Detailed Contents

List of Figures xi
List of Tables xvii
List of Abbreviations xix
About the Book xxi
Acknowledgements xxiii
About the Authors xxv

Chapter 1
Getting Started with Microsoft Excel 1
1.1. Glimpse of Excel 2019 2
1.2. Basic Data Manipulation 3
1.3. Finding the Number of Rows and Columns in Data 4
1.4. Formatting Decimal Places 4
1.5. Sorting Data 6
1.6. Unhide and Hide Rows and Columns 9
1.7. Filtering Data 10
1.8. Freezing Worksheet Panes 14

Chapter 2
Basic Arithmetic Operations in Excel 15
2.1. Addition 16
2.2. Subtraction 18
2.3. Multiplication 18
2.4. Division 21
2.5. Relative and Absolute References 22
2.6. AutoSum Function 23
2.7. SUMIF Function 23
2.8. MAX and MIN Function 25
2.9. COUNT and COUNTIF Function 25
2.10. Power Function 27
2.11. Pivot Tables 28

Chapter 3
Data: Input into Excel and Sources 37
3.1. Understanding Data 38
3.2. Data Entry in Excel 38
3.3. Importing Data in Excel 40
3.4. Data Sets Available in the Public Domain 41

CHAPTER 4
Visualization of Data Using Charts in Excel
4.1. Line Graph
4.2. Column Chart
4.3. Histogram
4.4. Pie Chart
4.5. Scatter Plot

CHAPTER 5
Measures of Central Tendency
5.1. Arithmetic Mean
5.2. Median
5.3. Mode
5.4. Weighted Arithmetic Mean
5.5. Geometric Mean
5.6. Harmonic Mean
5.7. Relationship between Arithmetic Mean, Geometric Mean and Harmonic Mean 1

CHAPTER 6
Measures of Dispersion 1
6.1. Range 1
6.2. Standard Deviation 1
6.3. Variance 1
6.4. Coefficient of Variation 1
6.5. Skewness 1
6.6. Kurtosis 1
6.7. Data Analysis Using Excel 1

CHAPTER 7
Compounding: Future Value of Money 1
7.1. Time Value of Money 1
7.2. Simple Interest 1
7.3. Compound Interest 1
7.4. Future Value of an Annuity 1

CHAPTER 8
Investment Decision Criteria: NPV and IRR 1
8.1. Net Present Value (NPV) 1
8.2. Merits and Demerits of NPV 1
8.3. Internal Rate of Return (IRR) 1
8.4. Common Applications of IRR 1
8.5. IRR versus Extended Internal Rate of Return (XIRR) 1
8.6. Merits and Demerits of IRR 1

CHAPTER 9
Loan Amortization 164
9.1. Type of Loan: Fixed Interest versus Reducing Balance Method 165
9.2. Loan Amortization with Table 168
9.3. Amount of Loan after n Periods 169
9.4. Payment or EMI Using Excel 171
9.5. Cumulative Principal and Interest Amount Using Excel 172
9.6. PPMT and IPMT 174
9.7. Effective Annual Interest Rate 177

CHAPTER 10
Correlation Coefficient 181
10.1. Covariance 181
10.2. Correlation 185
10.3. Karl Pearson's Coefficient of Correlation 187
10.4. Spearman's Coefficient of Correlation 195

CHAPTER 11
Regression Analysis 206
11.1. Regression Analysis 207
11.2. Developing a Linear Regression Model 207
11.3. Estimation of the Coefficients in the Regression Model 210
11.4. Interpretation of the Estimated Coefficients in the Regression Model 215
11.5. Forecast the Values of the Dependent Variable 215
11.6. Residuals or Errors and OLS Method 219
11.7. Coefficient of Determination 220

CHAPTER 12
Hypothesis Testing in Regression Analysis 231
12.1. Hypothesis Testing Using the *t*-Statistics 232
12.2. Hypothesis Testing Using the *p*-value 236
12.3. Hypothesis Testing Using the Confidence Interval 239

Appendix 1: Present Value and Future Value Tables 247
Appendix 2: Table of the Student's t-distribution 251
Index 252

LIST OF FIGURES

1.1	Layout of Excel Worksheet	2
1.2	Finding the Number of Rows and Columns in Data	4
1.3	Profile of Rural Workers	5
1.4	Formatting Decimal Points	6
1.5	Formatting Decimal Points Using Home Tab	6
1.6	Formatting up to Two Decimal Points	7
1.7	Step 1 in Sorting Data	7
1.8	Step 2 in Sorting Data	8
1.9	Step 3 in Sorting Data	8
1.10	Hidden Column	9
1.11	Filtering Data through Key Commands	10
1.12	Filtering Data through Tabs	11
1.13	Filtering Data Using Text Filters	12
1.14	Custom Filter	12
1.15	Number Filters	13
1.16	Copy and Paste of Filtered Data	13
1.17	Freezing /Unfreezing of Panes	14
2.1	Data of Workers Employed in Non-farm Activities	16
2.2	Step 1 in Adding of Cells	17
2.3	Step 2 in Adding of Cells	18
2.4	Subtracting of Cells	19
2.5	Method 1 of Multiplication of Cells	20
2.6	Method 2 of Multiplication of Cells	20
2.7	Division Function in Excel	21
2.8	Absolute Cell Reference	22
2.9	Usage of the AutoSum Function	23
2.10	Usage of SUMIF Function	24
2.11	The MAX Function	25
2.12	The MIN Function	26
2.13	The COUNT Function	26
2.14	The COUNTIF Function	27
2.15	The Power Function	28
2.16	Sales Data of a Company	29
2.17	Inserting a Pivot Table	30
2.18	Illustration of a Pivot Table Template	30
2.19	Summary of Results	31
2.20	Step 1 of Data Filtering in Pivot Table	32
2.21	Step 2 of Data Filtering in Pivot Table	32

2.22 Summarizing of Values 33
2.23 Forms of Summarized Values 34

3.1 Adding Form to Quick Access Toolbar 39
3.2 Data Entry Form 40
3.3 DBIE Website: Statistics 42
3.4 Data on Real Sector on DBIE Website 42
3.5 Exporting Data (Sector-wise) 44
3.6 The MoSPI Website 45
3.7 The Downloading of Reports at MoSPI 46
3.8 Statistical Publications Provided by MoSPI 47
3.9 Press Release 47
3.10 Downloading of Tables Data on MoSPI Website 48
3.11 Step 1 in Microdata Catalogue 49
3.12 Step 2 in Microdata Catalogue (National Data Archive) 49
3.13 Step 3 in Microdata Catalogue (National Data Archive) 50
3.14 National Data Bank 51
3.15 The New NSE Website 52
3.16 Step 1 in Downloading Historical Reports—NSE 52
3.17 Step 2 in Downloading Historical Reports—NSE 53
3.18 Step 3 in Downloading Historical Reports—NSE 53
3.19 Step 1 in Historical Data of a Company—NSE 54
3.20 Step 2 in Historical Data of a Company—NSE 54
3.21 Step 3 in Historical Data of a Company—NSE 55
3.22 Step 1 in Downloading Historical Data—BSE 55
3.23 Step 2 in Downloading Historical Data—BSE 56
3.24 Step 3 in Downloading Historical Data—BSE 56
3.25 Step 4 in Downloading Historical Data—BSE 57
3.26 Step 5 in Downloading Historical Data—BSE 57
3.27 World Bank Open Data 58
3.28 Downloading Data: Indicator-wise 59
3.29 Downloading Data: Country-wise 59
3.30 Downloading Data: Combination of an Indicator and Country 60
3.31 Downloading Data: Combination of an Indicator and Multiple Countries 60
3.32 UNdata Website 61
3.33 Step 1 in Downloading of Data Series—UNdata 62
3.34 Step 2 in Downloading of Data Series—UNdata 62
3.35 Step 3 in Downloading of Data Series—UNdata 63
3.36 Step 4 in Downloading of Data Series—UNdata 63

4.1 Per Capita Gross Domestic Product (2011 PPP $) 65
4.2 Line Graph in Excel 66
4.3 Labelling of Chart 67
4.4 Per Capita GDP of SAARC Countries, 2010–2018 68

4.5	Steps of Depicting Per Capita GDP of SAARC Countries, 2010–2018 in Column Chart	69
4.6	Column Chart Depicting Per Capita GDP	70
4.7	The Histogram: Direct Method	71
4.8	The Histogram: Another Method	73
4.9	Steps in Labelling Pie Chart, Motor Vehicle Production 2008–2009	75
4.10	Pie Chart of Motor Vehicle Production 2009–2010	75
4.11	The Scatter Plot of Company ABC	77
5.1	Step 1 in Calculating the Mean	83
5.2	Step 2 in Calculating the Mean	84
5.3	Step 3 in Calculating the Mean	84
5.4	Step 1 in Calculating the Median	88
5.5	Step 2 in Calculating the Median	89
5.6	Step 3 in Calculating the Median	89
5.7	Step 1 in Calculating the Mode	91
5.8	Step 2 in Calculating the Mode	91
5.9	Step 3 in Calculating the Mode	92
5.10	Calculating the Mode Using Graph	93
5.11	Step 1 in Calculating the Mode Using MAX	94
5.12	Step 2 in Calculating the Mode Using MAX	94
5.13	Step 1 in Calculating the Weighted Arithmetic Mean	97
5.14	Step 2 in Calculating the Weighted Arithmetic Mean	97
5.15	Step 1 in Calculating the Geometric Mean	98
5.16	Step 2 in Calculating the Geometric Mean	98
5.17	Step 1 in Calculating the Harmonic Mean	101
5.18	Step 2 in Calculating the Harmonic Mean	101
6.1	Scatter Plot for Scores of Class A	106
6.2	Scatter Plot for Scores of Class B	107
6.3	Step 1 in Calculation of Range	109
6.4	Step 2 in Calculation of Range	110
6.5	Step 1 in Calculation of Standard Deviation	113
6.6	Step 2 in Calculation of Standard Deviation	114
6.7	Step 1 in Calculation of Variance	117
6.8	Step 2 in Calculation of Variance	118
6.9	Step 1 in Calculation of Skewness	124
6.10	Step 2 in Calculation of Skewness	124
6.11	Histogram	125
6.12	Step 1 in Calculation of Kurtosis	127
6.13	Step 2 in Calculation of Kurtosis	128
6.14	Step 1 in Data Analysis (Descriptive Statistics)	130
6.15	Step 2 in Data Analysis (Descriptive Statistics)	130
6.16	Step 3 in Data Analysis (Descriptive Statistics)	131

7.1 Impact of Compounding in Investment 140
7.2 Calculating Future Value Using Excel 143
7.3 Calculating Future Value of Annuity Using Excel 146

8.1 Calculating NPV Using Excel 152
8.2 Calculating IRR Using Excel 156
8.3 Calculation of IRR of Endowment Plans 158
8.4 Calculating IRR in a Moneyback Policy 158
8.5 Calculating XIRR 159

9.1 Calculating EMI Using Fixed Rate Method 165
9.2 Calculating Outstanding Loan Amount Using PV Method 171
9.3 Calculating EMI Amount Using Excel 172
9.4 CUMPRINC Function in Excel 173
9.5 CUMIPMT Function in Excel 174
9.6 PPMT Function in Excel 175
9.7 IPMT Function in Excel 176
9.8 Effective Annual Rate Function 178

10.1 Step 1 in Calculation of Covariance 184
10.2 Step 2 in Calculation of Covariance 185
10.3 Positive Correlation 186
10.4 Negative Correlation 186
10.5 No Correlation 187
10.6 Step 1 in Calculation of Correlation 190
10.7 Step 2 in Calculation of Correlation 190
10.8 Step 1 in Calculation of Pearson Correlation Coefficient 191
10.9 Step 2 in Calculation of Pearson Correlation Coefficient 192
10.10 Step 1 in Calculation of Correlation using Data Analysis 192
10.11 Step 2 in Calculation of Correlation using Data Analysis 193
10.12 Step 3 in Calculation of Correlation using Data Analysis 193
10.13 Scatter Plot Showing Linear Relationship between the Monthly Rate of Returns on Sensex (*X*) and Nifty 50 (*Y*) 195
10.14 Step 1 in Calculation of Spearman Correlation Coefficient 199
10.15 Step 2 in Calculation of Spearman Correlation Coefficient 200
10.16 Step 3 in Calculation of Spearman Correlation Coefficient 200
10.17 Step 4 in Calculation of Spearman Correlation Coefficient 201
10.18 Relationship between Hours of Work and Average Cost 203

11.1 Graphical Representation of Regression Line 208
11.2 Regression Line with Positive Slope ($\beta_1 > 0$) 209
11.3 Regression Line with Negative Slope ($\beta_1 < 0$) 209
11.4 Regression Line with Constant Slope ($\beta_1 = 0$) 209
11.5 Step 1 in Calculating Regression Parameters 214

11.6	Step 2 in Calculating Regression Parameters	214
11.7	Step 3 in Calculating Regression Parameters	214
11.8	Step 4 in Calculating Regression Parameters	215
11.9	Step 1 in Regression Analysis (Predicting the Dependent Variable)	217
11.10	Step 2 in Regression Analysis (Predicting the Dependent Variable)	217
11.11	Population (PRF) and Sample Regression Functions (SRF)	218
11.12	Graphical Representation of Three Types of Variations	220
11.13	Regression Analysis Dialogue Box	223
11.14	Regression Analysis (Summary Output)	224
11.15	Regression Analysis (Residual Output)	224
11.16	Regression Analysis: Promotional Expense Line Fit Plots	225
12.1	Rejection Areas for a Two-tailed Test	234
12.2	Rejection Areas for a One-tailed Test	234
12.3	Regression Analysis Output: Residual Degree of Freedom	236
12.4	The *t*-distribution *p*-value > Alpha (α)	238
12.5	The *t*-distribution *p*-value < Alpha (α)	238
12.6	The *t*-distribution Computed *p*-value < Alpha (α)	238
12.7	Regression Analysis Output (95% Confidence Intervals)	239
12.8	Test Results of the Regression Equation Displayed in the Excel Sheet	240

