coefficient $(1 - \alpha)$ is given by

$$\bar{x} \pm Z_{\alpha/2} S_{\bar{x}}$$

or

$$\bar{x} \pm Z_{\alpha/2} \frac{s}{\sqrt{n}}$$

If sample size $n < 30$ (small) and σ is unknown, then interval estimation of population mean can be approximated by a t-distribution (probability distribution).

In t-distribution, more area in tails and less in centre. t-distribution depends on a parameter called degree of freedom. As degree of freedom increases, t-distribution approaches to normal distribution, as a result, sample standard deviation (s) becomes better estimate of population standard deviation (σ).

For sample size $n < 30$ (small), the interval estimator of a population mean (μ) with confidence coefficient $(1 - \alpha)$ is given by

$$\bar{x} \pm t_{\alpha/2} S_{\bar{x}}$$

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

$$\bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2} \frac{s}{\sqrt{n}}$$

Where $t_{\alpha/2}$ is the critical value of t-test statistic providing an area in the right and left tail of the t-distribution with $(n - 1)$ degree of freedom and $s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$

Table 10.1

Confidence Interval Estimation of Population Mean

Sample Size	σ	Interval Estimation
Large	Known	$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
Large	Estimated by s	$\bar{x} \pm Z_{\alpha/2} \frac{s}{\sqrt{n}}$
Small	Known	$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
Small	Estimated by s	$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$



Example 4

A random sample of 64 sales invoices was taken from a large population of sales invoices. The average value was found to be Rs. 2000 with a standard deviation of Rs. 540. Find a 95% confidence interval for the true mean value of all sales.

Solution - Given,

$$\bar{x} = 2000, s = 540, n = 64, \alpha = 5 \% \text{ and } Z_{\alpha/2} = 1.96$$

$$S_{\bar{x}} = \frac{s}{\sqrt{n}}$$

$$= \frac{540}{\sqrt{64}}$$

$$s_{\bar{x}} = \frac{540}{8}$$

$$= 67.50$$

Confidence interval of population mean (μ) is

$$\bar{x} \pm Z_{\alpha/2} \frac{s}{\sqrt{n}} = 2000 \pm 1.96 \times 64.50$$

$$= 2000 \pm 110.70$$

$$= 1889.30 \text{ (i.e., } 2000 - 110.70) \text{ to } 2110.70 \text{ (i.e., } 2000 + 110.70)$$

$$1889.30 \leq \mu \leq 2110.70$$



Example 5

The personnel department of an organization would like to estimate the family dental expenses of its employees to determine the feasibility of providing a dental insurance plan. A random sample of 10 employees reveals the family dental expenses (in thousand Rs.) in the previous year were 11, 37, 25, 62, 51, 21, 18, 43, 32 and 20. Set up a 99 % confidence interval of the average family dental expenses for the employees of this organization.

Solution- Given, $n = 10$ and $\sum x = 11 + 37 + 25 + 62 + 51 + 21 + 18 + 43 + 32 + 20 = 320$

$$\bar{x} = \frac{\sum x}{n} = \frac{320}{10} = 32$$

x	$(x - \bar{x}) = (x - 32)$	$(x - \bar{x})^2$
11	-21	441
37	5	25
25	-7	49
62	30	900
51	19	361
21	-11	121
18	-14	196
43	11	121
32	0	0
20	12	144
$\sum x =$	320	0
		$\sum (x_i - \bar{x})^2 = 2358$

Now, $\sum (x_i - \bar{x})^2 = 2359$ and $n = 10$

We know that $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$

$$s = \sqrt{\frac{2359}{10-1}}$$

$$s = \sqrt{\frac{2359}{9}}$$

$$s = 9.11$$

For $df = N-1 = 10-1 = 9$

$$\begin{aligned} \text{From table } t_{\alpha/2} &= 1.833, & \underline{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} &= 32 \pm 1.833 \times \frac{5.11}{\sqrt{10}} \\ & & &= 32 \pm 2.962 \end{aligned}$$

29.038 (i.e., $32 - 2.962$) to 34.962 (i.e., $32 + 2.962$)

$$29.038 \leq \mu \leq 34.962$$

10.12 Interval Estimation for Population Proportion

We know that $z = \sqrt{\frac{(\underline{p} - p)}{pq/n}}$

When sample size is large, then sample proportion $= \underline{p} = x/n$ is the best point estimator for the population proportion p .

The sampling distribution of sample distribution of sample proportion \underline{p} is approximately normal with mean $\mu_{\underline{p}}$ and standard error $\sqrt{pq/n}$

The confidence interval for population proportion at $(1 - \alpha)$ confidence is given by

$$\begin{aligned} & \underline{p} \pm z_{\alpha/2} \mu_{\underline{p}} \\ & \underline{p} \pm z_{\alpha/2} \sqrt{pq/n} \\ & \underline{p} - z_{\alpha/2} \sigma_{\underline{p}} \leq p \leq \underline{p} + z_{\alpha/2} \sigma_{\underline{p}} \end{aligned}$$

Where $q = 1 - p$ and $z_{\alpha/2}$ is the z-value corresponding to an area of $\alpha/2$ in the right tail of the standard normal probability distribution and the quantity $z_{\alpha/2} \sigma_{\underline{p}}$ is the margin of error (or error of the estimation). Since p and q are unknown, they are estimated using point estimators i.e., \underline{p} and \underline{q} .

Therefore, for a sample proportion, the standard error denoted by $SE(\underline{p})$ or $\sigma_{\underline{p}}$

$$\sigma_{\underline{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}$$



Example 6

Suppose we want to estimate the proportion of families in a town which have two or more children. A random sample of 144 families shows that 48 families have two or more children. Setup 95% confidence interval estimate of the population proportion of families having two or more children.

Solution -

$$\underline{p} = x/n = 48/144 = 1/3$$

$n = 144$, $\underline{p} \approx 1/3$, and $z_{\alpha/2} = 1.96$ (at 95% confidence coefficient)

$$\begin{aligned} \underline{p} \pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} &= (1/3) \pm 1.96 \sqrt{\frac{1/3(1-1/3)}{144}} \\ &= 1/3 \pm 1.96 \sqrt{\frac{1/3(2/3)}{144}} \\ &= 0.333 \pm 0.077 \end{aligned}$$

Therefore, population proportion of families who have two or more children is likely to be between 0.256 to 0.41 or 25.6 % to 41 %



Example 7

A shoe manufacturing company is producing 50,000 pair of shoes daily, from a sample of 500 pairs. 2% are found to be substandard quality. Estimate at 95% level of confidence the number of pairs of shoe that are reasonably expected to be spoiled in the daily production.

Solution -

$$\underline{p} = 0.02, \underline{q} = 1 - \underline{p} = 1 - 0.02 = 0.98$$

$$n = 500, N = 50000 \text{ and } z_{\alpha/2} = 1.96 \text{ at } 95\%$$

$$\begin{aligned} \underline{p} \pm z_{\alpha/2} \sqrt{\frac{\underline{p}(1-\underline{p})}{n} \times \frac{(N-n)}{N-1}} &= 0.02 \pm 1.96 \times \sqrt{\frac{0.02 \times 0.98}{500} \times \frac{(50000-500)}{50000-1}} \\ &= 0.02 \pm 1.96(0.0063). (0.9949) \\ &= 0.02 \pm 0.0122 \\ &= 0.0078 \text{ (i.e., } 0.02 - 0.0122) \text{ to } 0.0322 \text{ (i.e., } 0.02 + 0.0122) \end{aligned}$$

The expected number of shoes that are expected to be spoiled in the daily production

$$\begin{aligned} &= 50000 \times 0.0078 \leq p \leq 50000 \times 0.0322 \\ &= 390 \leq p \leq 1610 \end{aligned}$$

Summary

Estimation is a method to estimate the value of a population parameter from the value of the corresponding sample statistic.

Hypothesis Testing is a claim or belief about an unknown parameter value.

A hypothesis test addresses academic question concerning the existence of a treatment effect.

Estimation and hypothesis testing are similar in many respects, and they are complementary inferential processes.

There are two types of estimates i.e., point and confidence interval estimate, for the value of population parameter.

Point Estimate is the value of sample statistic, that is, used to estimate the most likely value of the unknown population.

Confidence Interval Estimate is the range of values that is likely to have a population parameter value with a specified level of confidence.

For estimating a parameter value, it is very much essential/important to know - a point estimate, the amount of possible error in the point estimate or an interval likely to contain the parameter value, and the statement/degree of confidence that the interval contains the parameter value. The knowledge of such information is called confidence interval or interval estimation.

A sample statistic, which is, calculated using sample data to estimate the most likely value of the corresponding unknown population parameter, is termed as point estimator. The point estimate/estimation is the numerical value of the estimator.

As we know that the sampling distribution of the estimator provides information about best estimator for a statistical point estimate. Therefore, before drawing any statistical inference, it is essential to resolve the following important issues - Select an appropriate statistic to serve as the best estimator of a population parameter, and the nature of the sampling distribution of the selected statistic.

Three important properties of a good point estimator are - Unbiasedness, Consistency, and Efficiency.

Bias/Unbiasedness, Consistency, and Efficiency are also considered as the criteria for selecting an estimator.

Unbiasedness is defined as the difference between the expected value of the estimator and the value of the parameter being estimated. When the estimated value of the parameter and the value of the parameter being estimated are equal, the estimator is considered unbiased.

Error of estimation - The distance between the true and estimate value of parameter is called Error of estimation.

Margin of error - For the unbiased estimators, difference between the point estimator and true value of parameter will be less than 1.96 standard deviation/error. It is called margin of error. It is also defined as the value added or subtracted from a point estimate to develop an interval estimate of a population parameter.

Margin of error = $1.96 \times \text{standard error (SE) of estimator} = 1.96 \times \text{SD}(\sigma) / \text{square root of } n$

If σ is unknown and sample size greater than or equals to 30 the sample standard deviation s can be used to approximate $\text{SD}(\sigma)$

A point estimator is said to be consistent if its value $\bar{\theta}$ tends to become closer to the population parameter θ as sample size increases.

Efficiency deals with the spread of the sampling distribution. It is the desirable characteristic of an unbiased estimator. If the spread (as measured by the variance) of the sampling distribution of an unbiased estimator is as small as possible then it is said to be efficient. In such situation, an individual estimate will fall close to the true value of population parameter with high probability. The reason for the same is that there is less variation in the sampling distribution of the statistic.

The limitations or drawbacks of point estimates are - Non-availability of the information related to the reliability of a point estimator or how close it is to its true population parameter; the chance/probability of a single sample statistic originally equals the population parameter is extremely small; and the better option is to offer range of values within which population parameters are expected to fall.

An interval estimate is a rule for calculating two numerical values that contains the required population parameter. It is generally referred as a confidence coefficient $(1-\alpha)$. It is also important to state-how confident you are that the interval estimate contain parameter value.

Confidence Interval Estimate is the range of values that is likely to have a population parameter value with a specified level of confidence.

The confidence interval estimate of a population parameter is obtained by applying the following formula - Point estimate \pm Margin of error

The Margin of error can be calculated by using the following formula -

$$\text{Margin of error} = Z_c \times \text{Standard error of a particular statistic}$$

Where Z_c = critical value of standard normal variable that represents confidence level (like

0.90, 0.95, etc.)

Interval Estimation of Population Mean (σ Known) - $\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

Interval Estimation for Difference of Two Means - $(\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha/2} \sigma_{\bar{x}_1 - \bar{x}_2}$

Interval Estimation of Population Mean (σ Unknown)

For Large Sample - $\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$

For Small Sample - $\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$

Interval Estimation For Population Proportion - $\underline{p} \pm z_{\alpha/2} \sqrt{\frac{\underline{p}(1-\underline{p})}{n}}$

Keywords

Estimation is a method to estimate the value of a population parameter from the value of the corresponding sample statistic.

Hypothesis testing is a claim or belief about an unknown parameter value.

Point Estimate is the value of sample statistic, that is, used to estimate the most likely value of the unknown population.

Confidence Interval Estimate is the range of values that is likely to have a population parameter value with a specified level of confidence.

A sample statistic, which is, calculated using sample data to estimate the most likely value of the corresponding unknown population parameter, is termed as **point estimator**. The point estimate/estimation is the numerical value of the estimator.

Bias of Point Estimator/Unbiasedness is defined as the difference between the expected value of the estimator and the value of the parameter being estimated.

Error of estimation is the distance between the true and estimate value of parameter.

Margin of error is defined as the value added or subtracted from a point estimate to develop an interval estimate of a population parameter.

A point estimator is said to be **consistent** if its value θ bar tends to become closer to the population parameter θ as sample size increases.

Efficiency deals with the spread of the sampling distribution. It is the desirable characteristic of an unbiased estimator. If the spread (as measured by the variance) of the sampling distribution of an unbiased estimator is as small as possible then it is said to be efficient. In such situation, an individual estimate will fall close to the curve value of population parameter with high probability. The reason for the same is that there is less variation in the sampling distribution of the statistic.

An **interval estimate** is a rule for calculating two numerical values that contains the required population parameter. It is generally referred as a confidence coefficient (1- α). It is also important to state-how confident you are that the interval estimate contain parameter value.

Confidence Interval Estimate is the range of values that is likely to have a population parameter value with a specified level of confidence.

Self Assessment

1. Which of the following is correct?
 - A. The estimation uses a sample statistic as the basis for estimating the value of the corresponding population parameter.
 - B. The estimation uses a population parameter as the basis for estimating the value of the corresponding sample statistic.
 - C. Estimation does not use a sample statistic as the basis for estimating the value of the corresponding population parameter.
 - D. Estimation does not use a population parameter as the basis for estimating the value of the corresponding sample statistic.

2. Point estimate and confidence interval estimate are the types of
 - A. parameters
 - B. estimates
 - C. populations
 - D. confidence intervals

3. As the sample statistic value varies from sample to sample,
 - A. the accuracy of the given estimator also varies from population to population.
 - B. the accuracy of the given estimator remains the same from sample to sample.
 - C. the accuracy of the given estimator also varies from sample to sample.
 - D. the accuracy of the given estimator remains the same from population to population.

4. Which of the following is/are the property/properties of a good point estimator?
 - A. Unbiasedness and Consistency
 - B. Consistency
 - C. Efficiency and Consistency
 - D. Unbiasedness, Consistency, and Efficiency

5. A point estimator is said to be consistent if its value $\hat{\theta}$ tends to become closer to
 - A. the population parameter θ as the sample size decreases.
 - B. the population parameter θ as the sample size increases.
 - C. the population parameter θ when there is no change in sample size.
 - D. the population parameter θ as the sample size decreases exponentially.

6. The formula to calculate margin of error is
 - A. $Z_c/\text{Standard Error}$
 - B. $Z_c - \text{Standard Error}$
 - C. $Z_c + \text{Standard Error}$
 - D. $Z_c \times \text{Standard Error}$

7. If the population mean is unknown and $\sigma_{\text{true population}}$ is given then for $n \geq 30$,
 - A. the population mean is the best point estimator for the sample mean.
 - B. the sample mean is the best point estimator for the sample.
 - C. the sample mean is the best point estimator for the population mean.
 - D. The population mean is the best point estimator for the population.