1

CHAPTER

Getting Started with Microsoft Excel

LEARNING OBJECTIVES

After reading this chapter, the readers will be able to understand

- Basics of the Excel program
- How to find the number of rows and columns
- Formatting the decimal places
- Sorting of data
- Unhide or hide some of the rows and columns
- Filtering the data using various methods
- Freezing worksheet panes

Excel is a spreadsheet/worksheet program which is used by most people, be it businessmen, college students, researchers, financial analysts and many more. Released in 1985, it is now widely used. Its popularity arises out of its ability to do mass arithmetic calculations like addition, subtraction, multiplication and division of thousands of numbers simultaneously. It automates calculations and even supports complex functions when used with other functions included in the software. Further, it is a perfect tool for creating graphs and tables for visual data analysis. In fact, it does much more. Not only it aggregates and simplifies numbers for data management, but it also processes raw data towards deeper data analysis. Hence, working knowledge of Excel is considered vitally important in today's world, irrespective of professionals in

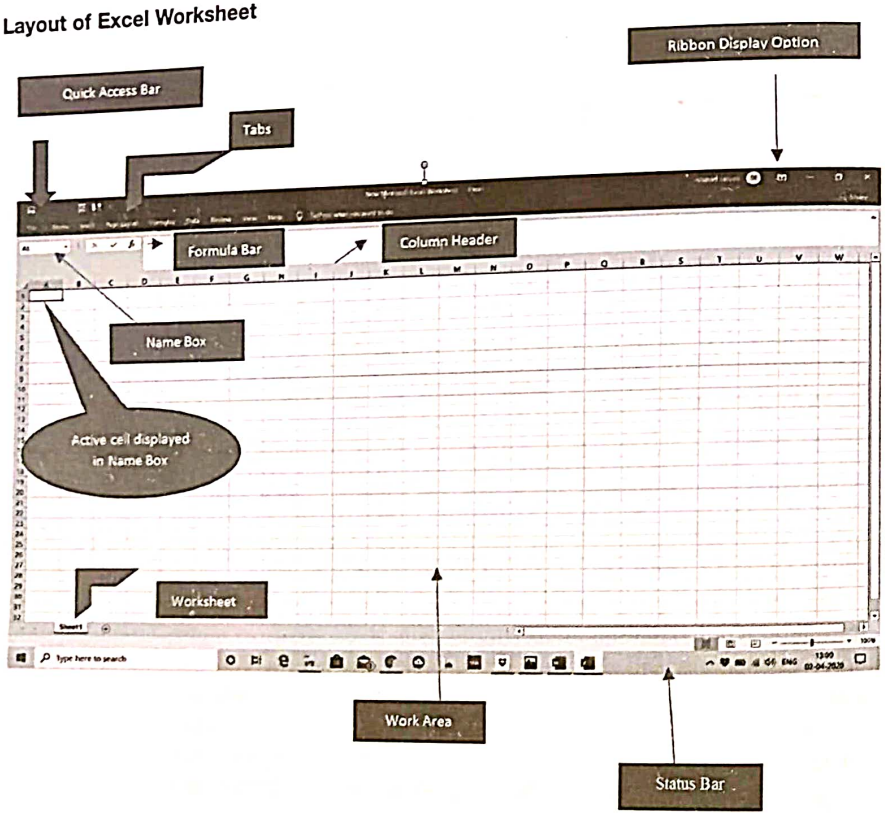
any field. The following section is an introduction to the Excel program. It covers topics that an undergraduate student of the computational techniques paper is expected to be familiar with. At this point, no prior knowledge of software is assumed.

1.1. GLIMPSE OF EXCEL 2019

The display of a newly opened Excel spreadsheet is presented in Figure 1.1. A quick display of a few outlines that help handle data analysis is described to assist in beginner's comprehension. It contains certain comment boxes with explanations. Before elucidation, click the **Ribbon Display Option** on the top-right corner of the document and select **Show Tabs**. Selecting **Auto-hide Option** will hide **tabs** from the display sheet which will make it difficult to go ahead with our data analysis.

FIGURE 1.1

Layout of Excel Worksheet



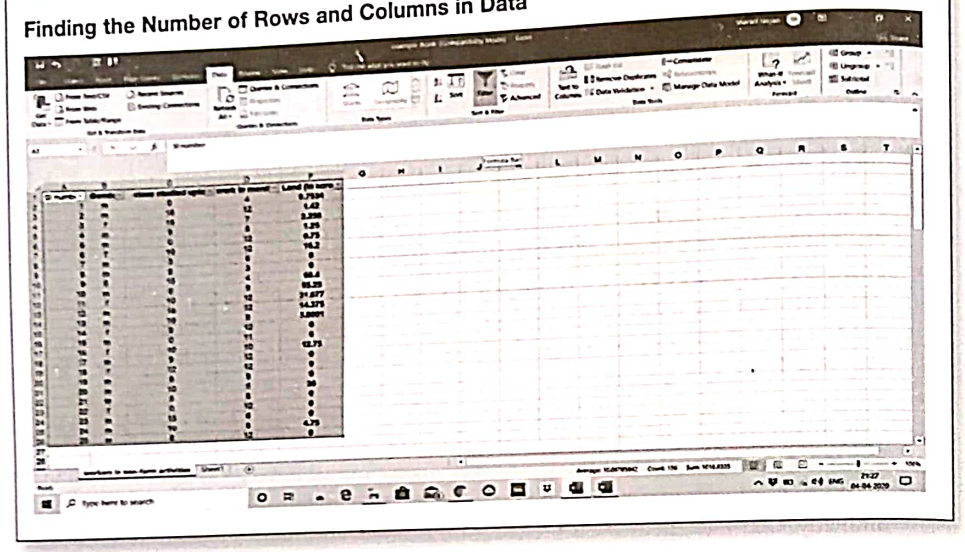
- **Quick Access Toolbar:** The **Quick Access Toolbar** consists of favourite commands. By default, it displays commands such as **Save**, **Undo** and **Redo**. As we will proceed with the data analysis, we may use more commands like **Sort**, **Copy** and **Format** etc. frequently which can be added to the **Quick Access Toolbar** using the drop-down arrow. From the drop-down list by selecting **More Commands**, we can choose a number of functions arranged in alphabetical order.
- **Tabs:** The tabs on the ribbon area are **File**, **Home**, **Insert**, **Page Layout**, **View** and **Help**. Each tab contains groups. The groups are further separated by a vertical line. Various functions are listed in these groups which can be used as per requirement. The **Home** tab contains the most frequently used commands in Excel.
- **Name Box:** It displays the selected cell in the spreadsheet. Alternatively, one can type cell box number and press **Enter** to reach the desired cell.
- **Formula Bar:** The **Formula Bar** displays the data or formula entered in the active cell. This **Formula Bar** may also be used to edit entered data or formula. The active cell displays the results of the formula entered while we see the formula itself in the **Formula Bar**.
- **Work area:** The entire cell area where data can be placed is called the **work area**. These cells are conspicuous in the format of rows and columns. The entire **work area** consists of 1,048,576 rows and 16,384 columns. The easiest way to find the number of rows can be observed by selecting any cell and then pressing **Ctrl + Down** arrow on the keyboard. Similarly, **Ctrl + Right** arrow is pressed for the number of columns.
- **Worksheet:** In Excel, a sheet is generally called a **Worksheet**. It is listed at the bottom of the document. A worksheet can be thought of as a single sheet of paper where calculations are performed. Any workbook contains at least one worksheet by default subject to the maximum of 255 worksheets. Each sheet can have its name. It can be renamed by right clicking on the worksheet tab. A new sheet can be created by clicking on the **Insert Sheet** icon if the data is large enough as well as heterogeneous. The sheets can be **deleted**, **renamed**, **moved** or **copied** by right clicking that worksheet tab.
- **Zoom in and zoom out:** In most cases, one can quickly use the minus and plus symbols in the **Status Bar** to zoom the document.
- **Adjustment of column width in the work area:** It may happen sometimes that the data/text entered in a cell exceeds the width of a cell. To adjust this overlapping to another cell, place the cursor at the end of the respective **column header** until this ↔ appears. Now double click this sign and the size of the column will adjust automatically to the cell having the lengthiest data.

1.2. BASIC DATA MANIPULATION

With this level of familiarity, this section illustrates the techniques necessarily required in handling of data. Specifically, we will see how to move across cells in a spreadsheet, for example, how to move from the top of a column to the bottom of the column, how to format cells, how to **Sort**

FIGURE 1.2

Finding the Number of Rows and Columns in Data



various columns in the spreadsheet and how to **hide/unhide** some of the columns. To explain it, in Figure 1.2, we have taken a hypothetical data of rural workers engaged in non-farm activities in a village. This data pertains to 25 workers only; in real life, the data may run into thousands of rows and hundreds of columns.

1.3. FINDING THE NUMBER OF ROWS AND COLUMNS IN DATA

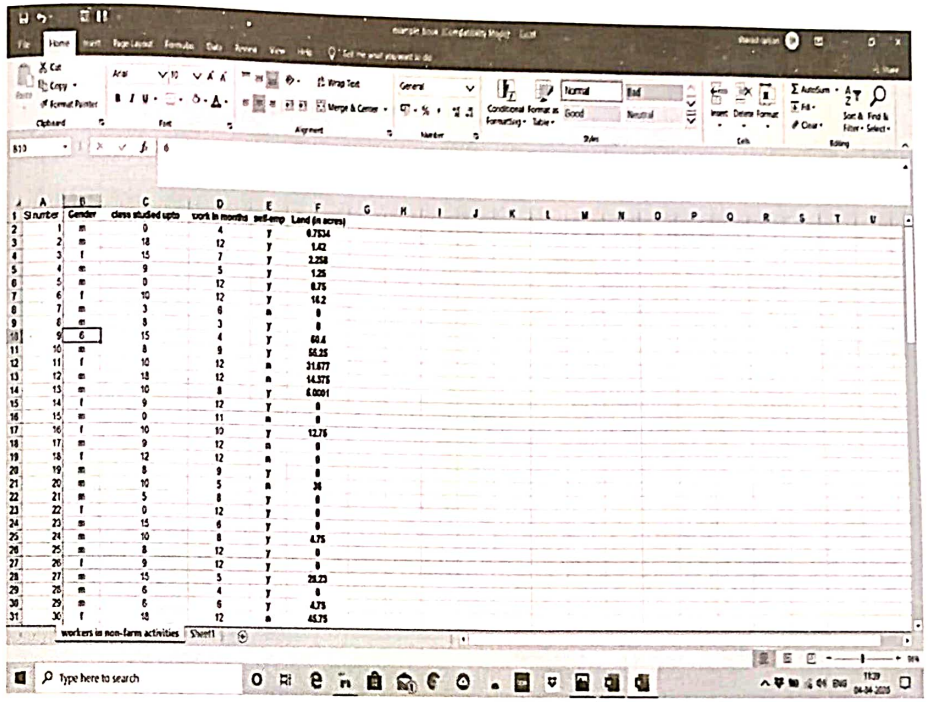
In a given data, to find out how many workers are engaged in non-farm activities. In other words, how many rows of data are in this file. An easy way to find this out is by putting the cursor on A1 and pressing **Ctrl + Down** arrow. Notice that the file has 25 workers. Remember the very first row is **Header** row. To find out the number of columns, press **Ctrl + Right** arrow, and it takes to the last column of the data. The data in the sheet can further be selected by pressing **Shift + Ctrl + Down** and **Shift + Ctrl + Right** keys (Figure 1.2).

1.4. FORMATTING DECIMAL PLACES

It may be noted in Figure 1.3 that the cells in landowned data have values up to different decimal places. For example, the worker with ID number 1 owns 0.7534 acres of land or value

FIGURE 1.3

Profile of Rural Workers



up to 4 decimal places, whereas worker with ID number 3 owns 2.258 acres of land which is up to 3 decimal places. If we wish to organize all the land data in a consistent manner such that all cells display values up to two decimal places, then select the entire **column** of data cells. Right click any cell in this column.

- Press **Format cells**
- Select **Number**, and choose the number of decimal places (Figure 1.4). If two decimal places are selected, data in all the cells will be transformed into two decimal places
- Press **OK**

Alternatively, as displayed in Figure 1.5, the data in the chosen cells may be converted directly into desired decimal places by choosing **Numbers** in the **Home** tab and selecting desired decimal points. Notice that all the decimal places are shortened or rounded off to two decimal places by using either of the two methods (Figure 1.6).

FIGURE 1.4

Formatting Decimal Points

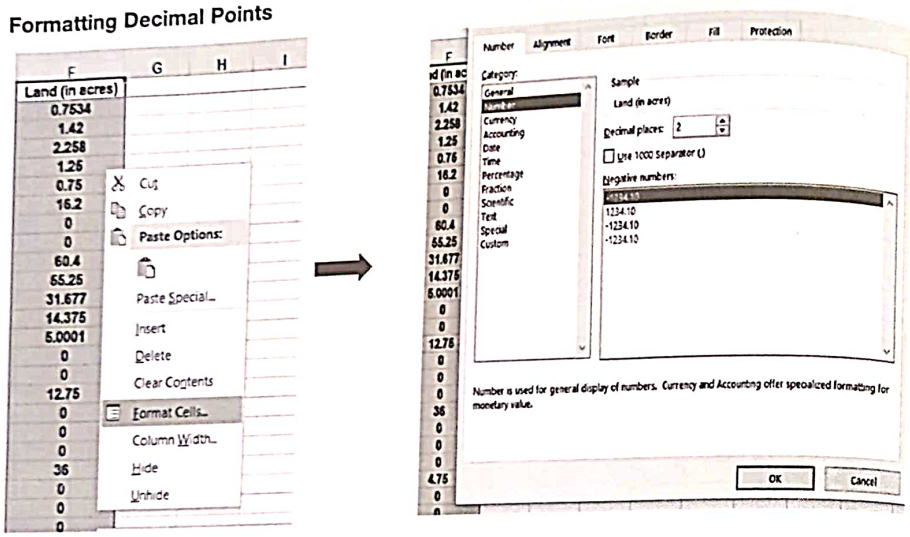
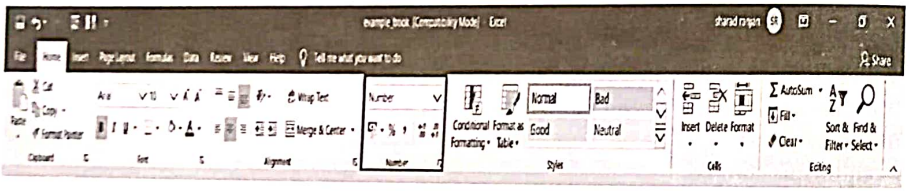


FIGURE 1.5

Formatting Decimal Points Using Home Tab

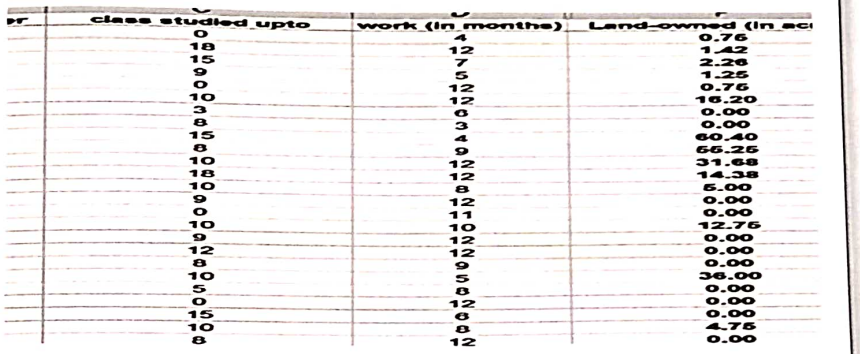


1.5. SORTING DATA

Let us **Sort** our data according to the number of months workers have worked in an ascending order. That is, we wish to list the workers who have not worked at all at the beginning of the list and those during the whole year at the end of the list. We wish to further **Sort** the workers as per their identity in Column 1 in ascending order. We start by selecting the entire data file. An easy way to select is to first place the cursor in the very first cell, which is A1 cell, press **Ctrl + Shift + Right** arrow and then **Ctrl + Shift + Down** arrow. The entire data is selected now.

FIGURE 1.6

Formatting up to Two Decimal Points

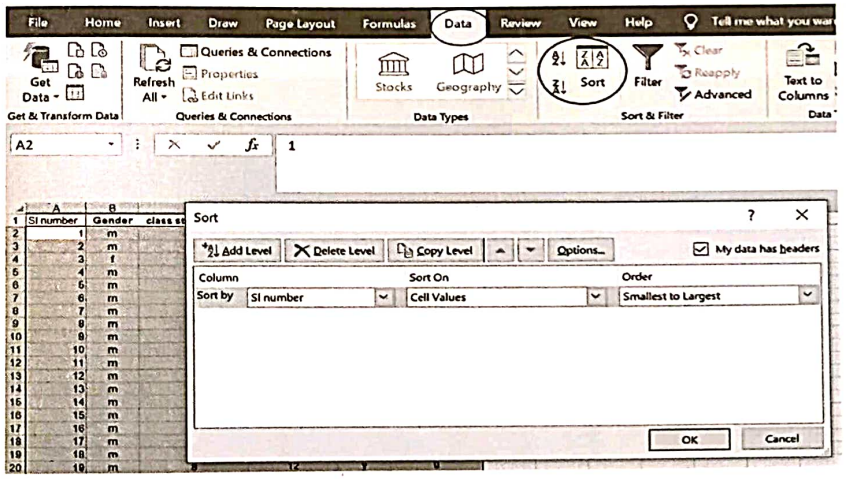


class studied upto	work (in months)	Land owned (in ac
0	4	0.75
18	12	1.42
15	7	2.26
9	5	1.25
0	12	0.76
10	12	16.20
3	6	0.00
8	3	0.00
15	4	60.40
8	9	55.25
10	12	31.68
18	12	14.38
10	8	5.00
9	12	0.00
0	11	0.00
10	10	12.75
9	12	0.00
12	12	0.00
8	9	0.00
10	5	36.00
5	8	0.00
0	12	0.00
15	9	0.00
10	8	4.75
8	12	0.00

- Click **Data**
- **Select Sort** opens up the following dialogue box (Figure 1.7)

FIGURE 1.7

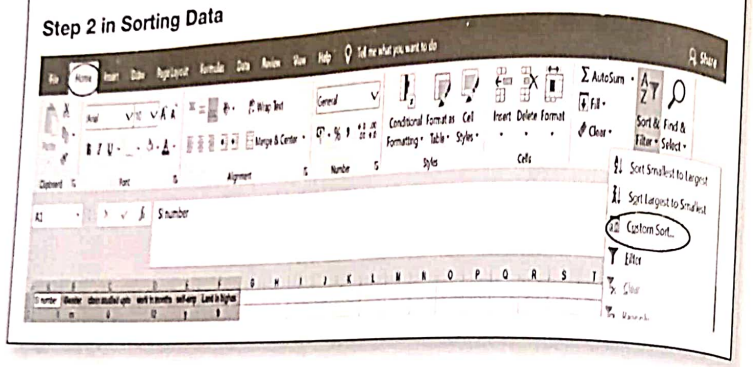
Step 1 in Sorting Data



Alternatively, the same dialogue box also opens up under *Custom Sort* in **Sort and Filter** in the **Home** tab (Figure 1.8).

FIGURE 1.8

Step 2 in Sorting Data



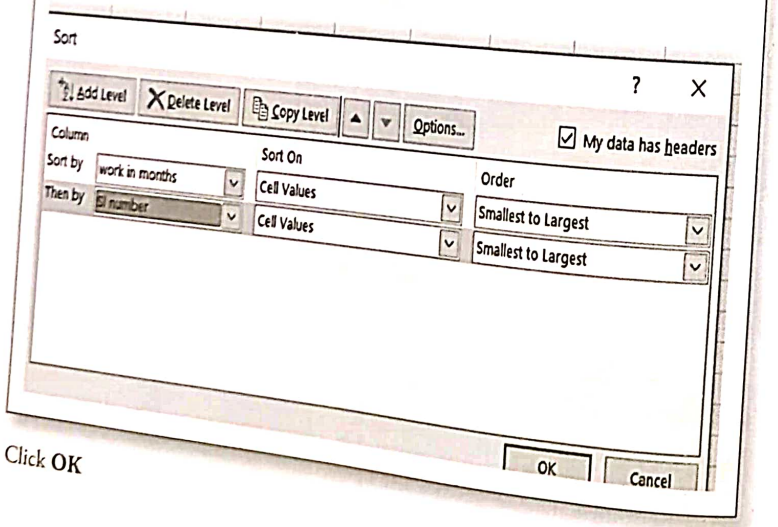
- Sort by work in months by order **Smallest to Largest**

Now we wish to **Sort** it further by worker ID.

- Click **Add Level**
- A new level (*Then by*) is added. Select ID number from the drop-down box, choose **Smallest to Largest** order because we want the serial IDs in ascending order. Since our data has headers, we need to make sure that the box with **My data has headers** is checked (Figure 1.9).

FIGURE 1.9

Step 3 in Sorting Data



- Click **OK**

Now the data has been sorted out according to the ascending order of work in months and *within* work in months, with their serial IDs again in ascending order.

1.6. UNHIDE AND HIDE ROWS AND COLUMNS

When we are dealing with a heavy spreadsheet, that is, with large number of **rows** and **columns**, it is sometimes helpful to **unhide** or **hide** some of the **rows** and **columns** to better see the information one needs to analyse. If we do not wish to see some of the **column/s** (e.g., self-employed in Figure 1.10),

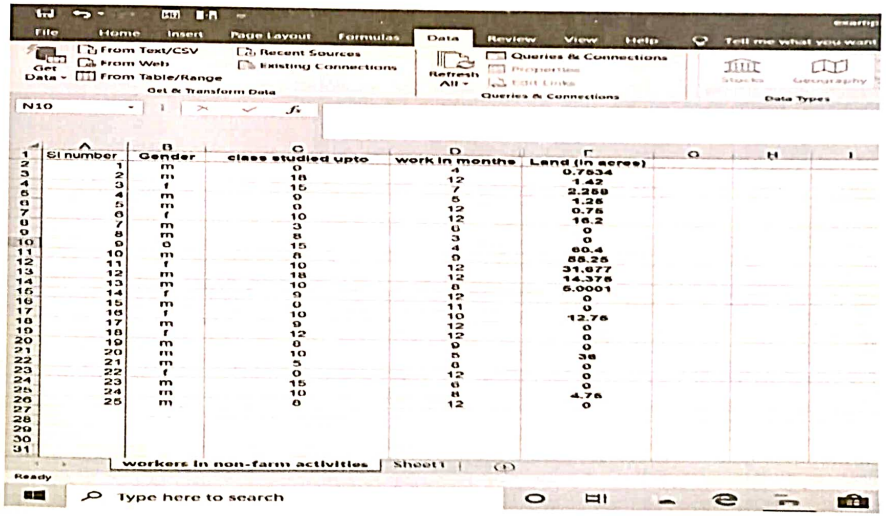
- Select the **column** you do not wish to see
- **Right click** the column
- Select **Hide**

It can be noted from the table that **Column E** is not displayed. It does not delete the **column** but simply **hide** it. To **unhide** it,

- Select both the adjacent **columns** of the hidden column
- **Right click** the column
- Select **Unhide**. The hidden data is now visible again. The same method can be followed in the case of **rows** also.

FIGURE 1.10

Hidden Column



1.7. FILTERING DATA

The **Filter** is an important tool in data analysis. It helps us in displaying data selectively, according to the criteria set by a researcher. It may be noted again that the data we have in our example is selective, but the real-life data may be huge. Now in our ongoing example, if we wish to see gender-related data, of female workers only, **Filter** needs to be applied. Similarly, other Headers can accordingly be **Filtered**. One simple course of action to apply **Filter** is to hold **Shift + Ctrl** and then press L. The columns should now have drop-down arrows next to each of the **Header** as exhibited in Figure 1.11.

Another way of doing it is by placing the cursor on any cell within the data range and select Filter in the Data tab. The outcome will be the same. One can also press the **Home** tab and select **Sort & Filter** (Figure 1.12). From the drop-down box, select **Filter**. The box has **Text Filters** and **Number Filters** if the data is in the form of text or numbers, respectively.

FIGURE 1.11

Filtering Data through Key Commands

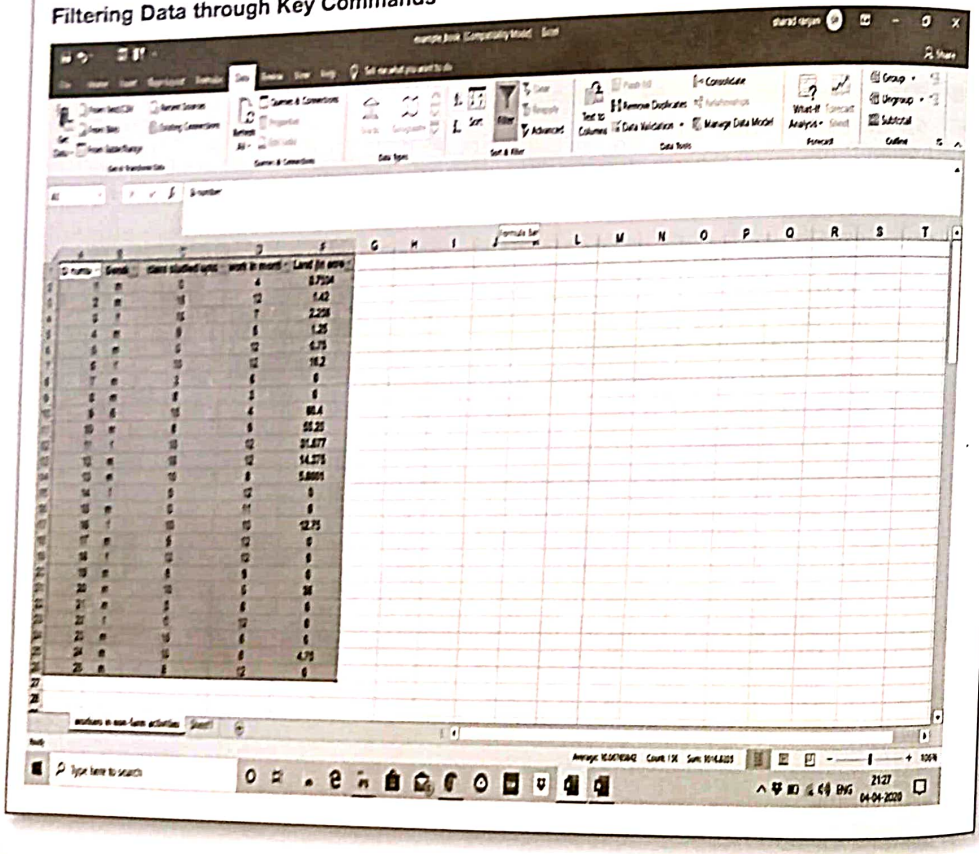
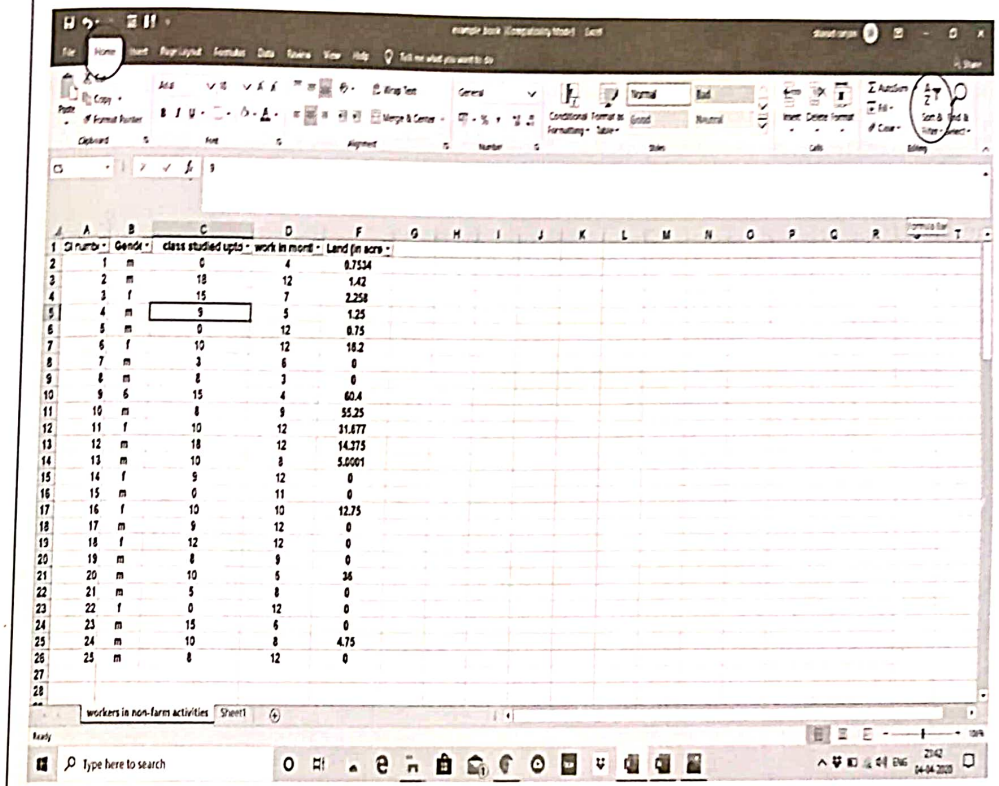