8. If $Z_{\alpha/2}$ is the z-value with an area $\alpha/2$ in the right tail of normal curve, then
 - A. $P\left[(-Z_{\alpha/2}) < \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}} < +Z_{\alpha/2}\right] =$
 - B. $P\left[(-Z_{\alpha/2}) < \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}} < +Z_{\alpha/2}\right] = 1 -$
 - C. $P\left[(-Z_{\alpha/2}) < \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}} < +Z_{\alpha/2}\right] = -1$

- D. $P [(-Z_{\alpha/2} < \frac{(\bar{x} - \mu)}{\frac{s}{\sqrt{n}}} < +Z_{\alpha/2}] = 1 -$
9. If the sample size is large and σ is estimated by 's' then interval estimate of the population mean is
- A. $\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$
- B. $\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$
- C. $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
- D. $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
10. If $\sum(x_i - \bar{x})^2 = 441$ and $n = 10$ then the value of sample standard deviation is
- A. 7
- B. 49
- C. 44.1
- D. 6.64
11. If $Z_c = 2.59$ and Standard Error = 2.5 then margin of error is
- A. 6.475
- B. 5.09
- C. 0.09
- D. 1.036
12. Error of estimation is the distance between the true and estimate value of parameter.
- A. the similarity between the true and estimate value of parameter.
- B. the similarity between the true and estimate value of statistics.
- C. the distance between the true and estimate value of statistics.
- D. the distance between the true and estimate value of parameter.
13. The formula $\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$ is related to
- A. Interval Estimation of Population Mean (σ Known)
- B. Interval Estimation of Population Mean (σ Unknown) for Large Sample
- C. Interval Estimation for Difference of Two Means
- D. For Small Sample
14. A random sample of 100 sales invoices was taken from a large population of sales invoices. The average value was found to be Rs. 1500 with a standard deviation of Rs. 136. The value of $s_{\bar{x}}$ is
- A. 0.136
- B. 1.36
- C. 1360
- D. 13.6

15. Suppose we want to estimate the proportion of families in a town which have two or more children. A random sample of 144 families shows that 48 families have two or more children. The confidence limits are
- 0.223 to 0.434
 - 0.223 to 0.443
 - 0.101 to 0.333
 - 0.232 to 0.434

Answers for Self Assessment

- | | | | | |
|-------|-------|-------|-------|-------|
| 1. A | 2. B | 3. C | 4. D | 5. B |
| 6. D | 7. C | 8. B | 9. C | 10. A |
| 11. A | 12. D | 13. B | 14. B | 15. D |

Review Questions

- Explain estimation, hypothesis testing, point estimate, confidence interval estimate, and point estimator.
- Distinguish between interval and point estimation.
- Describe the concept of unbiasedness with the help of an example.
- Discuss the concept of efficiency as an important criterion for selecting an estimator with the help of an example.
- Pen down various limitations of point estimates.
- If $\bar{x} = 124$, $n = 49$, $\sigma = 250$, and $Z_{\alpha/2} = 1.96$ then compute sampling error, interval estimate, and margin of error. Also, interpret the result.
- If $\bar{x} = 124$, $n = 49$, $\sigma = 250$, and $Z_{\alpha/2} = 1.96$ then compute confidence coefficient and confidence interval. Also, interpret the result.
- The personnel department of an organization would like to estimate the family dental expenses of its employees to determine the feasibility of providing a dental insurance plan. A random sample of 10 employees reveals the family dental expenses (in thousand Rs.) in the previous year were 1, 7, 5, 6, 5, 2, 8, 4, 3 and 9. Set up a 99 % confidence interval of the average family dental expenses for the employees of this organization.
- A ball pen manufacturer makes a lot of 10,000 refills. The procedure desires some control over these lots so that no lot will contain an excessive number of defective refills. He decides to take a random sample of 400 refills for inspection from a lot of 10,000 and finds 9 defectives. Obtain a 90% confidence interval for the number of defectives in the entire lot.
- The following data have been collected for a sample from a normal population - 5, 13, 6, 11, 8, 15, 10, and 12. Compute point estimate of population mean and standard deviation. What is the confidence interval for population mean at 95% and 99% confidence interval?



Further Reading

Fundamentals of Business Statistics by J. K. Sharma, Pearson Education (2014).



Web Links

- <https://online.stat.psu.edu/stat500/lesson/5/5.2>
- <https://egyankosh.ac.in/bitstream/123456789/20543/1/Unit-5.pdf>
- <https://www.egyankosh.ac.in/bitstream/123456789/20546/1/Unit-7.pdf>
- <https://corporatefinanceinstitute.com/resources/knowledge/other/point-estimators/>

Unit 11: Hypothesis

CONTENTS

Objectives

Introduction

11.1 Hypothesis

11.2 Importance of Hypothesis

11.3 Types of Hypotheses

11.4 Types of Errors

11.5 Testing of Hypothesis

11.6 Steps/Procedure of Hypothesis Testing

11.7 Concept of Confidence Interval

Summary

Keywords

Self Assessment

Answers for Self Assessment

Review Questions

Further Reading

Objectives

- Understand and explain the meaning and definition of hypothesis.
- Describe the importance of hypothesis.
- Analyze the differences in different types of hypotheses.
- Explain types of errors in hypothesis testing.
- Describe the procedure of hypothesis testing.
- Analyze the concept of confidence interval.

Introduction

This unit endeavors to make detailed discussion on the topic hypothesis. It includes description of meaning, importance, types of hypotheses; types of errors; hypothesis testing; and concept of confidence interval.

11.1 Hypothesis

As we know, research begins with a problem. Its function is to find a solution to the problem. The researcher should explore a set of suggested solutions or suggested explanations related to the problem at hand. These tentative solutions, which may or may not be the real solutions to the problem, formulated as a proposition are called hypotheses.

In general terms, a hypothesis is a logical prediction of certain occurrences without the support of empirical confirmation or evidence. In scientific terms, it is a tentative theory or testable statement about the relationship between independent and dependent variables. (two or more variables)

Meaning

Research Methods and Design

In general, the derivation of the hypothesis is from the problem statement. It should be framed in a positive and substantive form (before data collection). Its formulation is a creative task. It involves a lot of thinking, imagination, and innovation.

As we know, the construction of a hypothesis is considered very important after formulating a research problem in the research process. A hypothesis advances as a potential solution to the problem. As we know, any scientific inquiry starts with the statement of a solvable problem.

The researcher offered a tentative solution in the form of a testable proposition. A hypothesis is often considered a tentative and testable statement of the possible relationship between two or more variables under investigation.

Definitions

According to G. A. Lundberg - A hypothesis is a tentative generalization, the validity of which remains to be tested. In the most elementary stage, a hypothesis may be any guess, hunch, imaginative idea, which becomes the basis for action or investigation.

In the words of Goode and Hatt -It is a proposition that can be put to test to determine validity.

According to Rummel and Ballaine - A hypothesis is a question put in such a way that an answer of some kind can be forthcoming.

In the words of Kerlinger - a hypothesis is a conjectural statement of the relation between two or more variables.

According to Mcguigan - a testable statement of a potential relationship between two or more variables, i.e., advance as potential solution to the problem.

These definitions lead us to conclude that a hypothesis is a tentative solution or explanation or a guess or assumption or a proposition or a statement to the problem facing the researcher, adopted on a cursory observation of known and available data, as a basis of investigation, whose validity is to be tested or verified.

11.2 Importance of Hypothesis

Extension of Knowledge

Hypothesis facilitates the extension of knowledge in an area. They provide tentative explanations of facts and phenomena. It can be tested and validated. It sensitizes the researcher to certain aspects of the situations which are relevant from the standpoint of the problem at hand.

Logical Order of Relationships

Hypothesis provide the researcher with rational statements. It consists of elements expressed in a logical order of relationships. These relationships describe or explain conditions or events, that have not been confirmed yet by facts.

The hypothesis enables the researcher to relate logically known facts to intelligent guesses about unknown conditions. It is a guide to the thinking process and the process of discovery.

Direction

Hypothesis provides direction to the research. It defines what is relevant and irrelevant. The hypothesis tells the researcher what needs to do and find out in the study. Therefore, it prevents the review of not relevant literature.

Sample Selection & Research Procedure

It provides a basis for selecting the sample and the research procedure to be used in the study.

Statistical Techniques

Hypothesis guide the researcher in selecting the correct statistical techniques required in the data analysis and to test the relationship between the variables under study.

Delimitation

It also helps to delimit the study in scope so that it does not become broad or unwieldy.

Drawing & Reporting Conclusion

Hypothesis serves as a framework for drawing conclusions and provides the basis for reporting the conclusion of the study or provides the outline for setting conclusions in a meaningful way.

11.3 Types of Hypotheses

The various types of hypotheses are listed below:

- Null hypothesis
- Alternative hypothesis
- Simple Hypothesis
- Complex Hypothesis
- Research Hypothesis
- Logical Hypothesis
- Statistical Hypothesis

Null Hypothesis

Null hypothesis is written as H_0 . It is useful tool in testing the significance of difference. If there is no true difference between two population means or zero/no relationship, then the difference found between sample means or relationship is, accidental and unimportant, that is arising out of fluctuation of sampling and by chance. It is an important component of the decision-making methods of inferential statistics. If the difference between the samples of means is found significant, then the researcher can reject the null hypothesis. It should always be specific hypothesis i.e., it should not state about or approximately a certain value.

The null hypothesis is often stated in the following way: $H_0: \mu = 0$

Alternative Hypothesis

Alternative hypothesis is symbolized as H_1 or H_a . It is the hypothesis that specifies those values that are researcher believes to hold true, and the researcher hopes that sample data will lead to acceptance of this hypothesis as true. Alternative hypothesis represents all other possibilities, and it indicates the nature of relationship.

The alternative hypothesis is stated as $H_a: \mu \neq 0$

The alternative hypothesis may be Directional/Non-Directional hypothesis.

Simple Hypothesis

It predicts relationship between two variables i.e., the dependent and the independent variable

Complex Hypothesis

A Complex hypothesis examines relationship between two or more independent variables and two or more dependent variables.

Research Hypothesis

A research hypothesis is a specific, clear prediction about the possible outcome of a scientific research study based on specific factors of the population.

Logical Hypothesis

A logical hypothesis is a planned explanation holding limited evidence.

Statistical Hypothesis

A statistical hypothesis, sometimes called confirmatory data analysis, is an assumption about a population parameter.

Directional Hypothesis

A directional hypothesis is that which specify a particular direction.

Non-Directional Hypothesis

A non-directional hypothesis is that which does not specify a particular direction.

Declarative Hypothesis

Research Methods and Design

A declarative hypothesis is always positively worded.

Question Hypothesis

A question hypothesis is that in which inquiry is done about the expected solution or question is asked about outcome instead of suggesting expected outcome.

11.4 Types of Errors

There are two types of errors i.e.

- Type - I Errors
- Type - II Errors

Type - I Errors

A Type I error (α) is the probability of rejecting a true null hypothesis.

A Type I error (α) is the probability of telling you things are wrong, given that things are correct.

In type-I error, we may reject Null hypothesis when Null hypothesis is true.

In other words, Type-I error means rejection of hypothesis which should have been accepted. Type-I error is denoted by alpha known as alpha error, also called the level of significance of test.

We commit a Type 1 error if we reject the null hypothesis when it is true. This is a false positive.

Example

You decide to get tested for COVID-19 based on mild symptoms. The error that could potentially occur:

1. The test result says you have coronavirus, but you actually don't.
2. A fire alarm that rings when there's no fire.

Type - II Errors

A Type II error (β) is the probability of failing to reject a false null hypothesis.

A Type II error (β) is the probability of telling you things are correct, given that things are wrong.

In type-II error is when we accept Null hypothesis when the Null Hypothesis is not true. In other words, Type-II error means accepting the hypothesis which should have been rejected. Type-II error is denoted by beta known as beta error.

We commit a Type 2 error if we fail to reject the null hypothesis when it is not true. This is a false negative.

**Example**

You decide to get tested for COVID-19 based on mild symptoms. The error that could potentially occur:

1. The test result says you don't have corona virus, but you actually do.
2. A fire alarm that fails to ring/sound when there is a fire.

Table 11.1

Summary of Types of Errors

H_0 is	True	False
Rejected	Type I Error False Positive (α)	Correct Decision True Positive ($1-\alpha$)

Unit 11: Hypothesis

Not Rejected	Correct Decision True Negative ($1 - \beta$)	Type II Error False Negative (β)
--------------	---	---

The probability of Type-I error is usually determined in advance. It is understood as the level of significance of testing the hypothesis.

If Type-I error is fixed at 5%, it means that there are about 5 chance in 100 that we will reject Null hypothesis when Null hypothesis is true.

We can control Type-I error just by fixing at a lower level. For instance, if we fix it at 1%, we will say that the maximum probability of committing Type-I error would only be 0.01.

But with the fixed sample size, when we try to reduce Type-I error, the probability of committing Type-II error increases. Both types of errors cannot be reduced simultaneously.

the probability of making one type error can only be reduced if we are willing to increase the probability of making the other type of error.

You must set a very high level for Type-I error in one's testing technique of a given hypothesis. Hence, in the testing of hypothesis, you/one must make all possible efforts to strike an adequate balance between Type-I and Type-II errors.

11.5 Testing of Hypothesis

When the hypothesis has been framed in the research study, it must be verified as true or false. Verifiability is one of the important conditions of a good hypothesis. It/Verification of hypothesis means testing of the truth of the hypothesis in the light of facts.

If the hypothesis agrees with the facts, it is said to be true and it may be accepted as the explanation of the facts. But if it does not agree it is said to be false. Such a false hypothesis is either totally rejected or modified.

Verification is of two types viz., Direct verification and Indirect verification.

Direct verification

Observation and Experiments

When direct observation shows that the supposed cause exists where it was thought to exist, we have a direct verification.

When a hypothesis is verified by an experiment in a laboratory it is called direct verification by experiment.

When the hypothesis is not amenable for direct verification, we must depend on indirect verification.

Indirect verification

Indirect verification is a process in which certain possible consequences are deduced from the hypothesis and they are then verified directly.

Two steps are involved in indirect verification.

- Deductive development of hypothesis: By deductive development, certain consequences are predicted and
- Finding whether the predicted consequences follow. If the predicted consequences come true, the hypothesis is said to be indirectly verified.

Verification may be done through logical methods.

Testing of a hypothesis is done by using statistical methods.

Testing is used to accept or reject an assumption or hypothesis about a random variable by using a sample from the distribution. The assumption is the null hypothesis (H_0), and it is tested against some alternative hypothesis (H_1). Statistical tests of hypothesis are applied to sample data.

11.6 Steps/Procedure of Hypothesis Testing

The steps/procedure involved in testing a hypothesis are/is

- Select a sample and collect the data.
- Convert the variables or attributes into statistical form such as mean, proportion.
- Formulate hypotheses (null/alternative hypothesis).
- Select an appropriate test for the data such as t-test, Z-test.
- Perform computations.
- Finally draw the inference of accepting or rejecting the null hypothesis.

OR

- Formulating Hypotheses and Stating Conclusions.
- Formulate hypotheses (null/alternative hypothesis).
- Set the criteria for a decision
- Level of significance or alpha level for the hypothesis test - This is represented by α which the probability is used to define the very unlikely sample outcomes, if the null hypothesis is true. In hypothesis testing, the set of potential samples is divided into those that are likely to be obtained and those that are very unlikely if the hypothesis is true.
- Critical Region- The region composed of extreme samples values that are very unlikely outcomes if the null hypothesis is true. The boundaries for the critical region are determined by the alpha level. If sample data fall in the critical region, the null hypothesis is rejected. The α -level you set affects the outcome of the research.
- Collect data and compute sample statistics
- Make a decision and write down the decision rule.
 - State the hypothesis as the alternative hypothesis H_1 .
 - The null hypothesis, H_0 , will be the opposite of H_1 and will contain an equality sign.
 - If the sample evidence supports the alternative hypothesis, the null hypothesis will be rejected and the probability of having made an incorrect decision (when in fact H_0 is true) is α , a quantity that can be manipulated to be as small as the researcher wishes.
 - If the sample does not provide sufficient evidence to support the alternative hypothesis, then conclude that the null hypothesis cannot be rejected based on your sample. In this situation, you may wish to collect more information about the phenomenon under study.

11.7 Concept of Confidence Interval

A confidence interval displays the probability that a parameter will fall between a pair of values around the mean.

Confidence intervals measure the degree of uncertainty or certainty in a sampling method. They are most often constructed using confidence levels of 95% or 99%.

47.5% ($z = \pm 1.96$)

49.5% ($z = \pm 2.58$)

For example, we can calculate Confidence intervals as $\text{Mean} + Z\alpha/2$ (SD/square root of n or \sqrt{n}).