FIGURE 1.12

Filtering Data through Tabs



Now suppose we are interested in looking at gender-based data, that is, of female workers only which are employed in non-farm activities. To view it, click the drop-down arrow for the column of Gender to be filtered to display. **Uncheck** the box, **Select All** and **Check** the 'female' box again. Click **OK**. The data now displayed is of female workers only.

Additional filters can further be applied to other columns to narrow down the results. For instance, we wish to identify those female workers who have worked in all the months of the year. To do it, click on the Work in months **Filter**, **Uncheck** the box, **Select All** and **Check** the 12-month box again and click **OK**. The **Filters** can be cleared by pressing **Filter** again or **Clear** in **Data** tab.

Advanced **Filter** tools can be used for further insight into the data, both textual and numeral, according to the format of data given. For text data, click on the drop-down arrow of the column to be filtered (in our example, it is gender), click **Text Filters**, and a desired option such as **Equals**,

Does Not Equal, **Begins With**, **Ends With** and **Contains** may be chosen (Figure 1.13). A **Custom Filter** dialogue box pops up immediately after selecting the desired criteria from the option **Text filter** (Figure 1.14).

FIGURE 1.13

Filtering Data Using Text Filters

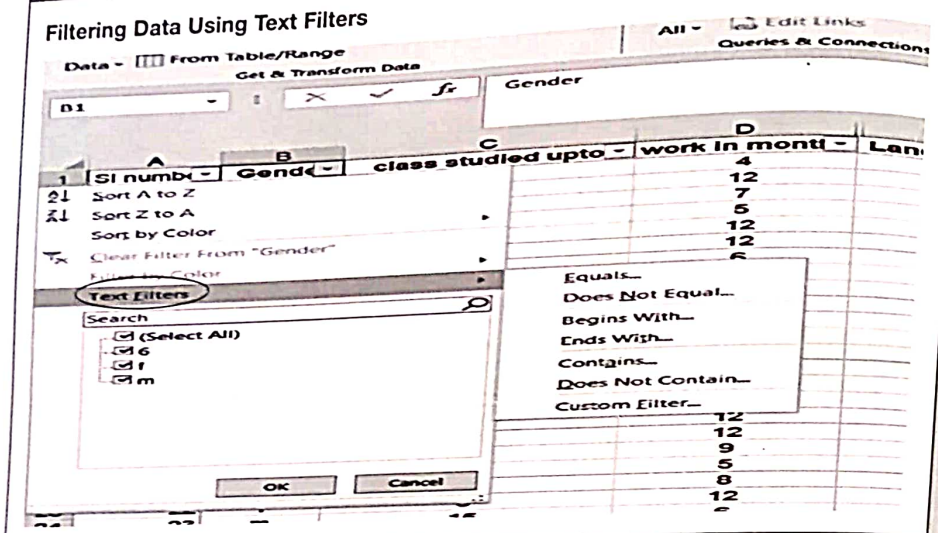
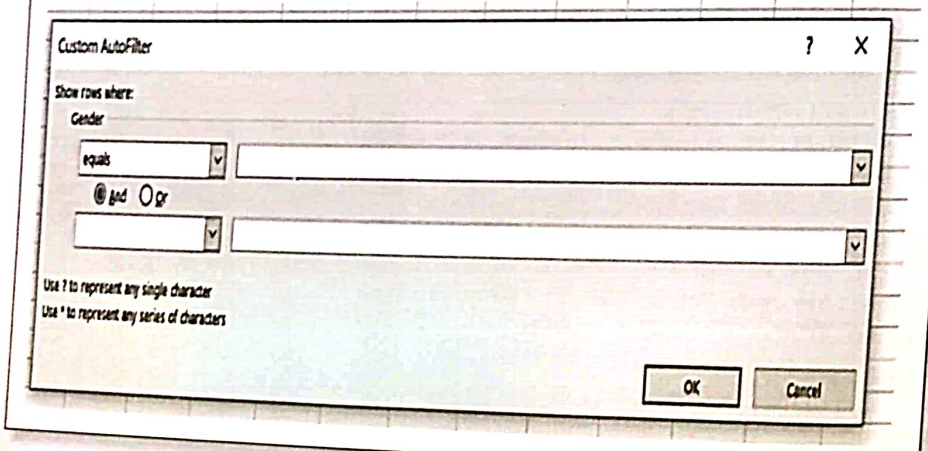


FIGURE 1.14

Custom Filter



Similarly, the relevant options in **Number Filters** such as **Equals**, **Does Not Equal**, **Greater Than**, **Greater Than or Equal To**, **Less Than or Equal To** and **Between** along with **Custom Filter** can be picked for analysis (Figure 1.15).

Once the data is filtered, we can now **copy** and **paste** the data into another sheet for deeper investigation. To do it, keep the cursor on any of the data cell, press **Ctrl + A** and again **Ctrl + C**. The data is now selected and copied. To paste it in the new sheet, open the new sheet and place the cursor on cell we want copied data to be pasted and press **Ctrl + V** (Figure 1.16). The data is now ready for scrutiny to the researcher.

FIGURE 1.15

Number Filters

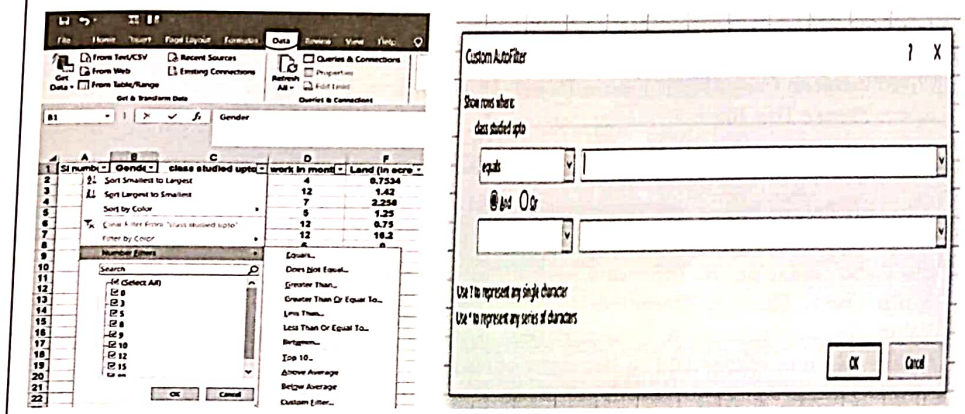
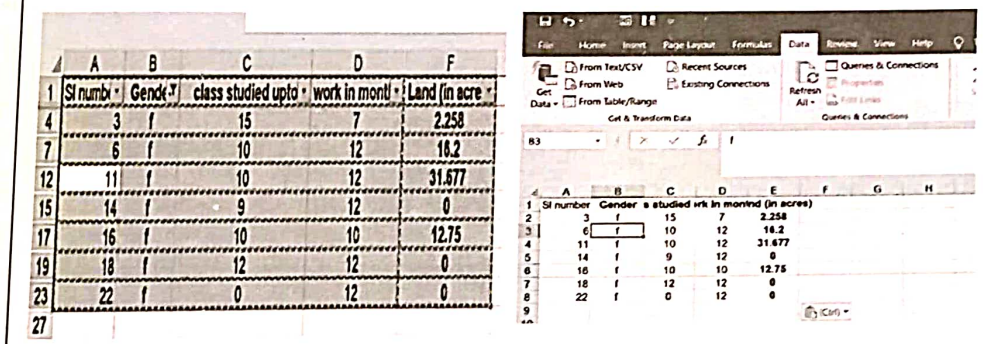


FIGURE 1.16

Copy and Paste of Filtered Data



1.8. FREEZING WORKSHEET PANES

The application of **Freezing Panes** utility is used when the size of data is large so that the complete data is not displayed at a point of time. Resultantly, the researcher tends to lose track of what each of these columns/rows belongs to while scrolling down or right. In such a situation, **Freezing Panes** is the technique to freeze the panes so that **Headers** of first **rows** and **columns** can be viewed even after scrolling across the worksheet. As shown in Figure 1.17, the steps to freeze the first **column** are given below:

- Click **View** tab
- Select **Freeze Panes**
- then from the drop-down box select **Freeze First Column**

The first **column** is **frozen** (or locked) now. Scrolling to the right, we will not lose sight of the first **column** any longer. Likewise, to **freeze** the top **row**,

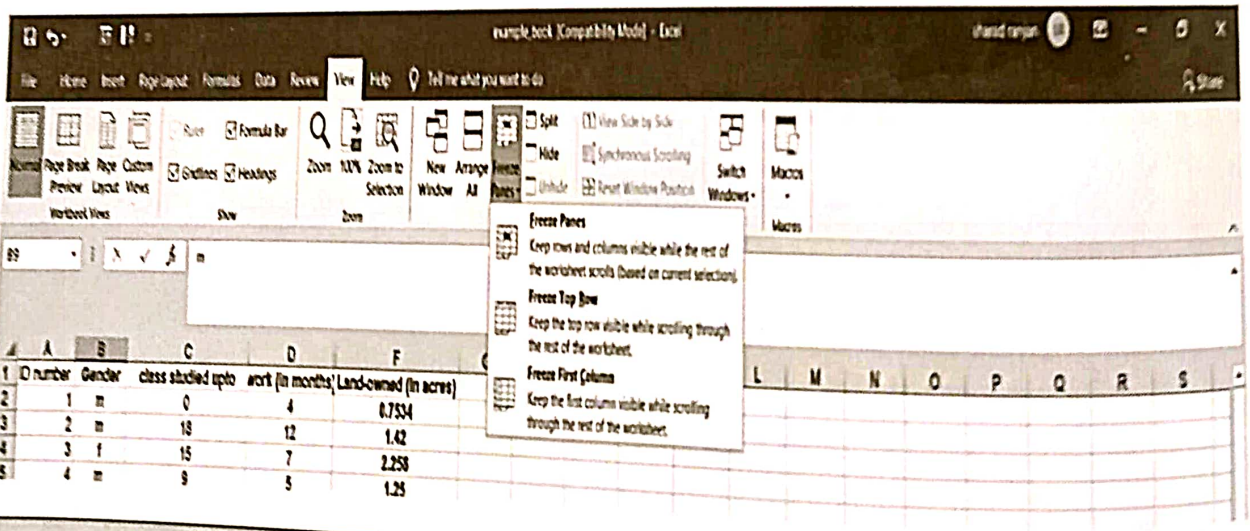
- First **Unfreeze Panes** from **Freeze Panes** drop-down box
- Select **Freeze Top Row**

Further, in a situation when we need to **freeze** both the top **row** and first **column**,

- Click on **Freeze Panes** in **View** tab
- Select **Unfreeze Panes** from the drop-down box
- Place the cursor on the intersection point of the first data point (Cell B2)
- Select **Freeze Panes** in **View** tab
- Select the first option of **Freeze Panes**. Now the first **row** and **column** will remain visible while scrolling data to the right or down.

FIGURE 1.17

Freezing /Unfreezing of Panes



With these introductory techniques, we will now learn the techniques to perform arithmetic manipulation in Excel in the next chapter.