Confidence intervals are conducted using statistical methods, such as a t-test. The confidence intervals are used to measure uncertainty in a sample variable.

For example, a researcher selects different samples randomly from the same population and computes a confidence interval for each sample to see how it may represent the true value of the population variable. The resulting datasets are all different; some intervals include the true population parameter and others do not.

Unit 11: Hypothesis

A confidence interval is a range of values, bounded above and below the statistic's mean, that likely would contain an unknown population parameter.

Confidence level refers to the percentage of probability, or certainty, that the confidence interval would contain the true population parameter when you draw a random sample many times.

Or

in the vernacular, we are 99% certain (confidence level) that most of these samples (confidence intervals) contain the true population parameter.

Confidence intervals represent the percentage of data from a given sample that falls between the upper and lower bounds. It does not mean that 99% of the data in a random sample fall between these bounds. But it means is that one can be 99% certain that the range will contain the population mean

In other words, it would be incorrect to assume that a 99% confidence interval means that 99% of the data in a random sample fall between these bounds. But it means is that one can be 99% certain that the range will contain the population mean.

Summary

Hypothesis is a tentative theory or testable statement about the relationship between independent and dependent variables. A hypothesis acts as a guide in the research process.

Hypothesis facilitates the extension of knowledge in an area. It consists of elements expressed in a logical order of relationships. It provides direction to the research. It provides a basis for selecting the sample and the research procedure. It guides the researcher in selecting the correct statistical techniques

The different types of hypotheses are null hypothesis, alternative hypothesis, simple hypothesis, complex hypothesis, research hypothesis, logical hypothesis, statistical hypothesis

In type-I error, we may reject null hypothesis when null hypothesis is true.

In type-II error is when we accept null hypothesis when the null hypothesis is not true.

Direct observation shows that the supposed cause exists where it was thought to exist, we have a direct verification.

When a hypothesis is verified by an experiment in a laboratory it is called direct verification by experiment.

Indirect verification is a process in which certain possible consequences are deduced from the hypothesis and they are then verified directly.

The procedure of hypothesis testing includes - formulate hypotheses (null/alternative hypothesis), set the criteria for a decision, level of significance or alpha level for the hypothesis test, decide critical region, Collect data and compute sample statistics, make a decision and write down the decision rule

A confidence interval displays the probability that a parameter will fall between a pair of values around the mean. They measure the degree of uncertainty or certainty in a sampling method.

Keywords

Hypothesis is a tentative theory or testable statement about the relationship between independent and dependent variables.

Null Hypothesis states that there is no true difference between two population means or zero/no relationship.

Alternative hypothesis is the hypothesis that specifies those values that a researcher believes to hold true, and the researcher hopes that sample data will lead to acceptance of this hypothesis as true.

Type I error (α) is the probability of rejecting a true null hypothesis.

Type II error (β) is the probability of failing to reject a false null hypothesis.

Research Methods and Design

Confidence level refers to the percentage of probability, or certainty, that the confidence interval would contain the true population parameter when you draw a random sample many times.

Self Assessment

1. A hypothesis is a ____ about the relationship between independent and dependent variables.
 - A. Testable statement
 - B. Tentative theory or testable statement
 - C. Tentative theory
 - D. Logic

2. Which of the following statement is correct?
 - A. Hypothesis does not facilitate the extension of knowledge in an area.
 - B. Hypothesis does not provide tentative explanations of facts and phenomena.
 - C. Hypothesis cannot be tested and validated.
 - D. Hypothesis provides the researcher with rational statements.

3. Which of the following represent an alternative hypothesis?
 1. $\mu_1 = \mu_2$
 2. $\mu_1 \neq \mu_2$
 3. $\mu_1 - \mu_2 \neq 0$
 - A. only 1
 - B. only 2
 - C. both 2 and 3
 - D. only 3

4. a _____ hypothesis examines the relationship between two or more independent and dependent variables.
 - A. A logical
 - B. A statistical
 - C. A research
 - D. A complex

5. _____ Hypothesis also called confirmatory data analysis.
 - A. A statistical
 - B. A Complex
 - C. A logical
 - D. A research

6. Type I errors are also called
 - A. Consumer's risk
 - B. β error
 - C. False alarm
 - D. Misdetection

7. In type-II error, researcher ____ the null hypothesis when the null hypothesis is ____.
- A. Reject; true
 - B. Reject; not true
 - C. Accepted; not true
 - D. Accepted; true
8. When a hypothesis is verified by an experiment in a laboratory, it is called ____.
- A. Direct verification
 - B. Indirect verification
 - C. Experiment hypothesis
 - D. Direct verification by experiment
9. A confidence interval displays
- A. Probability that a parameter will falls between a pair of values around the mean.
 - B. Probability of two values.
 - C. A parameter that falls between a pair of values around the mean.
 - D. Values of an interval as per the interest of the researcher.
10. Which of the following is correct?
- A. The confidence intervals help in measuring certainty in a sample variable.
 - B. The confidence intervals help in measure uncertainty in a sample variable.
 - C. The confidence intervals help in measuring interval for a sample variable.
 - D. The confidence intervals help in locating a sample variable in the population.
11. The formulation of hypothesis involves a lot of
- A. Thinking only
 - B. Imagination only
 - C. Innovation only
 - D. Thinking, imagination, and innovation.
12. A statistical hypothesis, sometimes called
- A. Data analysis
 - B. Confirmatory data analysis
 - C. Conclusion
 - D. Simple hypothesis
13. Verification of hypothesis is of
- A. Two types
 - B. Threetypes
 - C. Four types
 - D. Multiple types

Research Methods and Design

14. A confidence interval displays the probability that
- A parameter will fall outside a pair of values around the mean.
 - A parameter will fall between a pair of values around the median.
 - A parameter will fall between a pair of values around the mean.
 - A parameter will fall outside a pair of values around the median.
15. Which of the following is correct?
- A 99% confidence interval means that 99% of the data in a random sample fall between these bounds.
 - A 99% confidence interval means that 99% of the sample in a random sample fall between these bounds.
 - A 99% confidence interval means that 99% certain that the range will contain the population mean.
 - A 99% confidence interval means that 99% certain that the range will not contain the population mean.

Answers for Self Assessment

- | | | | | |
|-------|-------|-------|-------|-------|
| 1. B | 2. D | 3. C | 4. D | 5. A |
| 6. C | 7. C | 8. D | 9. A | 10. B |
| 11. D | 12. B | 13. A | 14. B | 15. C |

Review Questions

- Define hypothesis. Discuss the importance of hypothesis.
- Explain various types of hypotheses with the help of examples.
- Describe type -I and type -II errors with the help of an example.
- Enlist and describe the steps involved in testing a hypothesis.
- “Verification of hypothesis means testing of the truth of the hypothesis in the light of facts.” Justify.
- Elaborate the concept of confidence interval with the help of appropriate example.

Further Reading

- Fundamentals of Business Statistics by J. K. Sharma, Pearson Education.
- Research Methodology- Methods and Techniques by Gaurav Garg and C. R. Kothari, New Age International (P) Limited.
- Methodology of Educational Research by Lokesh Koul, Vikas Publishing House Pvt. Ltd.
- Essentials of Scientific Behavioral Research by R.A. Sharma, R. Lall Book Depot.

**Web Links**

- <https://www.publichealthnotes.com/hypothesis-in-research-definition-types-and-importance/>
- <https://www.weibull.com/hotwire/issue88/relbasics88.htm>
- <https://www.investopedia.com/terms/c/confidenceinterval.asp>

Unit 12: Hypothesis Testing I

CONTENTS

Objectives

Introduction

12.1 Hypothesis Testing

12.2 Procedure of Hypothesis Testing (t-test and z-test)

12.3 Hypothesis Testing – Single Population Mean with Small Samples ($n < 30$)

12.4 Hypothesis Testing - Population Mean with Small Samples -Independent Samples

12.5 Hypothesis Testing – Difference of Population Means (Related Samples)

12.6 Hypothesis Testing – Single Population Mean with Large Samples ($n > 30$)

12.7 Hypothesis Testing – Difference of Population Means (Independent Samples)

12.8 Hypothesis Testing – Difference of Population Means($n > 30$) - (Related Samples)

Summary

Keywords

Self Assessment

Answers for Self Assessment

Review Questions

Further Reading

Objectives

- Define hypothesis testing.
- Explain the procedure of hypothesis testing.
- Compute t-test and z-test statistic in different cases for population mean with small and large samples respectively.
- Interpret the result of t-test and z-test statistic.

Introduction

This unit endeavors to make detailed discussion on the topic hypothesis testing with the help of t-test and z-test. It includes the concept, procedure, and relating problems of t-test and z-test. This unit also includes the interpretation of results obtained after the application of t-test and z-test.

12.1 Hypothesis Testing

Hypothesis testing is the process that enables a decision maker to test the validity (or significance) of his claim by analyzing the difference between the value of sample statistic and the corresponding hypothesized population parameter value.

12.2 Procedure of Hypothesis Testing (t-test and z-test)

The procedure of hypothesis testing is as following:

- H_0 or H_a
 - μ = Population Mean & μ_0 = Hypothesized Mean

H_0	H_a
$\mu = \mu_0$	$\mu \neq \mu_0$
$\mu \leq \mu_0$	$\mu > \mu_0$
$\mu \geq \mu_0$	$\mu < \mu_0$

- Level of Significance (α)
 - Rejecting - H_0 when H_0 - True
 - $\alpha = 0.05$
 - $\alpha = 0.01$
 - $\alpha = 0.10$
- Critical/Rejection Region
 - If p - Low, H_0 - Rejected
 - or
 - $\text{prob}(H_0 - \text{True}) \leq \alpha$, Reject H_0
 - If p - high, H_0 - Accepted
 - or
 - $\text{prob}(H_0 - \text{True}) > \alpha$, Accept H_0
- Test of Significance /Test Statistic
 - Parametric (Interval/Ratio)
 - Non-parametric (Nominal/Ordinal)
 - One Sample/Two Samples/K Samples
 - Two/More Samples - Independent/Related
 - Measurement Scale
 - Sample Size
 - Number of Samples& Size
- Test of Significance /Test Statistic
 - One-sample Tests - Single Sample
 - Specified Population
 - Test Statistic =

(Sample Statistic - Hypothesized Population Parameter)/Standard Error of Sample Statistic

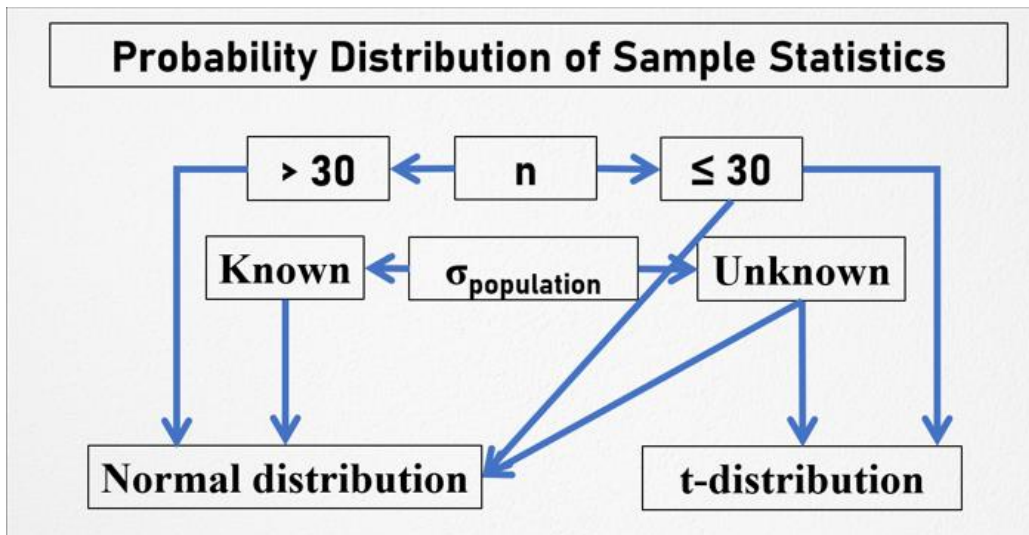


Figure 12.1

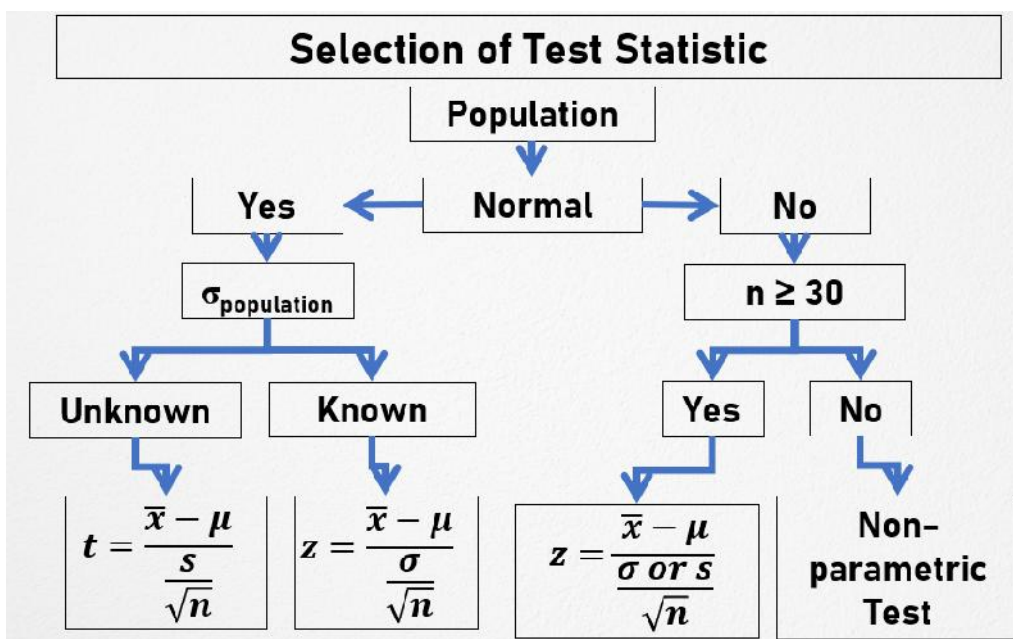


Figure 12.2

- Decision Rule
 - $t_{\text{calculated}} > t_{\text{table}}$
 - Significant & H_0 Rejected
 - $t_{\text{calculated}} \leq t_{\text{table}}$
 - Not Significant & H_0 Accepted

12.3 Hypothesis Testing - Single Population Mean with Small Samples ($n < 30$)

Example 1

The average breaking strength of steel rods is specified to be 28 thousand kg. A sample of 16 rods was tested. The mean and standard deviation were 26.5 and 4 respectively. Test the significance of the deviation.

Solution:

Given,

$$n = 16, \bar{x} = 26.5, \text{ and } s = 2,$$

$H_0: \mu = 28$ & $H_1: \mu \neq 28$ (Two-Tailed Test)

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$t = \frac{26.5 - 28}{\frac{2}{\sqrt{16}}} = \frac{-1.5}{\frac{2}{4}} = -1.5$$

$$df = n - 1 = 16 - 1 = 15$$

$\alpha = 0.05$ & 0.01 level of significance

t-value for $\alpha/2 = 0.025$ is $t_{\alpha/2} = 2.13$

$\alpha/2 = 0.005$ is $t_{\alpha/2} = 2.95$

$t_{\text{calculated}} (1.5) < t_{\text{table}} (2.13 \text{ \& } 2.95)$

Not Significant & H_0 Accepted

No significant deviation of sample mean from population mean



Example2

An automobile type manufacturer claims that the average life of a particular grade of type is more than 20,000 km when used under normal conditions. A random sample of 16 types was tested and a mean and standard deviation of 22,000 km and 5000 km, respectively were computed. Assuming the life of the types in km to be approximately normally distributed, decide whether the manufacturer's claim is valid.