2

CHAPTER

Basic Arithmetic Operations in Excel

LEARNING OBJECTIVES

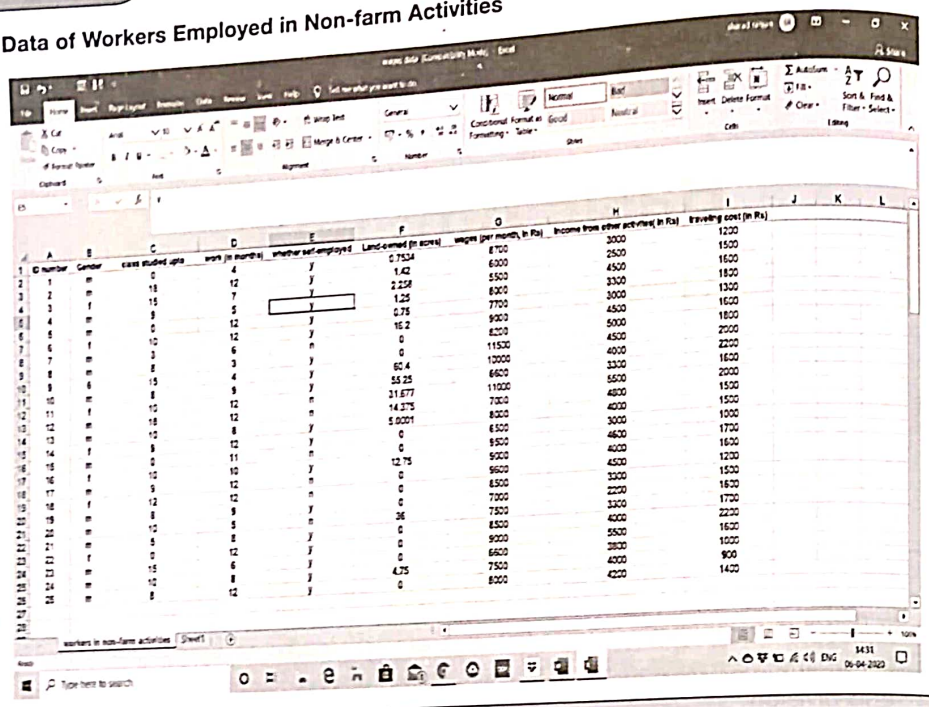
After reading this chapter, the readers will be able to understand

- Basic arithmetic operations in Excel such as addition, subtraction, multiplication and division
- Performing calculations using functions such as SUMIF, COUNT and COUNTIF, MAX and MIN
- Use of relative and absolute cell references
- Creating pivot tables in Excel

This chapter imparts an understanding of basic arithmetic operations in Excel. Such operations are necessary for undertaking any study. In particular, entering a formula in the spreadsheet to perform basic operations such as addition, subtraction, multiplication and division are introduced. Some other commands, namely **SUMIF**, **COUNT** and **COUNTIF**, **MAX** and **MIN** are also introduced. How to **Copy** and **Paste** the formulas across the spreadsheet is presented. To learn these functions, the data set of workers engaged in non-farm activities considered in Chapter 1 is used here. This time, however, additional information on their incomes and expenses are also presented as shown in Figure 2.1.

FIGURE 2.1

Data of Workers Employed in Non-farm Activities



2.1. ADDITION

We wish to determine the total income earned per month of worker with ID number 1. To figure it out, we need to aggregate his monthly wages and income earned from other sources which are given in Cells G2 and H2, respectively. Let us find the calculation in Column J and name it *total income earned*. Corresponding to G2 and H2 cells, place the cursor in Cell J2 where we want the result of the calculation to appear.

- Press the = key to start writing the formula
- Click on the first cell which is to be added (or write the name of the cell), press + (i.e., the plus sign) and click on the second cell that needs to be added (or simply write the name of the cell)
- Press **Enter** for the result

Another practical approach is the following:

- Press the **Insert Function** (f_x) key. It automatically brings in = in **Formula Bar** and a **dialogue box** pops up.

- Search for a function **SUM** or select a function straightaway from the given options (Figure 2.2).
- Select the cells intended to be **summed**. The cells may be added up by either giving a **Range** G2:H2 (as in our example) or may be entered separately in Number 2, Number 3 and so on (Figure 2.3).
- Press **OK** for the answer.

We can **Copy** the formula by selecting the cell and taking the cursor to the bottom right of the cell till we see a thin plus (+) sign and double click. Notice that the entire column gets populated by the formula displaying the total income of all the workers.

FIGURE 2.2

Step 1 in Adding of Cells

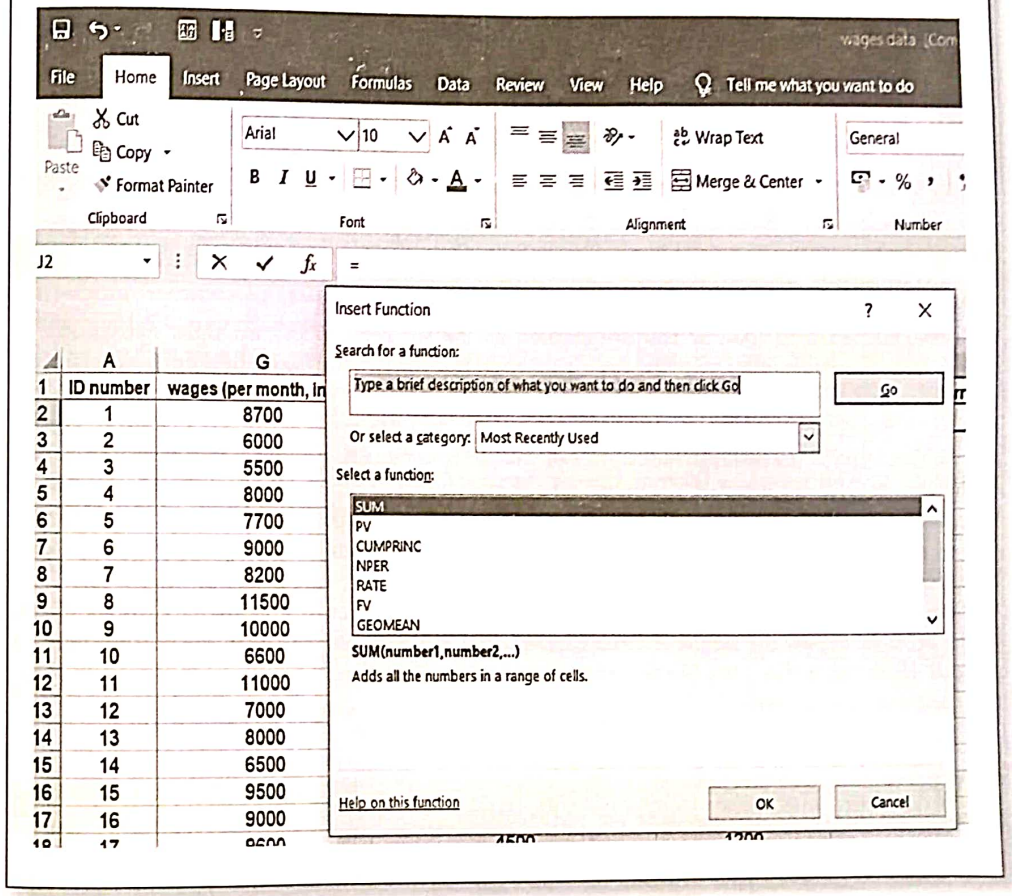
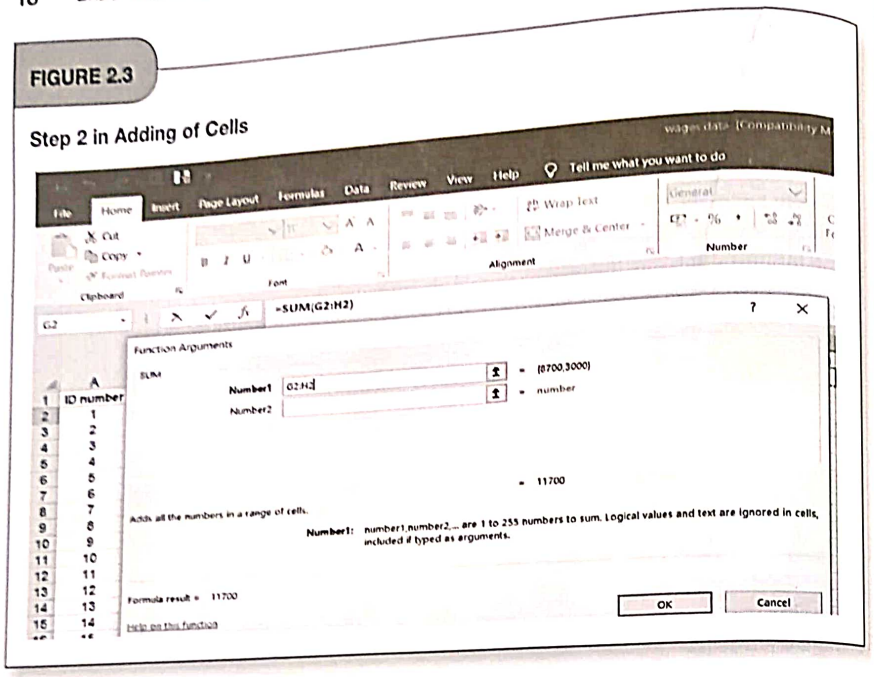


FIGURE 2.3

Step 2 in Adding of Cells



2.2. SUBTRACTION

We now may want to look at the net income of the workers. In our example, net income is obtained by subtracting the traveling cost from total income. In our Excel sheet example, the total income earned is given in Column J. So to find out, first, the net income of worker with ID number 1, we place the cursor in Cell K2, where we want the result of the calculation to appear and we then type '=' to begin the formula. We may next write either J2 (where total income earned of the worker with ID number 1 is given) straightaway or select the Cell J2 and then put a minus (–) sign. Write I2 (which represents the traveling cost) or select the cell I2. Press **Enter**. This will lead to the answer we are looking for. Thus, the formula comes out as **=J2–I2** (Figure 2.4). It may be noted that there is no **SUBTRACT** function in Excel.

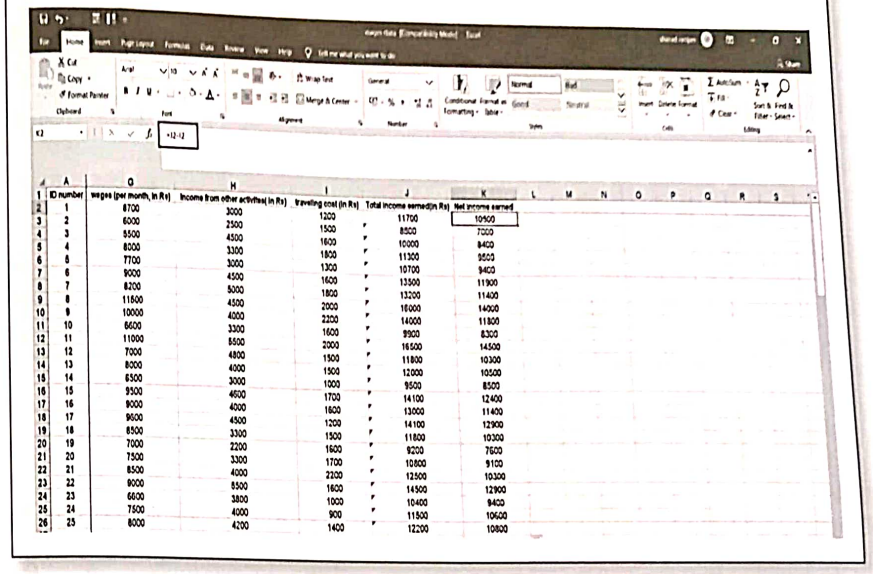
To know the net income earned of the rest of the workers, that is, to copy the formula in rest of the cells, we follow the step mentioned earlier, that is, taking the cursor to the bottom right of the cell till we see a thin plus (+) sign and then double clicking on it. Thus, we get the result for the remaining cells as well.

2.3. MULTIPLICATION

In addition to the above information attained, one may be curious to know the total income earned by these workers in a year. To arrive at final income earnings in a year, we need to multiply

FIGURE 2.4

Subtracting of Cells



the monthly income with the number of months they worked in a year. It can be seen that the data of the number of months worked by a worker is given in Column D. The worker with ID number 1 worked for four months a year, the worker with ID number 2 worked round the year, the worker with ID number 3 for seven months a year and so on. The total income earned per year is specified in Column J, as shown. Multiplication of cells of Columns D and J will give us the annual income levels. The results of the annual income of all workers are to be presented in cells of L.

For the first worker (with ID number 1) whose work in months and total income earned are placed in Cells D2 and J2, respectively, we place the cursor on the corresponding Cell L2.

- In this L2 cell, type '=', then click on the cell that contains the first number we want to multiply.
- Type '*' and click on the second cell that we want to multiply.
- Press **Enter** for the results (Figure 2.5).