Solution:

Given, $H_0: \mu \geq 20,000$ and $H_1: \mu < 20,000$ (Left-tailed test)

$H_0: \mu \geq 20,000$ & $H_1: \mu < 20,000$ (Left-tailed test)

$$n = 16, \bar{x} = 22,000, s = 5000,$$

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$t = \frac{22000 - 20000}{\frac{5000}{\sqrt{16}}} = \frac{2000}{\frac{5000}{4}} = 1.6$$

$$df = n - 1 = 16 - 1 = 15$$

$\alpha = 0.05$ & 0.01 level of significance

t-value for $\alpha = 0.05$ is $t_{\alpha} = 1.75$ & for $\alpha = 0.01$ is $t_{\alpha} = 2.60$

$t_{\text{calculated}} (1.6) < t_{\text{table}} (1.75 \text{ \& } 2.60)$

Not Significant & H_0 Accepted

Manufacturer's claim is valid.



Example3

A fertilizer mixing machine is set to give 12 kg of nitrate for every 100 kg of fertilizer. 10 bags of 100 each are examined. The percentage of nitrate so obtained is 11, 14, 13, 12, 13, 12, 13, 14, 11, and 12. Is there reason to believe that the machine is defective?

Solution:

Fertilizer mixing machine produces 12 kg of nitrate for every 100 kg of fertilizer and is not defective.

Unit 12: Hypothesis Testing I

$n = 10$, $H_0: \mu = 12$ & $H_1: \mu \neq 12$ (Two-Tailed Test)

Table 12.1

x	d = x - 12	d^2
11	-1	1
14	2	4
13	1	1
12	0	0
13	1	1
12	0	0
13	1	1
14	2	4
11	-1	1
12	0	0
$\sum x = 125$	$\sum d = 5$	$\sum d^2 = 13$

Now, $\bar{x} = \frac{\sum x}{n} = \frac{125}{10} = 12.5$

$$s_d = \sqrt{\frac{\sum d^2}{n-1} - \frac{(\sum d)^2}{n(n-1)}}$$

$$s_d = \sqrt{\frac{13}{10-1} - \frac{(5)^2}{10(10-1)}}$$

$$s_d = \sqrt{\frac{13}{9} - \frac{25}{90}}$$

$$s_d = \sqrt{1.44 - 0.28}$$

$$s_d = \sqrt{1.16}$$

$$s_d = 1.08$$

Now,

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$t = \frac{12.5 - 12}{\frac{1.08}{\sqrt{10}}}$$

$$t = \frac{0.5}{\frac{1.08}{3.16}}$$

$$t = -1.463$$

$$df = n - 1 = 10 - 1 = 9$$

$\alpha = 0.05$ & 0.01 level of significance

t-value for $\alpha/2 = 0.025$ is $t_{\alpha/2} = 2.26$ and

$\alpha/2 = 0.005$ is $t_{\alpha/2} = 3.25$

$t_{\text{calculated}} (1.466) < t_{\text{table}} (2.26 \text{ \& } 3.25)$

Not Significant & H_0 Accepted

Manufacturer's claim is valid, and machine is not defective.

12.4 Hypothesis Testing - Population Mean with Small Samples - Independent Samples



Example 4

In a test, the marks obtained by students of two groups are

Group A: 18, 20, 36, 50, 49, 36, 34, 49, and 41

Group B: 29, 28, 26, 35, 30, 44, and 46

Test the significance of difference between the mean marks secured by students of two groups.

Solution

There is no significant difference in the mean marks secured by students of two groups.

$H_0 - \mu_1 - \mu_2 = 0$ i.e., $\mu_1 = \mu_2$ & $H_1: \mu_1 \neq \mu_2$ (Two-Tailed Test)

Table 12.2

X_A	$d_A = \bar{X}_A - X_A$	d_A^2	X_B	$d_B = \bar{X}_B - X_B$	d_B^2
18	-19	361	29	-5	25
20	-17	289	28	-6	36
36	-1	1	26	-8	64
50	13	169	35	1	1
49	12	144	30	-4	16
36	-1	1	44	10	100
34	-3	9	46	12	144
49	12	144	-	-	-

Unit 12: Hypothesis Testing I

41	4	16	-	-	-
$\frac{41}{\Sigma x_A} = 333$	$\frac{4}{\Sigma d_A} = 0$	$\frac{16}{\Sigma d_A^2} = 1234$	$\frac{-}{\Sigma x_B} = 238$	$\frac{-}{\Sigma d_B} = 0$	$\frac{-}{\Sigma d_B^2} = 386$

$$\text{Now, } \bar{x}_A = \frac{\Sigma x_A}{n_A} = \frac{333}{9} = 37 \text{ \& } \bar{x}_B = \frac{\Sigma x_B}{n_B} = \frac{238}{7} = 34$$

$$s = \sqrt{\frac{\Sigma d_A^2 + \Sigma d_B^2}{n_1 + n_2 - 2}}$$

$$s = \sqrt{\frac{1234 + 386}{9 + 7 - 2}}$$

$$s = 10.76$$

$$t = \frac{\bar{x}_A - \bar{x}_B}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{37 - 34}{10.76} \sqrt{\frac{9 \times 7}{9 + 7}}$$

$$t = 0.551$$

$$df = n_A + n_B - 2 = 9 + 7 - 2 = 14$$

$\alpha = 0.05$ & 0.01 level of significance

t-value for $\alpha/2 = 0.025$ is $t_{\alpha/2} = 2.145$ & $\alpha/2 = 0.005$ is $t_{\alpha/2} = 2.977$

$t_{\text{calculated}} (0.551) < t_{\text{table}} (2.145 \text{ \& } 2.977)$

Not Significant & H_0 Accepted

There is no significant difference in the mean marks secured by students of two groups.

The students of two groups do differ significantly based on their mean marks.

12.5 Hypothesis Testing - Difference of Population Means (Related Samples)



Example 5

The HRD manager wishes to see if there has been any change in the ability of trainees after a specific training programme. The trainees take an aptitude test before the start of the programme and an equivalent one after they have completed it. The scores recorded are

Trainees	1	2	3	4	5	6	7	8	9
Scores Before Training	75	70	46	68	68	43	55	68	77
Scores after Training	70	77	57	60	79	64	55	77	76

Has any change taken place at 5 per cent significance level?

Solution:

There is no significant change that has taken place after the training.

$H_0 - \mu_d = 0$ (i.e., $\mu_1 - \mu_2 = \mu_1 = \mu_2$) & $H_1: \mu_d \neq 0$ ($\mu_1 \neq \mu_2$) (Two-Tailed Test)

Table 12.3

Trainee	Scores Before Training (X)	Scores After Training (Y)	d = X-Y	d^2
1	75	70	5	25
2	70	77	7	49
3	46	57	-11	121
4	68	60	8	64
5	68	79	-11	121
6	43	64	-21	441
7	55	55	0	0
8	68	77	-9	81
9	77	76	1	1
			$\bar{d} = -5$	$\frac{1}{\sum d^2} = 93$

$$\bar{d} = \frac{\sum d}{n} = \frac{-45}{9} = -5$$

$$s_d = \sqrt{\frac{\sum d^2}{n-1} - \frac{(\sum d)^2}{n(n-1)}}$$

$$s_d = \sqrt{\frac{903}{9-1} - \frac{(-45)^2}{9(9-1)}}$$

$$s_d = \sqrt{\frac{903}{8} - \frac{2025}{72}}$$

$$s_d = \sqrt{112.88} = 28.13$$

$$s_d = \sqrt{84.75}$$

$$s_d = 9.21$$

$$t = \frac{\bar{d} - \mu}{\frac{s_d}{\sqrt{n}}}$$

$$t = \frac{-5 - 0}{\frac{9.21}{\sqrt{9}}}$$

$$t = \frac{-5}{3}$$

$$t = \frac{-5}{3.07} = -1.63$$

$$df = n - 1 = 9 - 1 = 8$$

$\alpha = 0.05$ level of significance

t-value for $\alpha/2 = 0.025$ is $t_{\alpha/2} = 2.306$

$t_{\text{calculated}} (1.63) < t_{\text{table}} (2.306)$

Not Significant & H_0 Accepted

There is no significant change that has taken place after the training.

12.6 Hypothesis Testing - Single Population Mean with Large Samples ($n > 30$)



Example 6

The mean lifetime of a sample of 400 fluorescent light bulbs produced by a company is found to be 1600 hours with a standard deviation of 150 hours. Test the hypothesis that the mean lifetime of the bulbs produced in general is higher than the mean life of 1570 hours at $\alpha = 0.01$ level of significance.

Solution:

There is no significant change that has taken place after the training.

$H_0: \mu \leq 1570$ & $H_1: \mu > 1570$ (Right-Tailed Test)

$n = 400$, $\bar{x} = 1600$, $s = 150$ and $\alpha = 0.01$

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$z = \frac{1600 - 1570}{\frac{150}{\sqrt{400}}}$$

$$z = \frac{30}{7.5}$$

$$z = 4$$

$$df = n - 1 = 400 - 1 = 399$$

$\alpha = 0.01$ level of significance

t-value for $\alpha = 0.01$ is $t_{\alpha/2} = 2.58$

$z_{\text{calculated}} (4) > z_{\text{table}} (2.58)$

Significant & H_0 Rejected

Mean lifetime of bulbs produced by the company may be higher than 1570 hours.

12.7 Hypothesis Testing - Difference of Population Means (Independent Samples)



Example 7

A firm believes that the tyres produced by the process X on an average last longer than tyres produced by process Y. To test this belief, random samples of tyres produced by the two processes were tested and the results are

Process	Sample Size	Average Lifetime (in Km)	Standard Deviation (in Km)
---------	-------------	--------------------------	----------------------------

X	50	22400	1000
Y	50	21800	1000

Is there evidence at 5 per cent level of significance that the firm is correct in its belief?

Solution:

There is no significant in the average life of tyre produced by process X and Y.

$H_0: \mu_1 = \mu_2$ & $H_1: \mu_1 \neq \mu_2$

$\bar{x}_1 = 22400, \bar{x}_2 = 21800, \sigma_1 = \sigma_2 = 1000 \text{ Km}, n_1 = n_2 = 50$

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}}$$

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$z = \frac{22400 - 21800}{\sqrt{\frac{(1000)^2}{50} + \frac{(1000)^2}{50}}}$$

$$z = \frac{600}{\sqrt{\frac{1000000}{50} + \frac{1000000}{50}}}$$

$$z = \frac{600}{\sqrt{20000 + 20000}} = \frac{600}{40000} = \frac{600}{200}$$

$$z = 3$$

$$df = n - 1 = 50 - 1 = 49$$

$\alpha = 0.05$ level of significance

z-value for $\alpha = 0.01$ is $Z_{\alpha/2} = 2.58$

$Z_{\text{calculated}} (3) > z_{\text{table}} (1.645)$

Significant & H_0 Rejected

Tyre produced by process X last longer than those produced by process B.

12.8 Hypothesis Testing - Difference of Population Means($n > 30$) - (Related Samples)



Example8

In the first trail of a practice period, 35 students have mean score 80 and standard deviation 8 on a digit symbol learning test. On the tenth trial, the mean is 84 and standard deviation is 10. The correlation between scores on the first and tenth trails is 0.40. Test the significance difference between mean scores of two trails?

Solution -

$n_1 = n_2 = 35, M_1 = 80, M_2 = 84, \sigma_1 = 8, \sigma_2 = 10, r = 0.40$

$H_0: \mu_1 = \mu_2$ & $H_1: \mu_1 \neq \mu_2$

There is no significance difference in mean scores of two trials.

$$\sigma_D = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} - 2 \cdot r \cdot \frac{\sigma_1}{\sqrt{n_1}} \cdot \frac{\sigma_2}{\sqrt{n_2}}}$$

Unit 12: Hypothesis Testing I

$$\sigma_D = \sqrt{\frac{(8)^2}{35} + \frac{(10)^2}{35} - 2 \times 0.40 \times \frac{8}{\sqrt{35}} \times \frac{10}{\sqrt{35}}}$$

$$\sigma_D = \sqrt{1.83 + 2.86 - 2 \times 0.40 \times \frac{80}{35}}$$

$$\sigma_D = \sqrt{4.69 - 2 \times 0.40 \times 2.29}$$

$$\sigma_D = \sqrt{4.69 - 1.83}$$

$$\sigma_D = \sqrt{2.86}$$

$$\sigma_D = 1.69$$

$$z = \frac{M_1 - M_2}{\sigma_D}$$

$$z = \frac{80 - 84}{1.69}$$

$$z = \frac{4}{1.69}$$

$$z = 2.37$$

$$df = n - 1 = 35 - 1 = 34$$

$$\text{At } 0.05 - t = 1.71$$

$$\text{At } 0.01 - t = 2.49$$

$$Z_{\text{calculated}} (2.37) > Z_{\text{table}} (1.71 \text{ and } 2.49)$$

$t_{\text{cal}} > t_{\text{table}}$ at 0.05 - Significant

$t_{\text{cal}} < t_{\text{table}}$ at 0.01 - Not Significant



Example 9

Two groups of students (students of professional and academic course) are matched for mean and standard deviation on a group intelligence test. The records of two groups on a vocational ability test are:

	Professional Course	Academic Group
N	125	125
M	51.52	54.48
SD	6.25	7.24
r	0.32	

Test the significance difference in vocational ability of two groups.

Solution

$$H_0: \mu_1 = \mu_2 \text{ \& } H_1: \mu_1 \neq \mu_2$$

There is no significance difference in vocational ability of two groups.

$$n_1 = n_2 = 125, M_1 = 51.52, M_2 = 54.48, \sigma_1 = 6.25, \sigma_2 = 7.24, r = 0.32$$

$$\sigma_D = \sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \times (1 - r)^2}$$

$$\sigma_D = \sqrt{\left(\frac{(6.25)^2}{125} + \frac{(7.24)^2}{125}\right) \times (1 - 0.32)^2}$$

$$\sigma_D = \sqrt{\left(\frac{39.06}{125} + \frac{52.42}{125}\right) \times (0.68)^2}$$

$$\sigma_D = \sqrt{\left(\frac{39.06}{125} + \frac{52.42}{125}\right) \times (0.68)^2}$$

$$\sigma_D = \sqrt{\left(\frac{91.48}{125}\right) \times (0.46)} = \sqrt{0.73 \times 0.46} = \sqrt{0.34} = 0.58$$

$$\sigma_D = 0.58$$

$$z = \frac{M_1 - M_2}{\sigma_D}$$

$$z = \frac{51.52 - 54.48}{0.58}$$

$$z = \frac{-2.96}{0.58}$$

$$z = -5.10$$

$$z = -5.10$$

$$df = n_1 + n_2 - 2 = 125 + 125 - 2 = 250 - 2 = 248$$

$$\text{At } 0.05 - t = 1.97$$

$$\text{At } 0.01 - t = 2.59$$

$Z_{\text{calculated}}(5.10) > Z_{\text{table}}(1.97)$ and $Z_{\text{table}}(2.59)$

Significant and null hypothesis Rejected.

There is significant difference in vocational ability of two groups

Academic group have high vocational ability as compared to professional group.

Summary

The process that enables a decision maker to test the validity (or significance) of his claim by analyzing the difference between the value of sample statistic and the corresponding hypothesized population parameter value, is called hypothesis testing.

Procedure

- H_0 or H_a
- Level of Significance (α)
- Critical/Rejection Region
- Test of Significance / Test Statistic
- Decision Rule

Hypothesis Testing t-test (Small Samples i.e., $n < 30$)

Single Population Mean

$$\bullet \quad t = \frac{\bar{x} - \mu}{s}$$

$$\bullet \quad s_d = \sqrt{\frac{\sum d^2}{n-1} - \frac{(\sum d)^2}{n(n-1)}} \text{ and } t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Independent Samples

$$\bullet \quad s = \sqrt{\frac{\sum d_A^2 + \sum d_B^2}{n_1 + n_2 - 2}}$$

$$\bullet \quad t = \frac{\bar{x}_A - \bar{x}_B}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

Related Samples

$$\bullet \quad s_d = \sqrt{\frac{\sum d^2}{n-1} - \frac{(\sum d)^2}{n(n-1)}}$$

$$t = \frac{\bar{d} - \mu}{\frac{s_d}{\sqrt{n}}}$$

z-test Statistics (Large Sample i.e., $n > 30$)

- Hypothesis Testing

- H_0 or H_a

- $\sigma_D = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

- $\sigma_D = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} - 2r \cdot \frac{\sigma_1}{\sqrt{n_1}} \cdot \frac{\sigma_2}{\sqrt{n_2}}}$

- $\sigma_D = \sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \times (1 - r)^2}$

- $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ or $\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$ or $\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

- $z = \frac{\bar{x}_1 - \bar{x}_2}{\frac{\sigma_1}{\sqrt{n_1}} + \frac{\sigma_2}{\sqrt{n_2}}}$ or $z = \frac{M_1 - M_2}{\frac{\sigma_1}{\sqrt{n_1}} + \frac{\sigma_2}{\sqrt{n_2}}}$

- $df = n_1 + n_2 - 2$

- Interpret

Keywords

Hypothesis testing is the process that enables a decision maker to test the validity (or significance) of his claim by analyzing the difference between the value of sample statistic and the corresponding hypothesized population parameter value.

Self Assessment

1. If $\mu \leq \mu_0$ is the null hypothesis, then the corresponding alternative hypothesis is
 - A. $\mu > \mu_0$
 - B. $\mu < \mu_0$
 - C. $\mu = \mu_0$
 - D. $\mu \geq \mu_0$
2. Parametric test apply to
 - A. Interval and nominal data
 - B. Interval and ratio data
 - C. Ratio and ordinal data
 - D. Ratio and nominal data
3. If $n = 9$, $\mu = 25$, $\bar{x} = 20$, and $s = 3$ then t-statistics is equals to
 - A. 5
 - B. -5
 - C. 15
 - D. -15
4. The formula to calculate degree of freedom, in case of two small independent samples, is
 - A. $df = (N_1 + N_2) - 1$
 - B. $df = (N_1 + N_2)$
 - C. $df = (N_1 - 1) + (N_2 - 1)$
 - D. $df = N_1 + (N_2 - 2)$

5. Which of the following is the correct formula to calculate S_d in the case of the small, correlated sample?

A. $S_d = \sqrt{\frac{\sum d^2 - (\sum d)^2}{n_1 + n_2 - 2}}$

B. $S_d = \sqrt{\frac{\sum d^2 - (\sum d)^2}{n_1 + n_2 - 1}}$

C. $S_d = \frac{\sum d^2 - (\sum d)^2}{n_1 + n_2 - 2}$

D. $S_d = \sqrt{\frac{\sum d^2 - (\sum d)^2}{n_1 + n_2 - 1}}$

6. Which of the following is not the assumption of the z-test?

- A. Sample size > 30 .
- B. Distribution is not normal (mean = zero and variance = 1).
- C. Independent sample observations.
- D. Distribution is not normal (mean = zero and variance = 1).

7. If the population standard deviation σ is not known, then the value of the z-test statistic is

- A. $z = (\bar{x} - \mu) / s / \sqrt{n}$
- B. $z = (x - \mu) / \sigma / \sqrt{n}$
- C. $z = (x - \mu) / \sigma / \sqrt{n}$
- D. $z = (x - \mu) / s / \sqrt{n}$

8. If $n = 16$, $\bar{x} = 130$, $\sigma = 10$, $\mu = 128$, and z table - at $\alpha = 0.05 = 1.96$ then z_{cal} is ____ and null hypothesis is ____.

- A. Significant and rejected
- B. Significant and accepted
- C. Insignificant and rejected
- D. Insignificant and accepted

9. Which of the following is the correct formula to calculate z-test statistic for large independent samples?

A. $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

B. $z = \frac{\bar{x} - \mu}{\sigma}$

C. $z = \frac{\bar{x} - \mu}{s / \sqrt{n}}$

D. $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ OR $\frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ OR $\frac{\bar{x} - \mu}{s / \sqrt{n}}$

10. If $N_1 = 60$, $N_2 = 40$, $M_1 = 35$, $M_2 = 32$, $\sigma_1 = 6$, and $\sigma_2 = 4$, the value of z-test statistics is approximately

- A. 9.49
- B. 9.48
- C. 9.67
- D. 9.94

11. $t_{calculated} < t_{table}$, then null hypothesis

- A. Significant & H_0 accepted
- B. Not Significant & H_0 rejected

- C. Not Significant & H_0 accepted
 D. Significant & H_0 rejected
12. A random sample of 20 observations produced a sample mean of $\bar{x} = 92.4$ and $s = 25.8$.
 What is the value of the standard error of \bar{x} ?
- A. 5.8
 B. 4.8
 C. 8.5
 D. 8.4
13. Perform a one-sample t-test using the following statistics: $n = 5$, $\bar{x} = 3.871$, $s = 0.679$. The result for null hypothesis $\mu = 5.0$ is
- A. Accepted at the 5% level; accepted at the 1% level.
 B. Rejected at the 5% level; accepted at the 1% level.
 C. Accepted at the 5% level; rejected at the 1% level.
 D. Rejected at the 5% level; rejected at the 1% level.
14. If population is not normal and n is not greater than equals to 30 then use
- A. $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$
 B. $z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$
 C. Non-parametric test
 D. Parametric test
15. If population is normal and $\sigma_{\text{population}}$ is known, the $z =$
- A. $z = \frac{\bar{x} + \mu}{\frac{\sigma}{\sqrt{n}}}$
 B. $z = \frac{\mu - \bar{x}}{\frac{\sigma}{\sqrt{n}}}$
 C. $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$
 D. $z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

Answers for Self Assessment

1. A 2. B 3. B 4. C 5. D
 6. B 7. D 8. D 9. D 10. A
 11. C 12. A 13. B 14. C 15. C

Review Questions

- Explain the procedure of hypothesis testing in different case of small sample.
- Discuss the procedure of hypothesis testing in different case of large sample.
- Apply t-test on the following data and interpret the result:

Group X	31	34	38	35
Group Y	38	37	33	35

4. Apply t-test on the following data and interpret the result:

Group A	8	2	6	3
Group B	5	6	2	9

5. Two sections of 5 students each got the following scores on an achievement test in science:

Section X1: 6 9 7 4 10

Section X2: 15 11 4 9 5

Compute the mean for both the sections and test the significance of the difference between the two means. Also interpret the result.

6. A teacher of statistics gave a test in statistics to 50 students of his class. Then she induced a state of anxiety among them, and the achievement test was re-administered. The mean and standard deviation for pre-test is 45 and 5 and for post-test is 55 and 4. The value of coefficient of correlation between pre and post-test is 0.65. Find out the following:

i) Is there a significant difference between the two set of scores?

ii) Test the hypothesis that the population mean on the post-test is significantly lower than the population mean on the pre-test.

7. The mean produce of wheat of a sample of 100 fields is 200 lbs. per acre with a standard deviation of 10 lbs. Another sample of 150 fields gives the mean of 220 lbs. with a standard deviation of 12 lbs. Can the two samples be considered to have been taken from the same population whose standard deviation is 11 lbs? Use 1% and 5% per cent level of significance.
8. The mean and standard deviation of 100 boys and 144 girls in an achievement test are 150, 60 and 10, 12 respectively. Compute the significance of difference between means of boys and girls and interpret the result.
9. An achievement test is administered at the time of admission and same is also administered at the end of session. The data is given below:

Pre-test - Mean = 45, Standard deviation = 6 and N = 64

Post-test - Mean = 50, Standard deviation = 5 and N = 64

Coefficient of correlation = 0.75

Is the gain in post-test significant?

10. Apply t-test on the following data and interpret the result:

Sample - I Mean = 62, Standard deviation = 9.7, and N = 35

Sample - II Mean = 57, Standard deviation = 6.8, and N = 32



Further Reading

- Fundamentals of Business Statistics by J. K. Sharma, Pearson Education.
- Research Methodology- Methods and Techniques by Gaurav Garg and C. R. Kothari, New Age International (P) Limited.
- Methodology of Educational Research by Lokesh Koul, Vikas Publishing House Pvt. Ltd.
- Essentials of Scientific Behavioural Research by R.A. Sharma, R. Lall Book Depot.

**Web Links**

- <https://keydifferences.com/difference-between-t-test-and-z-test.html>

Unit 13: Hypothesis Testing II

CONTENTS

Objectives

Introduction

13.1 Chi-square Test

13.2 Properties of Chi-square Test

13.3 Conditions For Chi-square Test

13.4 Applications of Chi-Square Test

13.5 Chi-Square Test of Independence

13.6 Test of Goodness-of-fit

13.7 Yate's Correction for Continuity

13.8 Test For Population Variance

13.9 Test For Homogeneity

Summary

Keywords

Self Assessment

Answers for Self Assessment

Review Questions

Further Reading

Objectives

- Enlist various properties of χ^2 -test statistic.
- Explore conditions for application of χ^2 -test statistic.
- Compute χ^2 -test statistic in different cases.
- Interpret the result of χ^2 -test statistic.

Introduction

This unit endeavors to make detailed discussion on the topic hypothesis testing with the help of χ^2 -test statistic. It includes the concept, properties, conditions, and problems of χ^2 -test statistic. This unit also includes the interpretation of results obtained after the application of χ^2 -test statistic.

13.1 Chi-square Test

Chi-square test is a non-parametric test. The sampling distribution of Chi-square is called Chi-square distribution. It is a test for establishing the association between two categorical variables and find out the significance of difference groups.

The probability density function of chi-square distribution is

$$y = y_0 (\chi^2)^{\frac{\nu}{2}-1} (e)^{-\frac{\chi^2}{2}}$$

Where ν = degrees of freedom (df ν)

y_0 = a constant depending on degree of freedom ν

$$e = \text{a constant} = 2.71828$$

The χ^2 -distribution is a continuous probability distribution extending from 0 to ∞ as shown. Since it is the sum of squares so its values (cannot be negative) is non-negative.

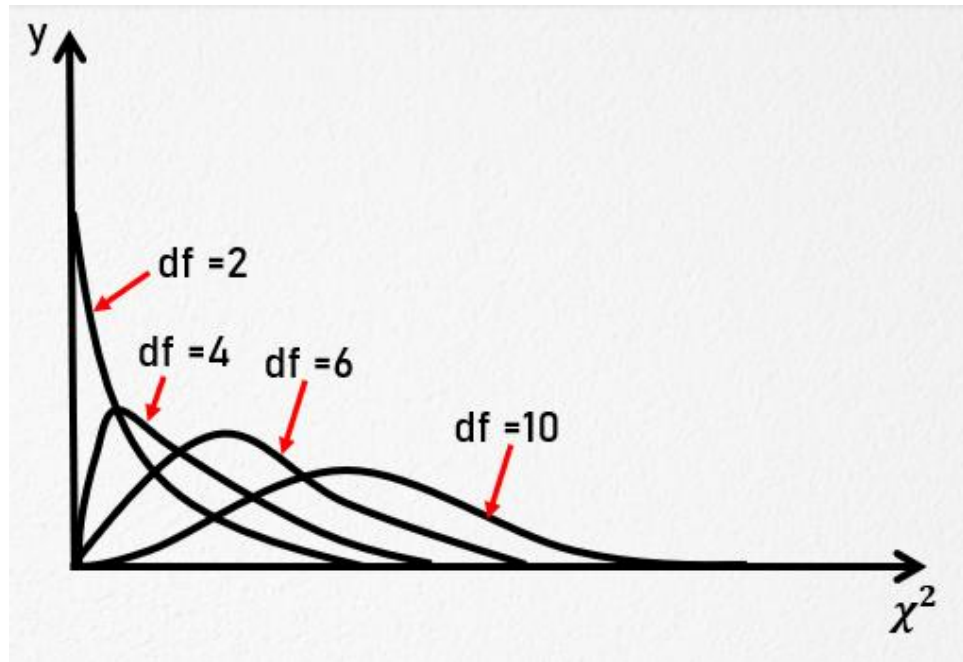


Figure 13.1

13.2 Properties of Chi-square Test

The various properties of chi-square test are discussed below:

- The shape of curve for various values of degrees of freedom as shown in figure 13.1
- For $\nu = 1$ the probability density function $y = y_0(\chi^2)^{\frac{\nu}{2}-1}(e)^{-\frac{\chi^2}{2}}$ reduces to $y = y_0(e)^{-\frac{\chi^2}{2}}$ which is the standard normal curve for positive values of the variate.
- The sampling distribution of χ^2 is a family of curves which vary with degrees of freedom.
- When $\nu = 1$, the curve is tangential to x-axis at origin, i.e., the curve attains its maximum value when

$$\left(\frac{dy}{d\chi^2}\right) = y_0[(\nu-1) - \chi^2] \chi^{\nu-2} (e)^{-\frac{\chi^2}{2}} = 0$$

$$\text{or } (\nu-1) - \chi^2 = 0$$

$$\text{or } \chi^2 = (\nu-1)$$

- When $\nu > 1$, the curve falls slowly and y approaches to 0 as χ^2 approaches to ∞ . In other words, sampling distribution of χ^2 is skewed towards higher values, that is, positively skewed.
- For $\nu = 3$, curve touches the y-axis at the origin.
- For $\nu > 4$, curve is tangential to χ^2 axis at origin.
- For degrees of freedom $\nu \geq 30$, the curve approximates to normal curve with mean ν and standard deviation $\sqrt{2\nu}$. In such a case the distribution of $\sqrt{2\nu}$ provides a better approximation to normality than χ^2 with mean $\sqrt{2\nu-1}$ and standard deviation one. This

characteristic helps to test the significance of the difference between observed and expected values of the variable.

- Since density function of χ^2 does not contain any parameter of population, χ^2 -test statistic is referred to as a non-parametric test. Thus χ^2 -distribution does not depend upon the form of the parent population.
- The mean and variance of χ^2 -distribution is- Mean, $\mu(\chi^2) = v$ and variance, $\sigma(\chi^2) = 2v$.

13.3 Conditions For Chi-square Test

The various conditions for chi-square test are discussed below:

- The experiment consists of 'n' identical but independent trials. The outcome of each trial falls into one of k categories. The observed number of outcomes in each category, written as O_1, O_2, \dots, O_n , with $O_1 + O_2 + \dots + O_n = n$ are counted.
- If there are only two cells, the expected frequency in each cell should be 5 or more. Because for observations less than 5, the value of chi-square shall be overestimated, resulting in the rejection of null hypothesis.
- For more than two cells, if more than 20 per cent of the cells have expected frequencies less than 5, then chi-square should not be applied.
- Samples must be drawn randomly from the population of interest. All the individual observations in a sample should be independent.
- The sample should contain at least 50 observations.
- The data should be expressed in original units, rather than in percentage or ratio form. Such precaution helps in comparison of attributes of interest.

13.4 Applications of Chi-Square Test

The various application of Chi-square test is as follows:

- Test of Independence
- Test of Goodness-of-fit
- Yate's Correction for Continuity
- Test for Population Variance
- Test for Homogeneity



Did you know?

Contingency Table - When observations are classified according to two qualitative variables or attributes and arranged in a table, the display is called a contingency table. It is a cross-table for displaying the frequencies of all possible groups of two variables. The test of independence uses the contingency table format and is also referred to as a contingency table analysis (or test).

13.5 Chi-Square Test of Independence

It is used to analyze the frequencies of two qualitative variables or attributes with multiple categories to determine whether the two variables are independent.

It is used to analyze any level of measurement, but it is particularly useful in analyzing nominal data.