The second method to approach it is as follows:

- Press the **Insert Function** (f_x) key. It automatically brings '=' in the **Formula Bar** and a **dialogue box** shows up
- Select category **Math & Trig**
- Select the function **Product** and
- Press **OK**. It leads to another pop up as shown in Figure 2.6

FIGURE 2.5

Method 1 of Multiplication of Cells

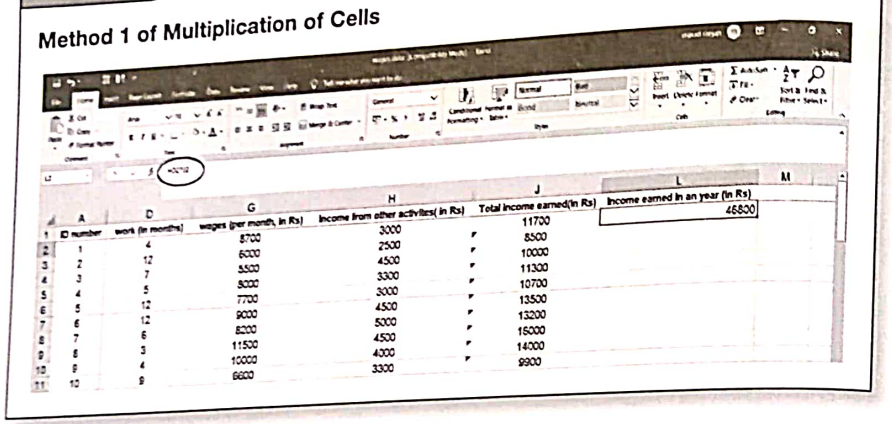
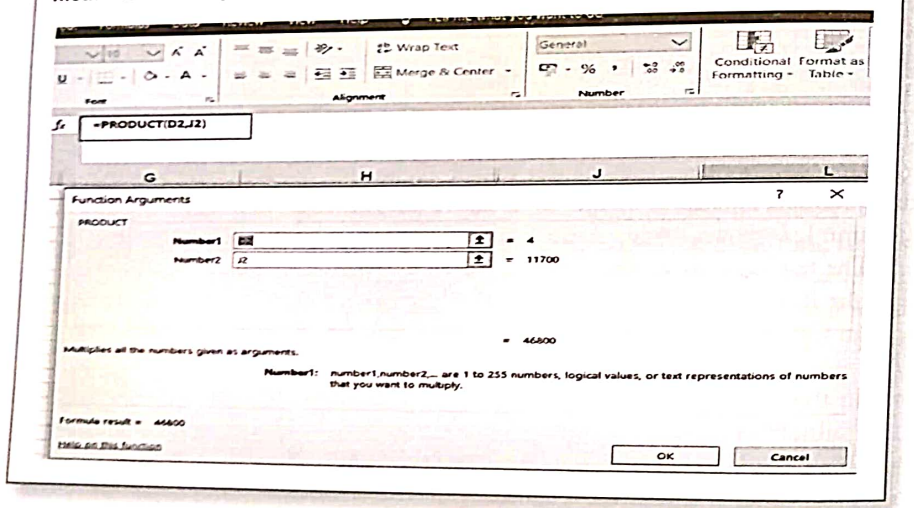


FIGURE 2.6

Method 2 of Multiplication of Cells



- In the **Number 1** and **Number 2** boxes, enter the numbers to be multiplied respectively. We can now see that the worker with ID number 1 collected ₹46,800 in an year.
- It may be noted that if multiplication figures are more than two, the **Number** boxes can be extended accordingly.
- Press **OK** for the report.
- We can **Copy** and **Paste** the formula down in the same manner followed earlier.

2.4. DIVISION

To find out the average monthly wages of the entire workforce in our data, the annual income earned by the worker needs to be divided by the total number of months in an year, that is, 12. To carry out this calculation, in our data set, cells of Column L2 are to be divided by respective cells in Column D2.

For this, we first select the column in which such calculations are to be pursued (Column M in our example). Place the cursor in cell M2, corresponding to the data figure of the worker with ID number 1. In our earlier illustrations, an arithmetic operation was carried out by first pressing '=' sign. We follow the same method here. Select the cell L2 which represents the total income earned in an year and press the forward slash (/) sign which does the job of division. We divide it by cell D2 representing the number of months worked in an year by the worker and press **Enter**.

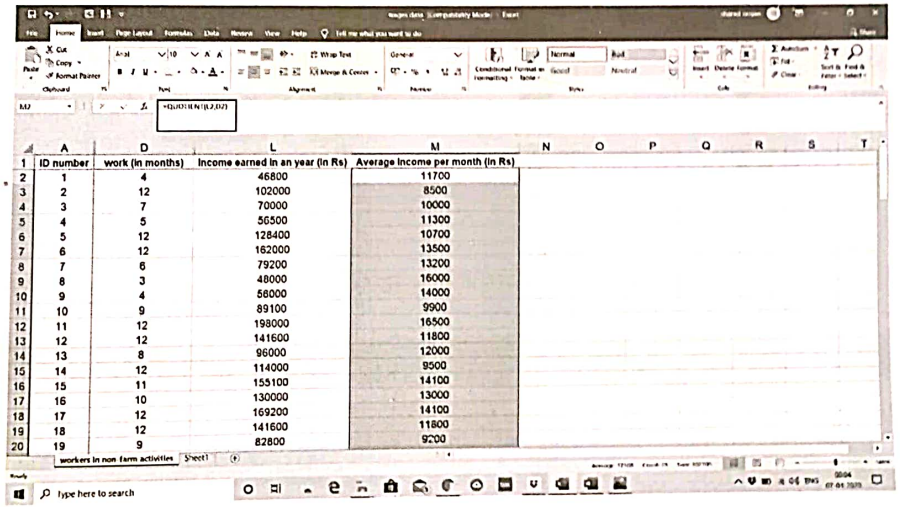
We can now see that the average income earned by the worker with ID number 1 is ₹11,700. The formula for the remaining cells can be copied accordingly. It may be noted that there is no **divide** function in Excel. However, if we are interested only in the integer portion of a division and discard the remainder, then we can use the **QUOTIENT** function in **Insert Function**.

> **QUOTIENT (numerator, denominator),** where
> the numerator (required)—the dividend, that is, the number to be divided and
> the denominator (required)—the divisor, that is, the number to divide by.

When two numbers divide evenly with no remainder, the division symbol and a **QUOTIENT** formula return the same result. For example, both **=46,800/4** and **=QUOTIENT(46,800, 4)** return 11,700 (Figure 2.7).

FIGURE 2.7

Division Function in Excel



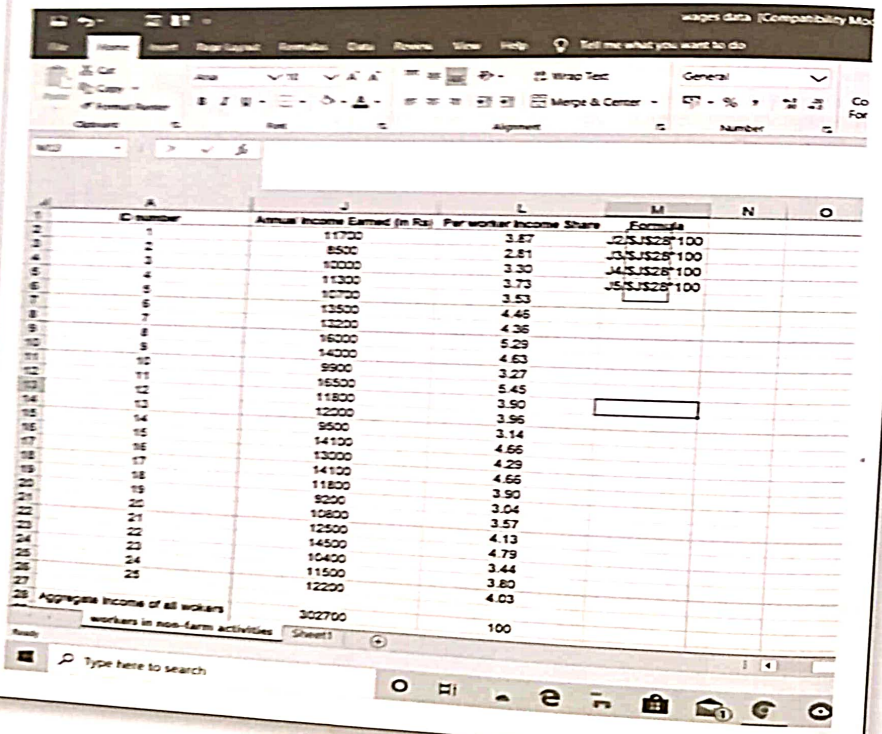
2.5. RELATIVE AND ABSOLUTE REFERENCES

A formula containing a cell address is called a cell reference. For example (=B1) refers to the value of the current cell, that is, the value of B1. So far, the calculations performed in our examples by **Copy** and **Paste** were *relative*. They were relative in the sense that after copying the function, the latter cell takes the *relative* reference of the preceding cell. For example, the Division function used for calculating average income per month for the worker with ID number 1 takes the formula M2 '=L2/D2'. Similarly, for the worker with ID number 2, the same formula applies in cell M3 (=L3/D3) with incremental row number.

However, when we talk about *absolute* cell reference, we fix a particular cell by inserting a **$** sign and then carry out calculations. In Figure 2.8, we examine per worker income share (in Column M) relative to aggregate income (in Cell J28). To determine it, we first calculate the aggregate

FIGURE 2.8

Absolute Cell Reference



ID number	Annual Income Earned (in Rs)	Per worker Income Share	Formula
1	11700	3.87	=J2/$J$28*100
2	8500	2.81	=J3/$J$28*100
3	10000	3.30	=J4/$J$28*100
4	11300	3.73	=J5/$J$28*100
5	10700	3.53	
6	13500	4.46	
7	13200	4.36	
8	16000	5.29	
9	14000	4.63	
10	9900	3.27	
11	16500	5.45	
12	11800	3.90	
13	12000	3.96	
14	9500	3.14	
15	14100	4.66	
16	13000	4.29	
17	14100	4.66	
18	11800	3.90	
19	9200	3.04	
20	10800	3.57	
21	12500	4.13	
22	14500	4.79	
23	10400	3.44	
24	11500	3.80	
25	12200	4.03	
Aggregate income of all workers	302700	100	

income of all workers. The next step is dividing an individual's annual income while keeping aggregate income constant ($J$28) as seen in Figure 2.6. In the absence of the $ sign, the formula in cell M3 will become '=J3/J29*100', which will be meaningless. By carefully going through Column M, we can figure out the denominator being fixed with the help of $ sign.

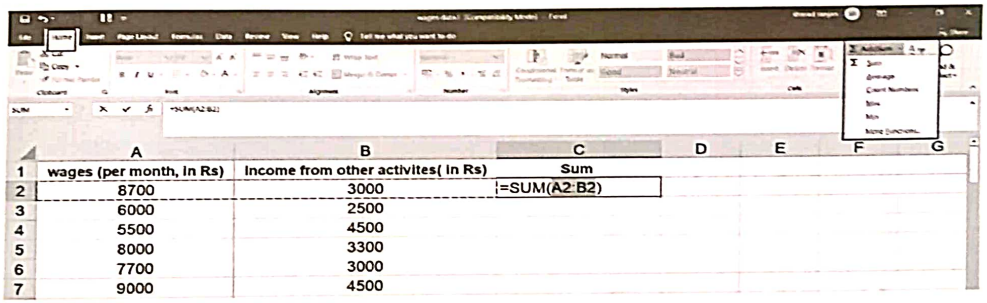
2.6. AUTOSUM FUNCTION

One of the basic functions used in Excel is the **SUM** function. Calculating the sum of many values can often be a complex task. We can use the **SUM** function to ease the task. There are many ways to perform this function. In our ongoing discussion on Excel data of workers employed in non-farm activities, we want to find the sum of the incomes of workers given in Columns B and C by using the SUM function. Place the cursor in cell C2, corresponding to the data of Worker 1. Now type =**SUM** and select the cells to be summed up. Press **Enter**.

An alternative easier way is also depicted in Figure 2.9. To operate this, place the cursor in the cell where we desire the answer to appear (e.g., C2). Select **AutoSum** in **Home** tab as is seen in the figure and press **Enter**. This function automatically selects the cells to be added. It may be appropriate to note here that **AutoSum** function also suggests other functions such as **Average**, **Count Numbers**, **Max** and **Min** number in the data range.

FIGURE 2.9

Usage of the AutoSum Function



	A	B	C
1	wages (per month, in Rs)	Income from other activites(in Rs)	Sum
2	8700	3000	=SUM(A2:B2)
3	6000	2500	
4	5500	4500	
5	8000	3300	
6	7700	3000	
7	9000	4500	

2.7. SUMIF FUNCTION

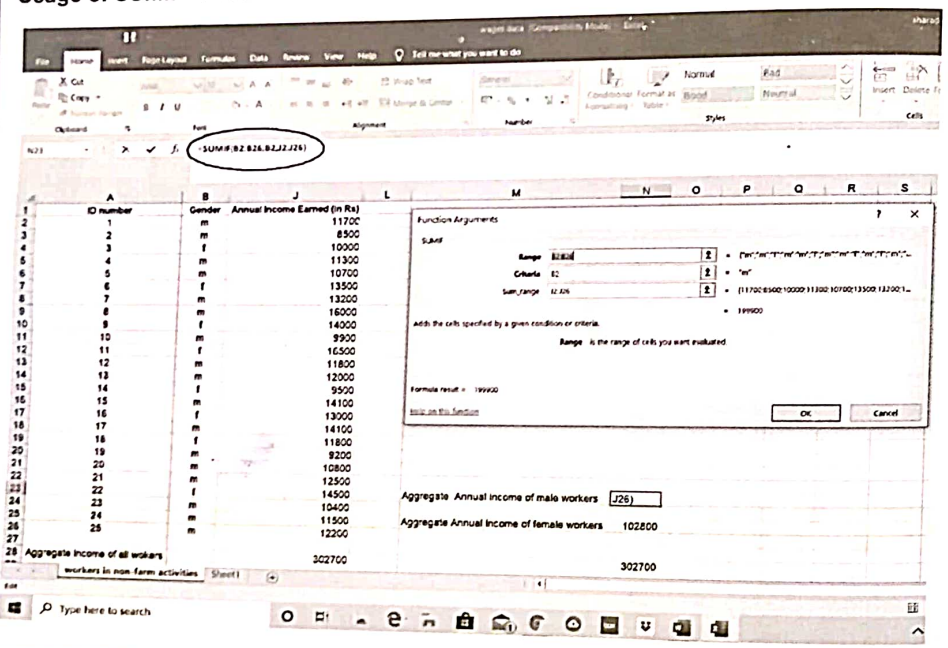
We can now calculate the total of aggregate annual income earned by all the workers using the **Sum** function. In our data set, this figure turns out to be ₹302,700. Our example categorically also specifies the gender-related data.