Example 1

A question (Whether TV shows are primarily waste of time/educational/entertaining?) was asked from a sample of randomly selected two hundred adults. Responses of the respondents were categorized by gender are

Gender	Opinion			
	<i>Waste of Time</i>	<i>Educational</i>	<i>Entertaining</i>	<i>Total</i>
<i>Male</i>	50	12	28	90
<i>Female</i>	30	28	52	110
<i>Total</i>	80	40	80	200

Is there a relationship between opinion of adults and gender?

Solution:

H₀ - There is no relationship between opinion of adults and gender.

Table 13.1

Gender	Opinion			
	<i>Waste of Time</i>	<i>Educational</i>	<i>Entertaining</i>	<i>Total</i>
<i>Male</i>	50	12	28	90
<i>Female</i>	30	28	52	110
<i>Total</i>	80	40	80	200

	<i>Waste of Time</i>	<i>Educational</i>	<i>Entertaining</i>
<i>Male</i>	$\frac{80 \times 90}{200} = 36$	$\frac{40 \times 90}{200} = 18$	$\frac{80 \times 90}{200} = 36$
<i>Female</i>	$\frac{80 \times 110}{200} = 44$	$\frac{40 \times 110}{200} = 22$	$\frac{80 \times 90}{200} = 44$

Table 13.2

f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
50	36	14	196	5.444
12	18	-6	36	2
28	36	-8	64	1.777
30	44	14	196	4.455
28	22	6	36	1.636
52	44	8	64	$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 16.766$

Now,

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 16.766$$

$$df = (c-1)(r-1)$$

$$df = (3-1)(2-1) = (2)(1) = 2$$

For $\alpha = 0.05$ and 0.01 , Table values – 5.99 & 9.21

$$16.766 > 5.99 \text{ \& } 9.21$$

Calculated value of χ^2 Significant & H_0 Rejected.

There is relationship between opinion of adults and gender.



Example 2

The mothers of 200 adolescents were asked whether they agreed/disagreed on a certain aspect of adolescent behaviour. The data collected are

	Agree	Disagree	Total
Graduate Mother	38	12	50
Non-graduate Mother	84	66	150

Is the attitude of these mothers related to their being graduates or non-graduates?

Solution:

The type of this problem is a 2×2 Contingency (Test of Independence).

H_0 - Attitude of mothers is independent to their being graduates or non-graduates.

Table 13.3

	Agree	Disagree	Total
Graduate Mother	A = 38	B = 12	A+B = 50
Non-graduate Mother	C = 84	D = 66	C+D = 150
Total	A+C = 122	B+D = 78	N = 200

$$\chi^2 = \frac{N(AD - BC)^2}{(A+B)(C+D)(A+C)(B+D)}$$

$$\chi^2 = \frac{200(38 \times 66 - 84 \times 12)^2}{(50)(150)(122)(78)}$$

$$\chi^2 = \frac{200(2508 - 1008)^2}{71370000}$$

$$\chi^2 = \frac{200(1500)^2}{71370000}$$

$$\chi^2 = \frac{200 \times 2250000}{71370000}$$

$$\chi^2 = \frac{450000000}{71370000}$$

$$\chi^2 = 6.305$$

$$df = (c-1)(r-1) = (2-1)(2-1) = (1)(1) = 1$$

For $\alpha = 0.05$ and 0.01 , Table values - 3.841 & 6.635

$$6.305 > 3.841 \text{ \& } 6.305 < 6.635$$

Calculated value of χ^2 Significant & H_0 Rejected at 0.05

Hence, attitude of mothers is dependent to their being graduates or non-graduates at 0.05 level of significance.



Did you know?

Coefficient of Contingency-

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

$$C = \sqrt{\frac{s-n}{s}}$$

$$\text{Where } s = \sum \left(\frac{O}{E}\right)^2 \text{ or } s = \sum \left(\frac{f_{ij}}{e_{ij}}\right)^2$$

Larger value of C = Greater degree of dependence/association between two attributes

If k = number of rows/columns in contingency table

If number of rows = columns, then upper limit/maximum value of $C = \sqrt{\frac{k-1}{k}}$

If number of rows \neq columns (i.e., $r = 3$ & $c = 4$) then to calculate upper limit/maximum value of

$$C = \sqrt{\frac{k-1}{k}}; \text{ use } k = \text{smaller value (i.e., } k = 3)$$



Example 3

Calculate Coefficient of Contingency for the data given in example 2.

Solution:

From example 2, we have

$$\chi^2 = 6.305 \text{ and } k = 2$$

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \text{ or } \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{6.305}{6.305 + 200}}$$

$$C = \sqrt{\frac{6.305}{6.305 + 200}}$$

$$C = \sqrt{\frac{6.305}{206.305}}$$

$$C = \sqrt{0.031}$$

$$C = 0.176$$

Which implies level of association between attitude and qualification.

$$C = \sqrt{\frac{k-1}{k}} = \sqrt{\frac{2-1}{2}} = 0.707 \text{ implies perfect dependence between attitude and qualification of mothers.}$$

13.6 Test of Goodness-of-fit

It is a statistical test of how well given data support an assumption about the distribution of a population or random variable of interest. The test determines how well an assumed distribution fits the given data.

It is a statistical test conducted to determine how closely the observed frequencies fit those predicted by a hypothesized probability distribution for population.

Problem Based on Equal Probability



Example4

A teacher is interested in trying to find out whether absenteeism is greater on day of week than on another. The sample distribution based on teacher's records for past year is

Day of Week	Monday	Tuesday	Wednesday	Thursday	Friday
No. of Absentees	54	66	48	75	57

Test whether absence is uniformly distributed over week.

Solution:

H_0 - Absence is uniformly distributed over week.

Based on equal probability, expected frequency $= \frac{300}{5} = 60$

Table 13.4

f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
54	60	-6	36	0.6
66	60	6	36	0.6
48	60	-12	144	2.4
75	60	15	225	3.75
57	60	-3	9	0.15
300	300			$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 7.5$

$$\chi^2 = 7.5$$

$$df = (c-1) \cdot (r-1) = (5-1)(2-1) = (4)(1) = 4$$

For $\alpha = 0.05$ and 0.01 ,


Table values - 9.49 and 13.3

$$7.5 < 9.49 \text{ and } 13.3$$

Calculated value of χ^2 is not Significant & H_0 - Accepted.

Hence, Absence is uniformly distributed over week.

Problem Based on Normal Distribution

 Example5

384 school teachers were classified into six categories of adjustment ranging from a high level of adjustment to a low level of adjustment as

Categories	A	B	C	D	E	F	Total
No. of School Teachers	48	61	82	91	57	45	384

Does this classification differ significantly from one expected if adjustment is supposed to be distributed normally in our population of school teachers?

Solution:

H_0 - No difference exists between observed and expected frequencies.

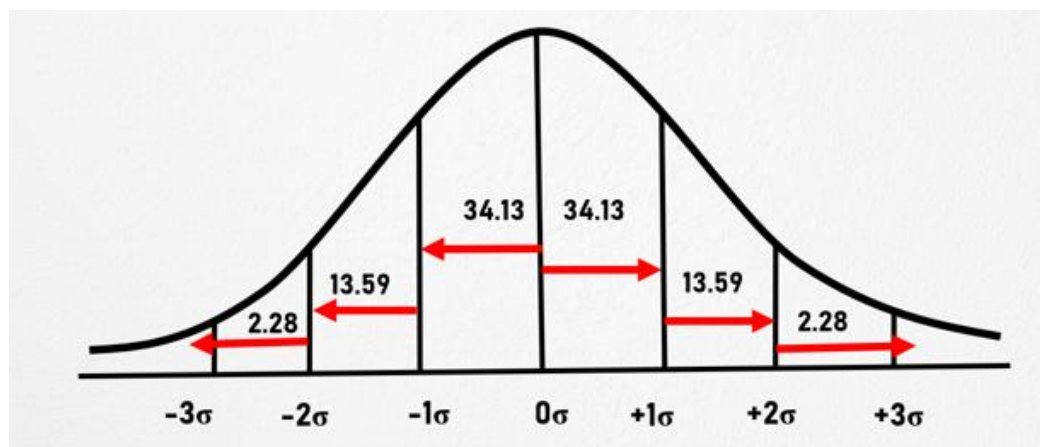


Figure 13.2

Table 13.5

Categories	Area of Normal Curve	% of Cases	No. of Cases (f_e)	f_o	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
I	+3σ to +2σ	2.28	9	48	39	1521	169
II	+2σ to +1σ	13.59	52	61	9	81	1.56
III	+1σ to 0σ	34.13	131	82	-49	2401	18.32
IV	0σ to -1σ	34.13	131	91	-40	1600	12.21
V	-1σ to -2σ	13.59	52	57	5	25	0.48
VI	-2σ to -3σ	2.28	9	45	36	1296	144
Total			384	384			$\chi^2 = 345.57$

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 345.57$$

$$df = (c-1)(r-1) = (2-1)(6-1) = (1)(5) = 5$$

For $\alpha = 0.05$ and 0.01 , Table values - 11.070 and 15.086

$$345.57 > 11.070 \text{ and } 15.086$$

Calculated value of χ^2 is Significant & H_0 Rejected

Hence, significant difference exists between observed and expected frequencies.

Problem Based on Binomial Distribution



Example 6

A survey of 800 families with 4 children each revealed following distribution

No. of Boys:	0	1	2	3	4
No. of Girls:	4	3	2	1	0
No. of Families:	32	178	290	236	64

Is this result consistent with the hypothesis that boy's and girl's birth are equally probable?

Solution:

H0-Boy and girl births are equally probable.

$$P(\text{Boy or Girl}) = p = \frac{1}{2} \text{ and } q = \frac{1}{2}$$

$$P(x=r) = {}^4C_r p^r q^{4-r}$$

$$P(x=r) = {}^4C_r \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{4-r}$$

$$P(x=r) = {}^4C_r \left(\frac{1}{2}\right)^{r+4-r}$$

$$P(x=r) = {}^4C_r \left(\frac{1}{2}\right)^4$$

Where r = 0, 1, 2, 3 and 4

Table 13.6

Category	P(x=r)	$f_e = nP(x)$	f_0	$f_0 - f_e$	$(f_0 - f_e)^2$	$\frac{(f_0 - f_e)^2}{f_e}$
0	${}^4C_0 \left(\frac{1}{2}\right)^4 = \frac{1}{16}$	$800 \times \left(\frac{1}{16}\right) = 50$	32	-18	324	6.48
1	${}^4C_1 \left(\frac{1}{2}\right)^4 = \frac{4}{16}$	$800 \times \left(\frac{4}{16}\right) = 200$	178	-22	484	2.42
2	${}^4C_2 \left(\frac{1}{2}\right)^4 = \frac{6}{16}$	$800 \times \left(\frac{6}{16}\right) = 300$	290	10	100	0.33
3	${}^4C_3 \left(\frac{1}{2}\right)^4 = \frac{4}{16}$	$800 \times \left(\frac{4}{16}\right) = 200$	236	36	1296	6.48
4	${}^4C_4 \left(\frac{1}{2}\right)^4 = \frac{1}{16}$	$800 \times \left(\frac{1}{16}\right) = 50$	64	14	196	3.92
Total						$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e} = 19.63$

$$\chi^2 = 19.63$$

$$df = (c-1)(r-1) = (2-1)(5-1) = (1)(4) = 4$$

For $\alpha = 0.05$ and 0.01

Table values - 9.488 and 13.277

$$19.63 > 9.488 \text{ and } 13.277$$

Calculated value of χ^2 – Significant & H0- Rejected

Hence, boy and girl births do not seem to be equally probable.

13.7 Yate's Correction for Continuity

A continuity correction made when calculating the chi-square test statistic for a 2 x 2 contingency table.

Expected/given frequencies < 5

$$\chi^2 = \sum \frac{(f_o - f_e - 0.5)^2}{f_e}$$

$$\chi^2 = \frac{N(|AD - BC| - \frac{N}{2})^2}{(A+B)(C+D)(A+C)(B+D)}$$



Example 7

The following information was obtained in a sample 50 small general shops:

	Shops in		
Owner	Urban Areas	Rural Areas	Total
Men	17	18	35
Women	3	12	15
Total	20	30	50

Can it be said that there are relatively more women owners of small general shops in rural than in urban areas?

Solution:

H_0 —There are an equal number of men and women owners of small general shops in both rural and urban areas.

Table 13.7

	Urban Areas	Rural Areas	Total
Men	A = 17	B = 18	A+B = 35
Women	C = 3	D = 12	C+D = 15
Total	A+C = 20	B+D = 30	N = 50

We know that

$$\chi^2 = \frac{N(|AD - BC| - \frac{N}{2})^2}{(A+B)(C+D)(A+C)(B+D)}$$

$$\chi^2 = \frac{50(|17 \times 12 - 18 \times 3| - \frac{50}{2})^2}{(35)(15)(20)(30)}$$

$$\chi^2 = \frac{50(|204 - 54| - 25)^2}{315000}$$

$$\chi^2 = \frac{50(150 - 25)^2}{315000}$$

$$\chi^2 = \frac{50(125)^2}{315000}$$

$$\chi^2 = \frac{781250}{315000}$$

$$\chi^2 = 2.48$$

$$df = (c-1)(r-1) = (2-1)(2-1) = (1)(1) = 1$$

For $\alpha = 0.05$ and 0.01 , Table values - 3.841 & 6.635

$$2.48 < 3.841 \text{ \& } 6.635$$

Calculated value of χ^2 – Not Significant & H_0 -Accepted

There are an equal number of men and women owners of small general shops in both rural and urban areas or Shops owned by men and women in both areas are equal in number.

13.8 Test For Population Variance

σ^2 = Variance of normal population

σ_0^2 = Hypothesized value of variance

$$\sigma^2 = \sigma_0^2$$



Example8

A random sample of size 20 from a population gives sample standard deviation of 9. Test the hypothesis that the population standard deviation is 12.

Solution:

Given, $n = 20$, $s = 9$, and $\sigma = 12$

H_0 : Population standard deviation (σ) is 12.

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

$$\chi^2 = \frac{(20-1)(9)^2}{(12)^2}$$

$$\chi^2 = \frac{(19)(81)}{(144)}$$

$$\chi^2 = 10.6875 = 10.69$$

$$df = n - 1 = 20 - 1 = 19$$

For $\alpha = 0.05$ and 0.01 , Table values - 30.1 and 36.2

$$10.69 < 30.1 \text{ \& } 36.2$$

Calculated value of χ^2 - Not Significant & H_0 - Accepted

Therefore, population standard deviation is 12.

13.9 Test For Homogeneity

It is useful in a case when we intend to verify whether several populations are homogeneous with respect to some characteristics of interest.

It is useful in testing a null hypothesis that several populations are homogeneous with respect to a characteristic.



Example 8

A movie producer is bringing out a new movie. To develop an advertising strategy, the producer wants to determine whether the movie will impact/appeal equal to all age groups or most to a particular age group. Based on random sample of 500 persons who attended preview of new movie, the results obtained are

Opinion	Age Group
---------	-----------

	<i>Below 20</i>	<i>20-30</i>	<i>31-50</i>	<i>Above 50</i>
<i>Liked</i>	78	146	28	48
<i>Disliked</i>	22	54	22	42
<i>Indifferent</i>	10	20	20	10

Test the significance of difference in the opinion of all age group towards new movie?

Solution:

H_0 : There is no significant difference in opinion of all age groups towards new movie.

or

H_0 : The opinion of all age groups is same about new movie.

Table 13.8

	<i>Below 20</i>	<i>20-30</i>	<i>31-50</i>	<i>Above 50</i>	<i>Total</i>
<i>Liked</i>	78	146	28	48	300
<i>Disliked</i>	22	54	22	42	140
<i>Indifferent</i>	10	20	20	10	60
<i>Total</i>	110	220	70	100	500

We know that Expected Frequency = $\frac{\text{Column Total} \times \text{Row Total}}{\text{Grand Total}}$

Table 13.9

	Below 20	20-30	31-50	Above 50
Liked	$\frac{110 \times 300}{500} = 66$	$\frac{220 \times 300}{500} = 132$	$\frac{70 \times 300}{500} = 42$	$\frac{100 \times 300}{500} = 60$
Disliked	$\frac{110 \times 140}{500} = 30.8$	$\frac{220 \times 140}{500} = 61.6$	$\frac{70 \times 140}{500} = 19.6$	$\frac{100 \times 140}{500} = 28$
Indifferent	$\frac{110 \times 60}{500} = 13.2$	$\frac{220 \times 60}{500} = 26.4$	$\frac{70 \times 60}{500} = 8.4$	$\frac{100 \times 60}{500} = 12$

Table 13.10

f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
78	66	12	144	2.182
22	30.8	-8.8	77.44	2.514
10	13.2	-3.2	10.24	0.776
146	132	14	196	1.485
54	61.6	-7.6	57.76	0.938
20	26.4	-6.4	40.96	1.552
28	42	-14	196	4.667
22	19.6	2.4	5.76	0.294
20	8.4	11.6	134.56	16.019
48	60	-12	144	2.4
42	28	14	196	7
10	12	-2	4	0.333
500	500			$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$ = 40.16

Now,

$$\chi^2 = 40.16$$

$$df = (c-1)(r-1) = (4-1)(3-1) = (3)(2) = 6$$

For $\alpha = 0.05$ and 0.01 , Table values - 12.592 and 16.812

$$40.16 > 12.592 \text{ and } 16.812$$

Calculated value of χ^2 - Significant and H_0 - Rejected

Hence, there is no significance difference in opinion of all age groups towards new movie or the opinion of all age groups is not same about new movie.

Summary

Chi-square test is a non-parametric test. It is a test for establishing the association between two categorical variables and find out the significance of difference between groups.

The various properties of chi-square test are - (i) when $\nu = 1$, the curve is tangential to x-axis at origin, (ii) when $\nu > 1$, the curve falls slowly and y approaches to 0 as χ^2 approaches to ∞ , (iii) for $\nu = 3$, curve touches the y-axis at the origin, (iv) for $\nu > 4$, curve is tangential to χ^2 axis at origin, and (v) for degrees of freedom $\nu \geq 30$, the curve approximates to normal curve with mean ν and standard deviation $\sqrt{2\nu}$.

The various conditions for chi-square test are - (i) the experiment consists of 'n' identical but independent trials, (ii) if there are only two cells, the expected frequency in each cell should be 5 or more, (iii) for more than two cell, if more than 20 per cent of the cells have expected frequencies less than 5, then chi-square should not be applied, (iv) all the individual observations in a sample should be independent, (v) the sample should contain at least 50 observations, and (vi) the data should be expressed in original units.

The various applications of Chi-square test are -test of independence, test of goodness-of-fit, yate's correction for continuity, test for population variance, and test for homogeneity.

Keywords

Chi-square test is a test for establishing the association between two categorical variables and find out the significance of difference between two groups

Test of independence is used to analyze the frequencies of two qualitative variables or attributes with multiple categories to determine whether the two variables are independent.

Test of goodness-of-fit is a statistical test of how well given data support an assumption about the distribution of a population or random variable of interest.

Yate's correction for continuity is a continuity correction made when calculating the chi-square test statistic for a 2 x 2 contingency table.

Test for homogeneity is useful in a case when we intend to verify whether several populations are homogeneous with respect to some characteristics of interest.

Self Assessment

- The probability density function of chi-square distribution is given by
 - $y = y_0(\chi^2)^{\frac{\nu}{2}-1} (e)^{-\frac{\chi^2}{2}}$
 - $y = y_0(\chi^2)^{\frac{\nu}{2}+1} (e)^{-\frac{\chi^2}{2}}$
 - $y = y_0(\chi^2)^{\frac{\nu}{2}-1} (e)^{-\frac{\chi^2}{2}}$
 - $y = y_0(\chi^2)^{\frac{\nu}{2}+1} (e)^{-\frac{\chi^2}{2}}$

- The chi-square curve approximates to normal curve for $df \geq 30$ with
 - The values of mean = 1 and standard deviation = 2ν
 - The values of mean = ν and standard deviation = $\sqrt{2\nu}$
 - The values of mean = ν and standard deviation = 1
 - The values of mean = 2ν and standard deviation = $\sqrt{2\nu}$

- Which of the following is the correct condition for the application of the chi-square test?
 - The sample should contain at least 30 observations.
 - The sample should contain at least 40 observations.
 - The sample should contain at least 50 observations.
 - The sample should contain at least 20 observations.

- A cross table for displaying the frequencies of all possible groups of two variables is known as
 - Homogeneity table
 - Yate's correction table
 - goodness-of-fit table
 - contingency table

- The correct formula to calculate expected frequencies is
 - $(\text{Row total} \times \text{Column total}) / \text{Grand total}$
 - $(\text{Row total} + \text{Column total}) / \text{Grand total}$
 - $(\text{Row total} \times \text{Grand total}) / \text{Column total}$
 - $(\text{Row total} + \text{Grand total}) / \text{Column total}$

-
6. The chi-square curve is tangential to x-axis at origin for
- A. $\nu > 4$
 - B. $\nu = 3$
 - C. $\nu > 1$
 - D. $\nu = 1$
7. The chi-square curve falls slowly and y approaches to 0 as χ^2 approaches to ∞ for
- A. $\nu = 3$
 - B. $\nu = 1$
 - C. $\nu > 1$
 - D. $\nu > 4$
8. The chi-square curve touches the y-axis at the origin for
- A. $\nu = 1$
 - B. $\nu = 3$
 - C. $\nu > 1$
 - D. $\nu > 4$
9. The chi-square curve is tangential to χ^2 axis at origin for
- A. $\nu > 4$
 - B. $\nu = 3$
 - C. $\nu > 1$
 - D. $\nu = 1$
10. If there are only two cells, the expected frequency in each cell should be 5 or more
- A. 5
 - B. 5 or more
 - C. 3 or more
 - D. 3
11. The full form of f_e in the formula of chi-square is
- A. Experimental frequency
 - B. Observed frequency
 - C. Expected frequency
 - D. Chi-frequency
12. Chi-square test is used when we have data that are expressed in
- A. Frequencies
 - B. Proportion
 - C. Percentages
 - D. Numbers

13. Chi-square test is used in
- Formulating hypothesis
 - Writing hypothesis
 - Observing hypothesis
 - Testing hypothesis

14. In the following equation,

$$y = y_0(\chi^2)^{\frac{v}{2}} - 1 (e)^{-\left(\frac{\chi^2}{2}\right)}$$

y_0 means

- Degrees of freedom
- A constant depending on degree of freedom
- Variable
- A constant depending on variable

15. Which of the following is correct?

- $C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$
- $C = \sqrt{\frac{\chi^2}{\chi^2 - n}}$
- $C = \sqrt{\frac{\chi^2 + n}{\chi^2}}$
- $C = \sqrt{\frac{\chi^2 - n}{\chi^2}}$

Answers for Self Assessment

- | | | | | |
|-------|-------|-------|-------|-------|
| 1. C | 2. B | 3. C | 4. D | 5. A |
| 6. D | 7. C | 8. B | 9. A | 10. B |
| 11. C | 12. A | 13. D | 14. B | 15. A |

Review Questions

- Explain the properties of chi-square test.
- Discuss various conditions for chi-square test.
- Explain the following: Chi-square test, Test of independence, Test of goodness-of-fit, and Test for homogeneity.
- A random sample of size 30 from a population gives sample standard deviation of 8. Test the hypothesis that the population standard deviation is 16.
- 500 instructors were classified into five categories of adjustment ranging from a high level of adjustment to a low level of adjustment as

Categories	A	B	C	D	E	Total
No. of School Teachers	78	90	132	143	57	500

Does this classification differ significantly from one expected if adjustment is supposed to be distributed normally in our population of instructors?

6. A survey of 400 families with 3 children each revealed following distribution

No. of Boys:	0	1	2	3
No. of Girls:	3	2	1	0
No. of Families:	16	89	145	150

Is this result consistent with the hypothesis that boy's and girl's birth are equally probable?

7. A teacher is interested in trying to find out whether absenteeism is greater on day of week than on another. The sample distribution based on teacher's records for past year is

Day of Week	Monday	Tuesday	Wednesday	Thursday	Friday
No. of Absentees	27	33	24	38	28

Test whether absence is uniformly distributed over week.



Further Reading

- Fundamentals of Business Statistics by J. K. Sharma, Pearson Education.
- Research Methodology- Methods and Techniques by Gaurav Garg and C. R. Kothari, New Age International (P) Limited.
- Methodology of Educational Research by Lokesh Koul, Vikas Publishing House Pvt. Ltd.
- Essentials of Scientific Behavioral Research by R.A. Sharma, R. Lall Book Depot.



Web Links

- <https://egyankosh.ac.in/bitstream/123456789/12294/1/Unit-17.pdf>

Unit 14: Hypothesis Testing III

CONTENTS

Objectives

Introduction

14.1 Analysis of Variance

14.2 Assumptions

14.3 One-Way Analysis of Variance (One-Way ANOVA)

14.4 Procedure of One-Way ANOVA

14.5 Advantages

14.6 Limitations

14.7 Application

Summary

Keywords

Self Assessment

Answers for Self Assessment

Review Questions

Further Reading

Objectives

- Enlist assumptions of analysis of variance.
- Elaborate the procedure of one-way analysis of variance.
- Understand the interpretation of one-way analysis of variance result.
- Analyze the advantages and limitations of analysis of variance.
- Compute one-way analysis of variance appropriately.
- Interpret the result of one-way analysis of variance.

Introduction

The fourteenth unit endeavors to make detailed discussion on the topic hypothesis testing with the help of one-way analysis of variance (ANOVA). It includes the concept, assumptions, procedure, advantages, limitations, and problems of one-way ANOVA. This unit also includes the interpretation of results obtained after the application of one-way ANOVA.

14.1 Analysis of Variance

The technique of analysis of variance was first devised by Sir Ronald Fisher, an English statistician who is also known as the father of modern statistics as applied to social and behavioural sciences. It was first reported in 1923 and its early applications were in the field of agriculture. Since then, it has found wide application in many areas of experimentation.

Analysis of variance is a statistical procedure for determining whether the means of several different populations are equal.

The analysis of variance is an important method for testing the variation observed in experimental situation into different part each part assignable to a known source, cause, or factor.

In its simplest form, the analysis of variance is used to test the significance of the differences between the means of several different populations.

The problem of testing the significance of the differences between the number of means results from experiments designed to study the variation in a dependent variable with variation in independent variable.

Thus, the analysis of variance, as the name indicates, deals with variance rather than with standard deviations and standard errors.

It is a method of dividing the variation observed in experimental data into different parts, each part assignable to a known source, cause, or factor.

14.2 Assumptions

The main assumptions of ANOVA are:

- The distribution of the variable (dependent) in the population from which samples are drawn is normal (Normality).
- Variance in the populations from which the samples are drawn are equal (Homogeneity of variance).
- The total variance is equal to the sum of between/among variance and within variance.

Or

Total Variance = Between Variance + Within Variance

Total Variance = Sum of Squares for treatment + Sum of Squares for error

- Sampling should be done through random process.

14.3 One-Way Analysis of Variance (One-Way ANOVA)

One-way analysis of variance is an analysis of variance in which only one criterion (variable) is used to analyze the difference between more than two population means.

14.4 Procedure of One-Way ANOVA

The procedure to compute one-way ANOVA is as following:

- Null Hypothesis
 - $H_0: \mu_1 = \mu_2 = \mu_3$
 - There is no significant difference in means of groups
- Alternative Hypothesis
 - $H_a: \mu_1 = \mu_2 \neq \mu_3$
 - At least mean of one group is not equal to the means of other two.

Table 14.1 should be constructing for calculating the square of all scores from given score for each group which is shown as follows:

Table 14.1

X_1	X_2	X_3	X_1^2	X_2^2	X_3^2
X_{11}	X_{21}	X_{31}	X_{11}^2	X_{21}^2	X_{31}^2
X_{12}	X_{22}	X_{32}	X_{12}^2	X_{22}^2	X_{32}^2
---	---	---	---	---	---
X_{1n}	X_{2n}	X_{3n}	X_{1n}^2	X_{2n}^2	X_{3n}^2
$\sum X_1$	$\sum X_2$	$\sum X_3$	$\sum X_1^2$	$\sum X_2^2$	$\sum X_3^2$

Now,

G = Number of Groups

n_1 = Number of subjects in first group

n_2 = Number of subjects in second group

n_3 = Number of subjects in third group

N = Total number of Subjects

Compute Grand Total of Scores is given by

$$T = \sum X = \sum X_1 + \sum X_2 + \sum X_3$$

Compute Correction Factor is given by

$$C = (T^2)/N$$

Compute Total Sum of Squares is given by

$$SS_t = \sum X_1^2 + \sum X_2^2 + \sum X_3^2 - C$$

Compute Sum of Squares Between the Groups is given by

$$SS_b = (\sum X_1)^2/n_1 + (\sum X_2)^2/n_2 + (\sum X_3)^2/n_3 - C$$

Compute Sum of Square Within Groups is given by

$$SS_w = SS_t - SS_b$$

Table 14.2

Summary of ANOVA

S.O.V.	df	SS	MS	F
Between	$G - 1$	SS_b	$MS_b = SS_b / G - 1$	$F = MS_b / MS_w$
Within	$N - G$	SS_w	$MS_w = SS_w / N - G$	
Total	$N - 1$	SS_t	-	

Interpretation

From table 14.2,

$$df_1 = G - 1 \text{ for } MS_b$$

$$df_2 = N - G, MS_w$$

F-value

From F-table,

Write the F-value corresponding to (df_1, df_2) , $\alpha = 0.05$ and $\alpha = 0.01$ (or at 0.05 and at 0.01)

If $F_{\text{calculated}} < F_{\text{table at } 0.05} \& F_{\text{table at } 0.01}$

$F_{\text{calculated}}$ is not significant.

H_0 is accepted or H_a is rejected.

If $F_{\text{calculated}} > F_{\text{table at } 0.05} \& F_{\text{table at } 0.01}$

$F_{\text{calculated}}$ is significant.

H_0 is rejected or H_a is accepted.

Confidence Interval

(Bonferroni Multiple Comparison Method)

$$\underline{x}_1 - \underline{x}_2 \pm t_{\alpha/2} \text{Sqrt}(s_2(1/n_1 + 1/n_2))$$

Where,

$S_2 = \text{MSE (Mean Square Error) or}$

MS_w (Mean Sum of Square Within Groups)

$n_1 \& n_2 = \text{number of observations in sample 1 \& 2}$

$\underline{x}_1 \& \underline{x}_2 = \text{mean of sample population 1 \& 2}$

If no significant difference, then zero included in interval.

If zero is included in the interval, we may conclude that there is no significant difference in the selected population means. i.e., there is no difference between the means of groups.

If significant difference, then end points of interval have same sign.

If the end points of the confidence interval have the same sign, then we may conclude that there is a significant difference between the selected population means.



Did you know?

- Mean Square Error (MSE) - The mean of the squared errors used to judge the quality of a set of errors.

14.5 Advantages

The advantages of one-way ANOVA are:

- It is an improved technique over the t or z - test.
- It evaluates both types of variances between and within.
- This technique is used for ascertaining the difference among several groups or treatments at time.
- It is an economical device.
- The experimental designs i.e., simple random and Level x treatments designs are based on one way analysis of variance technique.

14.6 Limitations

The limitations of one-way ANOVA are:

- The analysis of variance technique has certain assumptions - normality & homogeneity of distribution of data. The departure of the data from these assumptions may effect adversely on the inferences.
- The F- value provides global findings of difference among groups but it cannot specify the inference. When F value is significant, t test is followed for specifying the statistical inferences.
- Statistical table of F value is essential for the use of F test because without it results cannot be interpreted.

14.7 Application



Example

A study investigated the perception of corporate ethical values among individual specializing in marketing. The data is given below:

Advertising	5	6	5	4	4	6
Marketing Manager	4	5	4	5	5	4
Marketing Research	7	6	6	6	5	6

Test for significant differences in perception among three groups.