We now want to find out the aggregate income of male and female workers separately. To do this, we use the **SUMIF** function. The **SUMIF** function adds all the numbers in a range of cells based on criteria (e.g., criteria being male workers). To carry this out, we type **=SUMIF** in the cell where we want our result. This command in parenthesis requires three inputs as seen in Figure 2.10.

- The first one is the **Range** of cells we want to evaluate. In our example, it is Column B (from B2 to B26) as shown in Figure 2.10.
- Next is the **Criteria**. It is the condition or criterion in the form of a number, expression or text that defines which cells will be added. We define the **Criteria** m for males and f for female workers. Notice that if the **Criteria** is text, it automatically is taken in *double quotation marks* (or if we have to write a formula directly, we need to put the **Criteria** in *double quotation marks*). However, numbers do not need to be wrapped in quotes, but they can be.
- The last input is **Sum_range**, which are the data cells to be summed. In our example, Sum_range is J2 to J26 (J2:J26). This results in ₹199,900, which is the aggregate annual income of male workers. The same exercise can be undertaken for female workers, which according to our data, sums up to ₹102,800. The addition of male and female annual income wages will aggregate to ₹302,700. It is the same figure obtained by summing that of both male and female workers.

FIGURE 2.10

Usage of SUMIF Function



2.8. MAX AND MIN FUNCTION

There are many functions in **Excel** which can be found helpful in research and business statistics. Two such functions are **MAX** and **MIN** functions. The **MAX** function gives the largest value in the set of data. Similarly, the **MIN** function provides the minimum value in a data set. The syntax (which is the layout and order of the function and its arguments) is **=MAX** and the data range. In our data set of annual income earned, the maximum annual income earned is ₹16,500 (Figure 2.11) and the minimum figure is ₹8,500 (Figure 2.12).

2.9. COUNT AND COUNTIF FUNCTION

As the name suggests, the **COUNT** function counts the number of cells in the data range consisting of *numeral figures only*. The function starts with **=COUNT**, and then it requires the data range to be selected which then provides us with the count of numeral figures. In the Figure 2.13, the income information of three workers are unknown (given in the text). We can see that the output after applying the formula turns out to be 22. The output has not considered text values.

FIGURE 2.11

The MAX Function

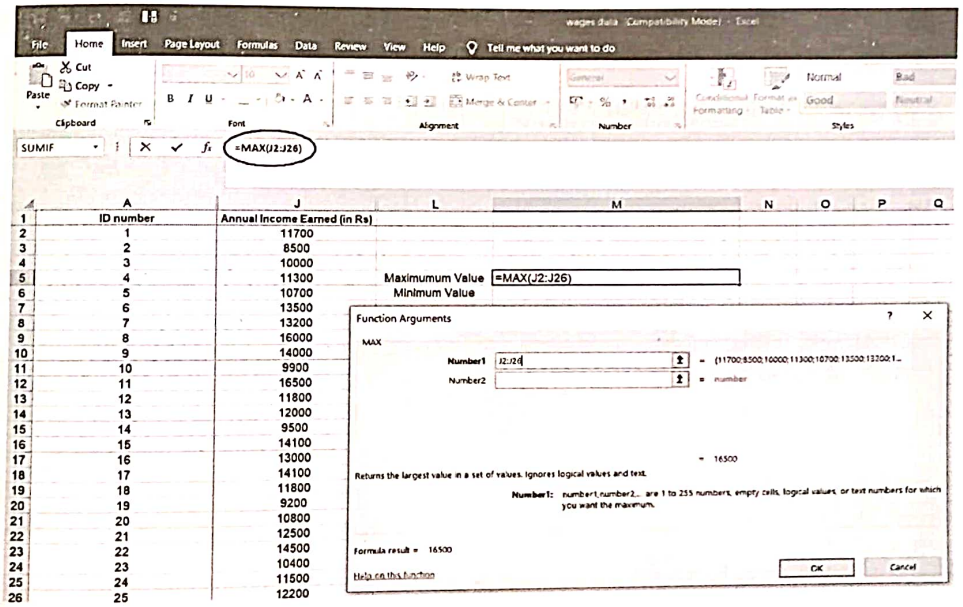


FIGURE 2.12

The MIN Function

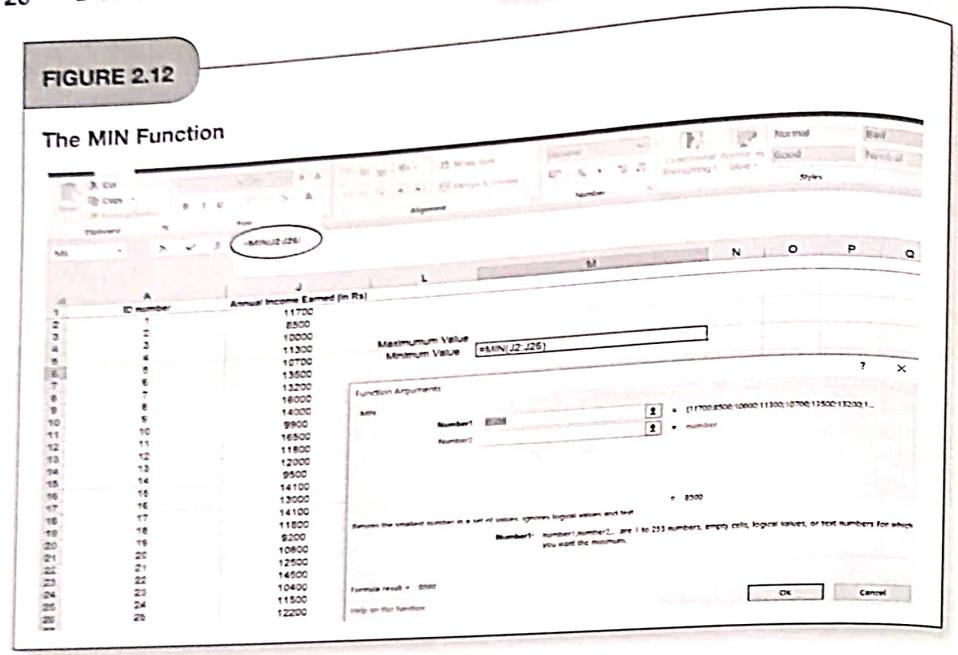


FIGURE 2.13

The COUNT Function

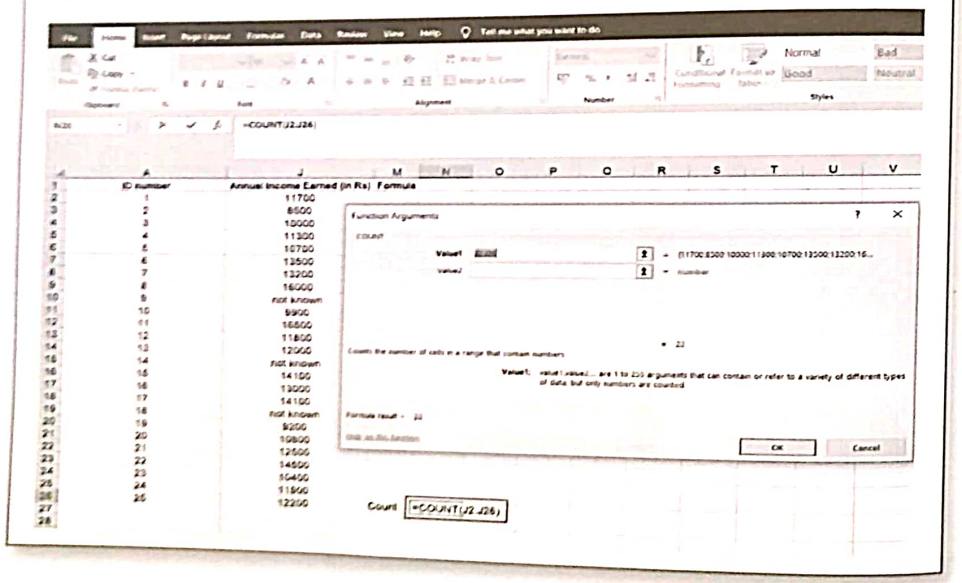
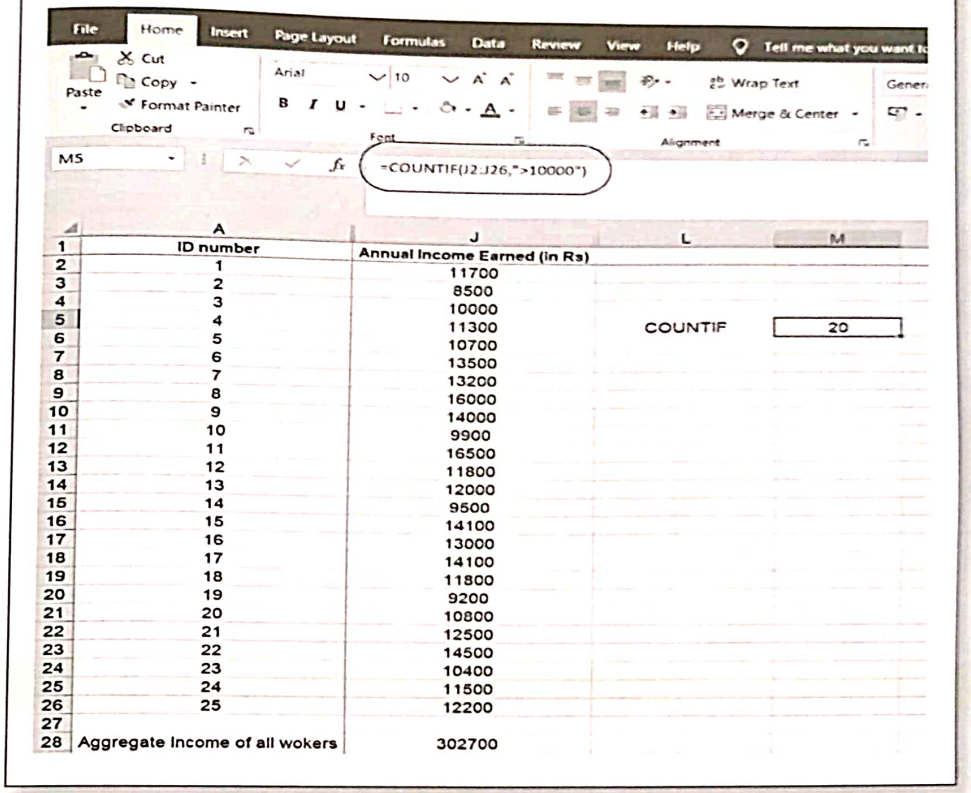


FIGURE 2.14

The COUNTIF Function



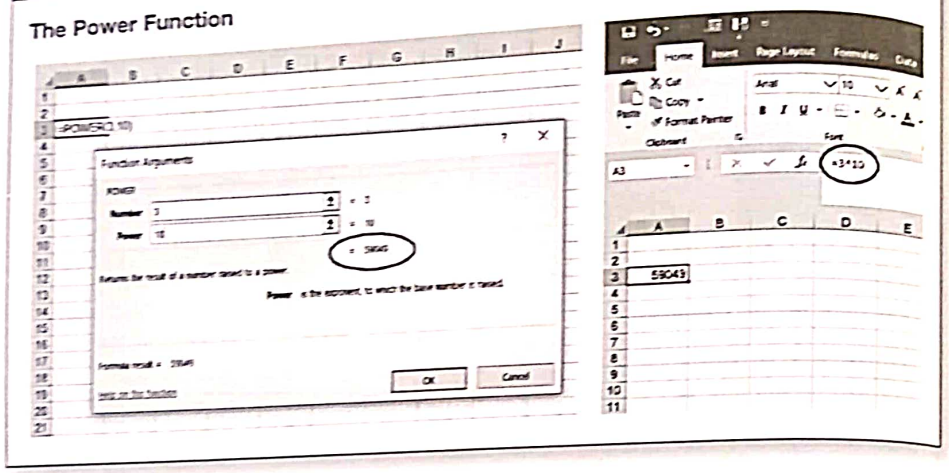
The **COUNTIF** function is an extension of **COUNT** function and is of considerable use. It counts the numbers based on the **criteria**. For example, we want to find out numbers which are greater than 10,000. We apply =**COUNTIF**, select the data range and the **criteria** is fixed as greater (>) than 10,000. The answer we get is 20. That means, in a given range, there are 20 numbers which are greater than 10,000 (Figure 2.14).

2.10. POWER FUNCTION

The **Power** function computes and returns the result of a number raised to a power. **Power Excel function** takes two arguments. The first one is a base (any real number), and the second one is an

FIGURE 2.15

The Power Function



exponent (power, which signifies how many times the given number will be multiplied by itself). For example, 3 raised to the power of 10 is explained in Figure 2.15. The answer is arrived by typing =**Power**, writing 3 in **Number box** and 10 in **Power box** which is equal to 5,909. Alternatively, it can also be done by typing =3^ (called hat) and typing 10, that is, to the number to which it is raised. Hence, we get the same result by using these two approaches.