Solution:

- $H_0: \mu_1 = \mu_2 = \mu_3$ and $H_a: \mu_1 = \mu_2 \neq \mu_3$

Table 14.3

X_1	X_2	X_3	X_1^2	X_2^2	X_3^2
5	4	7	25	16	49
6	5	6	36	25	36
5	4	6	25	16	36
4	5	6	16	25	36
4	5	5	16	25	25
6	4	6	36	16	36
$\sum X_1 = 30$	$\sum X_2 = 27$	$\sum X_3 = 36$	$\sum X_1^2 = 154$	$\sum X_2^2 = 123$	$\sum X_3^2 = 218$

Now,

G = Number of Groups = 3

n_1 = Number of subjects in first group = 6

n_2 = Number of subjects in second group = 6

n_3 = Number of subjects in third group = 6

N = Total number of Subjects = $n_1 + n_2 + n_3$ = 18

Compute Grand Total of Scores is given by

$$\begin{aligned} T &= \sum X = \sum X_1 + \sum X_2 + \sum X_3 \\ &= 30 + 27 + 36 \\ &= 93 \end{aligned}$$

Compute Correction Factor is given by

$$\begin{aligned} C &= (T^2)/N \\ &= (93)^2/18 \\ &= 8649/18 \\ &= 480.5 \end{aligned}$$

Compute Total Sum of Squares is given by

$$\begin{aligned} SS_t &= \sum X_1^2 + \sum X_2^2 + \sum X_3^2 - C \\ &= 154 + 123 + 218 - 480.5 \\ &= 495 - 480.5 \\ &= 14.5 \end{aligned}$$

Compute Sum of Squares Between the Groups is given by

$$\begin{aligned} SS_b &= (\sum X_1)^2/n_1 + (\sum X_2)^2/n_2 + (\sum X_3)^2/n_3 - C \\ &= (30)^2/6 + (27)^2/6 + (36)^2/6 - 480.5 \\ &= 900/6 + 729/6 + 1296/6 - 480.5 \\ &= 150 + 121.5 + 216 - 480.5 \\ &= 487.5 - 480.5 \\ &= 7 \end{aligned}$$

Compute Sum of Square Within Groups is given by

$$\begin{aligned} SS_w &= SS_t - SS_b \\ &= 14.5 - 7 \\ &= 7.5 \end{aligned}$$

Table 14.4

Summary of ANOVA

S.O.V.	df	SS	MS	F
Between	$G - 1 = 3 - 1 = 2$	$SS_b = 7$	$MS_b = SS_b / G - 1 = 7 / 2 = 3.5$	$F = MS_b / MS_w$ $= 3.5 / 0.5$ $= 7$
Within	$N - G = 18 - 3 = 15$	$SS_w = 7.5$	$MS_w = 7.5 / 15 = 0.5$	
Total	$N - 1 = 18 - 1 = 17$	$SS_t = 14.5$	-	

Interpretation

$df_1 = 2, MS_b = 3.5$

$df_2 = 15, MS_w = 0.5$

F-value from table -

at 0.05

at 0.01

3.68

6.36

From Table, For df - ($df_1 = 2, df_2 = 15$) and $\alpha = 0.05$

F - value = 3.68

Since $F_{\text{calculated}}(7) > F_{\text{table at 0.05}}(3.68)$ & $F_{\text{table at 0.01}}(6.36)$

$F_{\text{calculated}}$ is significant.

H_0 is rejected or H_a is accepted.

$$\text{Now, } \bar{X}_1 = \sum X_1 / n_1 = 30/6 = 5$$

$$\bar{X}_2 = \sum X_2 / n_2 = 27/6 = 4.5$$

$$\bar{X}_3 = \sum X_3 / n_3 = 36/6 = 6$$

Confidence Interval (1 and 2)

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2} \text{Sqrt}(s_2(1/n_1 + 1/n_2))$$

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2} \text{Sqrt}(MS_w(1/n_1 + 1/n_2)) \quad (s_2 = MS_w)$$

$$= (5-4.5) \pm 2.13 \text{Sqrt}(0.5(1/6 + 1/6))$$

$$= 0.5 \pm 2.13 \text{Sqrt}(0.5(2/6))$$

$$= 0.5 \pm 2.13 \text{Sqrt}(0.167)$$

$$= 0.5 \pm 2.13 \times 0.41$$

$$= 0.5 \pm 0.87$$

$$= 0.5 - 0.87 \text{ and } 0.5 + 0.87$$

$$= -0.37 \text{ and } +1.37$$

No significance difference (Mean ethical value of gp1 is not significantly differ from mean ethical value of gp2).

Confidence Interval (1 and 3)

$$\bar{x}_1 - \bar{x}_3 \pm t_{\alpha/2} \text{Sqrt}(s_2(1/n_1 + 1/n_2))$$

$$= (5-6) \pm 2.13 \text{Sqrt}(0.5(1/6 + 1/6))$$

$$= -1 \pm 2.13 \text{Sqrt}(0.5(2/6))$$

$$= -1 \pm 2.13 \text{Sqrt}(0.167)$$

$$= -1 \pm 2.13 \times 0.41$$

$$= -1 \pm 0.87$$

$$= -1 - 0.87 \text{ and } -1 + 0.87$$

$$= -1.87 \text{ and } -0.13$$

Significance difference (Mean ethical value of gp1 is significantly lower than mean ethical value of gp 3 or mean ethical value of gp3 is significantly higher than mean ethical value of gp 1).

Confidence Interval (2 and 3)

$$\bar{x}_2 - \bar{x}_3 \pm t_{\alpha/2} \text{Sqrt}(s_2(1/n_1 + 1/n_2))$$

$$= (4.5-6) \pm 2.13 \text{Sqrt}(0.5(1/6 + 1/6))$$

$$= -1.5 \pm 2.13 \text{Sqrt}(0.5(2/6))$$

$$= -1.5 \pm 2.13 \text{Sqrt}(0.167)$$

$$= -1.5 \pm 2.13 \times 0.41$$

$$= -1.5 \pm 0.87$$

$$= -1.5 - 0.87 \text{ and } -1.5 + 0.87$$

$$= -2.37 \text{ and } -0.63$$

Significance difference (Mean ethical value of gp2 is significantly lower than mean ethical value of gp 3 or mean ethical value of gp3 is significantly higher than mean ethical value of gp 2).

Summary

Analysis of variance is used to test the significance of the differences between the means of several different populations.

The main assumptions of ANOVA are - normality, homogeneity of variance, random samples, and total variance equals to the sum of between/among variance and within variance.

One-way analysis of variance is an analysis of variance in which only one criterion (variable) is used to analyze the difference between more than two population means.

The procedure to compute one-way ANOVA is as following:

- Null Hypothesis
 - $H_0: \mu_1 = \mu_2 = \mu_3$
 - There is no significant difference in means of groups
- Alternative Hypothesis
 - $H_a: \mu_1 = \mu_2 \neq \mu_3$
 - At least mean of one group is not equal to the means of other two.
- Construct the table for calculating the square of all scores from given score for each group and compute following:
 - G = Number of Groups
 - n_1 = Number of subjects in first group
 - n_2 = Number of subjects in second group
 - n_3 = Number of subjects in third group
 - N = Total number of Subjects
 - Compute Grand Total of Scores is given by $T = \sum X = \sum X_1 + \sum X_2 + \sum X_3$
 - Compute Correction Factor is given by $C = (T^2)/N$
 - Compute Total Sum of Squares is given by $SS_t = \sum X_1^2 + \sum X_2^2 + \sum X_3^2 - C$
 - Compute Sum of Squares Between the Groups is given by

$$SS_b = (\sum X_1)^2/n_1 + (\sum X_2)^2/n_2 + (\sum X_3)^2/n_3 - C$$
 - Compute Sum of Square Within Groups is given by $SS_w = SS_t - SS_b$
 - Summary of ANOVA
 - Calculate degrees of freedom, compare calculated F-value with table F-value, and interpret the results
 - If calculated F-value is significant, then apply Post hoc (Pair-wise comparisons) to check the significance of difference between groups.

The advantages of one-way ANOVA are - an improved technique over the t or z - test; evaluates variances between and within group; used for ascertaining the difference among several groups or treatments at time; an economical device; and simple random & (Level x treatments) designs are based on one way analysis of variance technique.

The limitations of one-way ANOVA are - the departure of the data from assumptions may effect adversely on the inferences; the F- value provides global findings of difference among groups but it cannot specify the inference; and statistical table of F value is essential for the use of F test.

Keywords

Analysis of variance is a statistical procedure for determining whether the means of several different populations are equal.

One-way analysis of variance is an analysis of variance in which only one variable is used to analyze the difference between more than two population means.

Self Assessment

- ANOVA is a statistical procedure for determining whether the
 - means of two different populations are equal.
 - means of several different populations are unequal.
 - means of two different populations are unequal.
 - means of several different populations are equal.
- Which of the following expression is correct?
 - Total Variance = Sum of Squares for within treatment + Sum of Squares for variance
 - Total Variance = Sum of Squares for treatment + Sum of Squares for error
 - Total Variance = Sum of Squares for between treatment + Sum of Squares for error
 - Total Variance = Sum of Squares for treatment + Sum of Squares for variance error
- The correct formula to calculate the total Sum of Squares is
 - $(\sum X_1)^2/n_1 + (\sum X_2)^2/n_2 + (\sum X_3)^2/n_3 - C$
 - $\sum X_1^2 + \sum X_2^2 + \sum X_3^2 + C$
 - $(\sum X_1)^2/n_1 + (\sum X_2)^2/n_2 + (\sum X_3)^2/n_3 + C$
 - $\sum X_1^2 + \sum X_2^2 + \sum X_3^2 - C$
- The correct formula to calculate the sum of squares between the groups is
 - $SS_t - SS_w$
 - $SS_t + SS_w$
 - $SS_w - SS_t$
 - $SS_w + SS_t$
- If $F_{\text{calculated}} < F_{\text{table at } 0.05} \& F_{\text{table at } 0.01}$ then
 - $F_{\text{calculated}}$ is significant.
 - The null hypothesis is rejected.
 - $F_{\text{calculated}}$ is not significant.
 - An alternate hypothesis is accepted.
- If $N = 12$ and $T = 56$ then the value of the correction factor is

- A. 261.33
B. 2.57
C. 4.67
D. 214.28
7. If $n_1 = n_2 = n_3 = n_4 = 5$, $\sum X_1 = 12$, $\sum X_2 = 40$, $\sum X_3 = 21$, and $\sum X_4 = 27$ then the value of SS_b is
- A. 88.20
B. 87.80
C. 82.80
D. 81.80
8. If $n_1 = n_2 = n_3 = n_4 = 5$, $\sum X_1^2 = 46$, $\sum X_2^2 = 334$, $\sum X_3^2 = 105$, and $\sum X_4^2 = 167$ then the value of SS_t is
- A. 150
B. 152
C. 252
D. 142
9. If $N = 20$, $G = 4$, $SS_b = 1853.60$, and $SS_w = 32.58.40$ then the value of F-ratio is
- A. 6.39
B. 9.36
C. 8.42
D. 2.84
10. If $F_{\text{calculated}} = 5.24$, $F_{\text{table at } 0.05} = 3.24$ & $F_{\text{table at } 0.01} = 6.39$ then
- A. $F_{\text{calculated}}$ is significant at 0.01 and the null hypothesis is rejected at 0.01.
B. $F_{\text{calculated}}$ is significant at 0.05 and the null hypothesis is accepted at 0.05.
C. $F_{\text{calculated}}$ is significant at 0.05 & 0.01. The null hypothesis is rejected at 0.05 & 0.01.
D. $F_{\text{calculated}}$ is significant at 0.05 and the null hypothesis is rejected at 0.05.
11. Which of the following is the limitation of one-way ANOVA?
- A. It is an improved technique over the t or z - test.
B. It evaluates both types of variances between and within.
C. The departure of the data from these assumptions may effect adversely on the inferences.
D. It is an economical device.
12. Which of the following is the assumption of ANOVA?
- A. Normality and homogeneity of variance
B. Non-random samples and normality
C. Homogeneity of variance and Non-random samples
D. Data does not follow normality.

13. If F-ratio is significant then

- A. apply descriptive statistics to test significance difference between pairs of groups.
- B. apply correlation to test significance difference between pairs of groups.
- C. apply chi-square to test significance difference between pairs of groups.
- D. apply post-hoc to test significance difference between pairs of groups.

14. If $T = 98$ and $N = 20$ then the value of correction factor is

- A. 4.08
- B. 480.20
- C. 4.9
- D. 0.25

15. If $SS_w = 35$ and $SS_b = 45$, then the value of SS_t is

- A. 10
- B. 20
- C. 80
- D. 90

Answers for Self Assessment

1. D 2. B 3. D 4. A 5. C
 6. A 7. C 8. B 9. D 10. D
 11. C 12. A 13. D 14. B 15. C

Review Questions

1. Explain the procedure of one-way ANOVA.
2. Discuss the advantages of one-way ANOVA.
3. Analyze the limitations of one-way ANOVA.
4. Apply one-way ANOVA on the following data and interpret the result:

Observations	1	2	3	4
Observer				
A	8	2	6	3
B	5	6	2	9

C	2	9	6	2
D	8	4	6	6

5. A certain manure was used on four plots of land A, B, C and D. Four beds were prepared in each plot and the manure used. The output of the crop in the beds of plots A, B, C and D is given below:

Output on Plots

A	31	34	38	35
B	37	36	32	39
C	35	39	36	34
D	38	37	33	35

Find out whether the difference in the means of the production of crops of the plots is significant or not.

6. A simple random design of experiments is used for testing the difference among three treatments and four subjects are assigned to each treatment. The obtained data are as follows:

A	2	4	2	3
B	5	6	2	6
C	2	5	9	3

Is the difference among the treatment significant?

7. The following data give the scores of 5 students on three successive tests of a mathematics subject:

Test I	20	25	21	24	26
Test II	18	24	19	30	17
Test III	22	20	34	26	32

Is the difference between scores in successive three tests significant?

8. A simple random design of experiments is used for testing the difference among five treatments and three subjects are assigned to each treatment. The obtained data are as follows:

X	12	18	20
Y	13	17	15
Z	15	19	11

A	19	13	17
B	16	21	13

Is the difference among the treatment significant?

9. The aim of an experimental study was to find out the effect of four different techniques of training on the learning of a particular skill. Four groups, each consisting of five students of class XII assigned randomly & were given training through these different techniques. The scores obtained on a performance test were recorded as follows:

Group I -	3	5	1	7	6	
Group II -	4	5	3	8	4	
Group III	-	5	4	1	3	2
Group IV	-	5	2	4	1	3

Apply the analysis of variance to test the null hypothesis at both the levels of significance.

10. The following table illustrates the sample psychological health ratings of corporate executives in the field of Banking, Manufacturing and Fashion retailing:

<i>Banking</i>	41	53	54	55	43
<i>Manufacturing</i>	45	51	48	43	39
<i>Fashion retailing</i>	34	44	46	45	51

Can we consider the psychological health of corporate executives in the given three fields to be equal at 1% and 5% level of significance?



Further Reading

1. Research Methodology- Methods and Techniques by Gaurav Garg and C. R. Kothari, New Age International (P) Limited.
2. Methodology of Educational Research by Lokesh Koul, Vikas Publishing House Pvt. Ltd.
3. Essentials of Scientific Behavioural Research by R.A. Sharma, R. Lall Book Depot.



Web Links

<https://courses.lumenlearning.com/suny-natural-resources-biometrics/chapter/chapter-5-one-way-analysis-of-variance/>

LOVELY PROFESSIONAL UNIVERSITY

Jalandhar-Delhi G.T. Road (NH-1)

Phagwara, Punjab (India)-144411

For Enquiry: +91-1824-521360

Fax.: +91-1824-506111

Email: odl@lpu.co.in