2.11. PIVOT TABLES

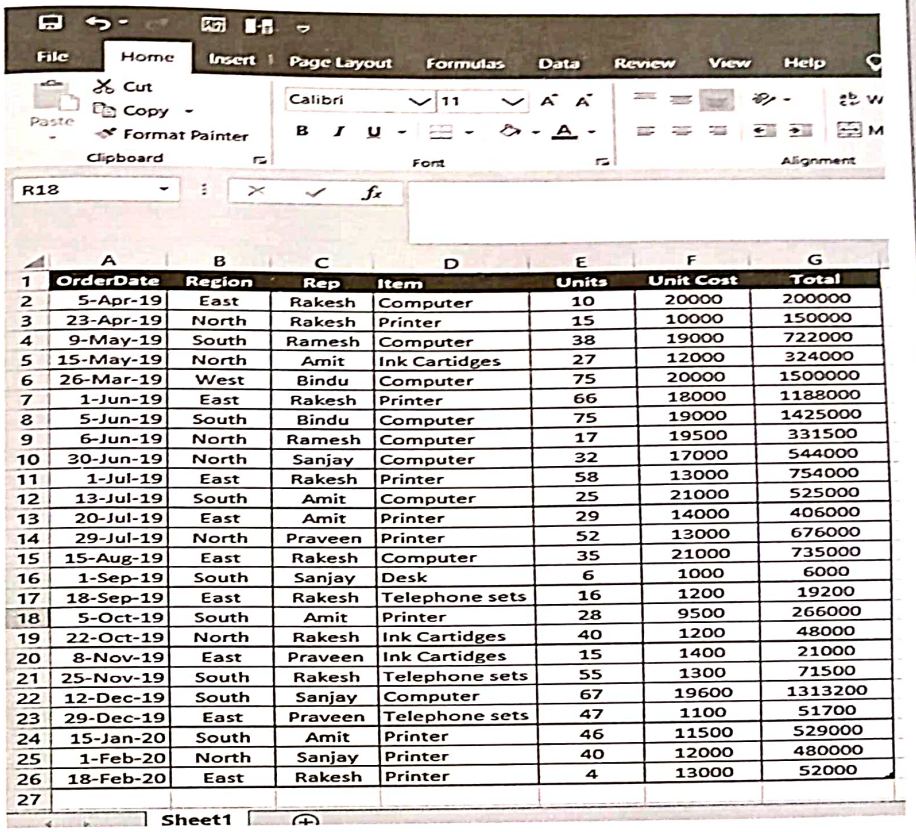
We now introduce a **Pivot Table** which is a very powerful tool for summing up data. We will understand here what a **Pivot Table** is, its uses and how to construct one.

Pivot Table is a data summarization tool that helps to concise the data and makes it more comprehensive. In this, we can view data summaries in various formats as per our requirement. **Pivot Tables** are easy to create and use. **Pivot Tables** are also dynamic. It is dynamic in the sense that as data changes in the data file, the summary in the **Pivot Tables** automatically gets updated.

We illustrate it with an example of a fictitious company that supplies various articles through various representatives in different regions in the year 2019–2020. It may be noted that the size of the currently used data is a limited one, whereas the real-life data may be a big file running into scores of **rows** and **columns**. The data presented in Figure 2.16 does not convey any meaning per se. We may draw some useful information from the data, like a summary of all items dispatched in various regions in the year mentioned using the **Pivot Table**. For this summary table, which needs to be drawn, the columns may represent the items dispatched and the rows represent the regions. We will now use **Pivot Tables** to create such a summary.

FIGURE 2.16

Sales Data of a Company



	A	B	C	D	E	F	G
1	OrderDate	Region	Rep	Item	Units	Unit Cost	Total
2	5-Apr-19	East	Rakesh	Computer	10	20000	200000
3	23-Apr-19	North	Rakesh	Printer	15	10000	150000
4	9-May-19	South	Ramesh	Computer	38	19000	722000
5	15-May-19	North	Amit	Ink Cartidges	27	12000	324000
6	26-Mar-19	West	Bindu	Computer	75	20000	1500000
7	1-Jun-19	East	Rakesh	Printer	66	18000	1188000
8	5-Jun-19	South	Bindu	Computer	75	19000	1425000
9	6-Jun-19	North	Ramesh	Computer	17	19500	331500
10	30-Jun-19	North	Sanjay	Computer	32	17000	544000
11	1-Jul-19	East	Rakesh	Printer	58	13000	754000
12	13-Jul-19	South	Amit	Computer	25	21000	525000
13	20-Jul-19	East	Amit	Printer	29	14000	406000
14	29-Jul-19	North	Praveen	Printer	52	13000	676000
15	15-Aug-19	East	Rakesh	Computer	35	21000	735000
16	1-Sep-19	South	Sanjay	Desk	6	1000	6000
17	18-Sep-19	East	Rakesh	Telephone sets	16	1200	19200
18	5-Oct-19	South	Amit	Printer	28	9500	266000
19	22-Oct-19	North	Rakesh	Ink Cartidges	40	1200	48000
20	8-Nov-19	East	Praveen	Ink Cartidges	15	1400	21000
21	25-Nov-19	South	Rakesh	Telephone sets	55	1300	71500
22	12-Dec-19	South	Sanjay	Computer	67	19600	1313200
23	29-Dec-19	East	Praveen	Telephone sets	47	1100	51700
24	15-Jan-20	South	Amit	Printer	46	11500	529000
25	1-Feb-20	North	Sanjay	Printer	40	12000	480000
26	18-Feb-20	East	Rakesh	Printer	4	13000	52000
27							

We begin by inserting a **Pivot Table** template given in the **Insert** tab. Excel asks for the data for which we wish to create a **Pivot Table**. We select the data from cell A1 to cell G26 (Figure 2.17). Then Excel asks us if where we want the **Pivot Table Report** to be placed—New Worksheet or Existing Worksheet—and then press **OK** after selecting it. We, for example, select our data to be placed in a New Worksheet (Figure 2.18). We may rename the New Worksheet as a **Pivot Table**.

In the left-hand side area, Excel has a template of the **Pivot Table**. We may also notice that on the right-hand side, fields of the various columns in the data table are also given. However, if we place the cursor anywhere outside the **Pivot Table**, the fields disappear but appear again as soon as we are back in the **Pivot Table** area. At the bottom, we have certain tools that allow us to organize our **Pivot Table Report**. Since we wish to find the data summary of items dispatched in various regions, we

FIGURE 2.17

Inserting a Pivot Table

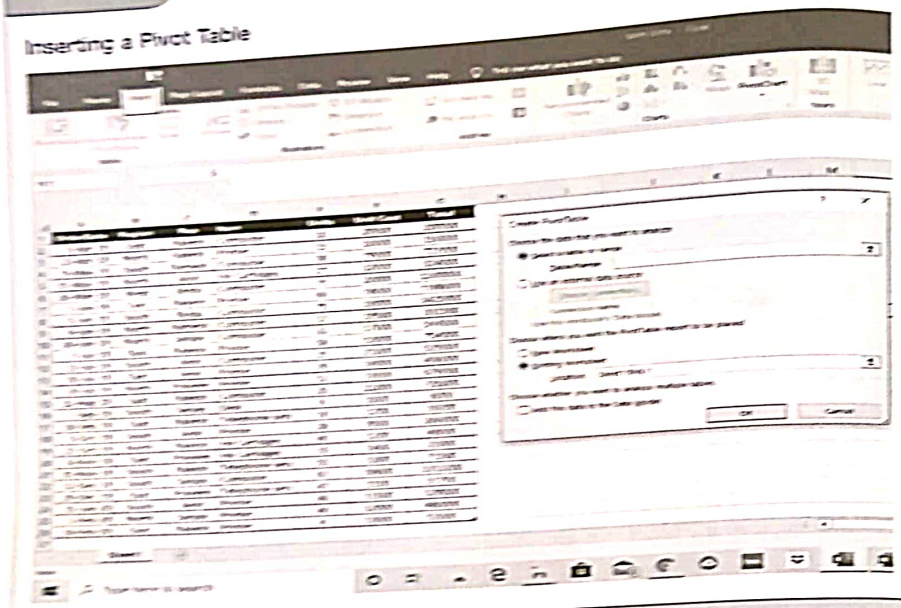
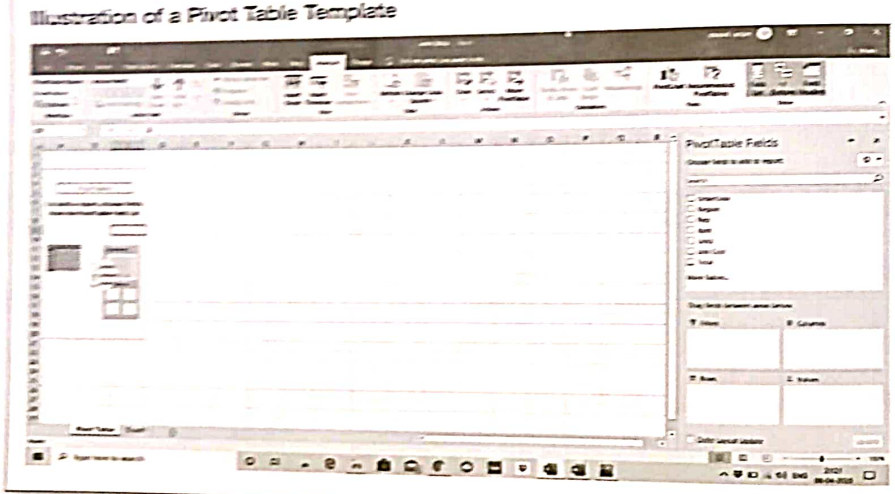


FIGURE 2.18

Illustration of a Pivot Table Template



select the items and drag them to the **column box**. Similarly, we select the region and drag it to the **rows**. Since we wish to analyse the items dispatched, we again drag the items to the **Values** field. Excel has now categorized the data. We can now make it out that in the

- Eastern region, 9 items were dispatched
- North Region, 7 items were dispatched
- South Region, 8 items were dispatched
- West Region, 1 item was dispatched

and the other details can be figured out from the given Pivot Table (Figure 2.19).

We can make further processing in the summarized data reported in the **Pivot Table**. For instance, we may wish to see the details of items dispatched region-wise for the first quarter only. We can achieve this by dragging the order date, in the **Pivot Table Fields**, to the **Filter** box. This gives the option of **Filter** on the top of the **Pivot Table**. Check the box **Select Multiple Items** first and uncheck the boxes mentioning dates not falling in the first quarter. Click **OK**. The **Pivot Table** report has now been filtered and it reports the figures for the first quarter only (Figure 2.20).

We can add yet another filter to the report by placing it in the **Report Filter area** as in Figure 2.21. Further, if we wish a **Pivot Table** to report sales figures by the name of representatives, we can drag the variable *representative* from the **Pivot Table Field** to the **Filter** field. Another filter box appears below the former **Filter** and we select the name of the representative in whose sales figures we are interested in following the same procedure just mentioned. The **Filter** can also be removed by selecting and dragging the variable from the *Column* field back to the **Pivot Table Fields**.

FIGURE 2.19

Summary of Results

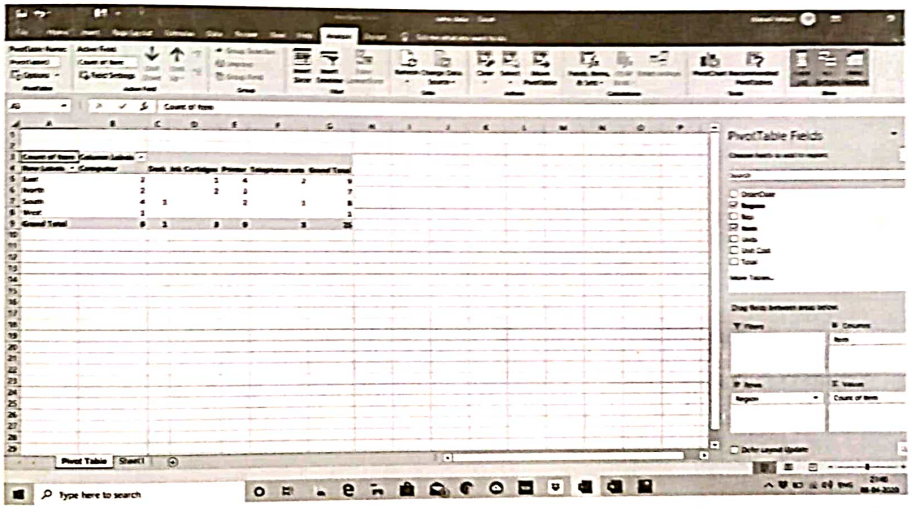


FIGURE 2.20

Step 1 of Data Filtering in Pivot Table

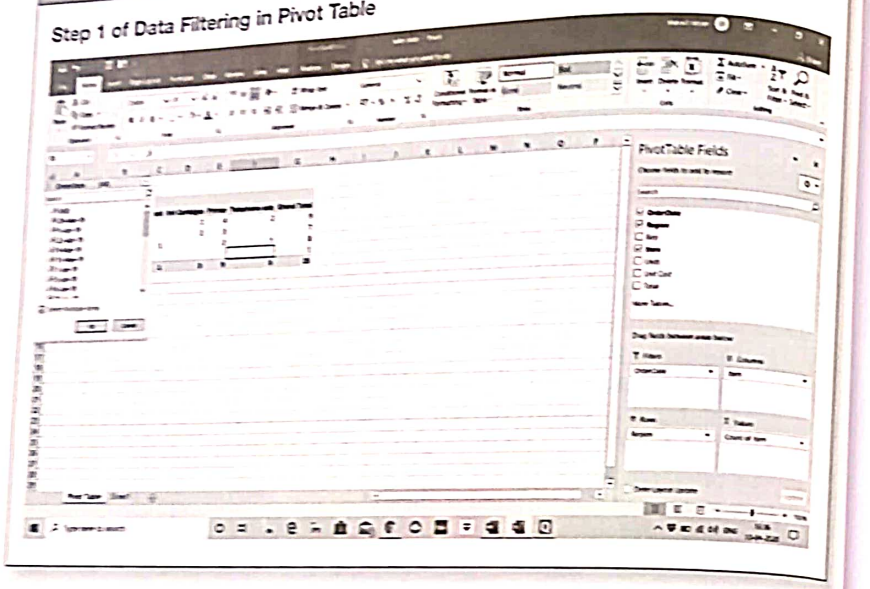
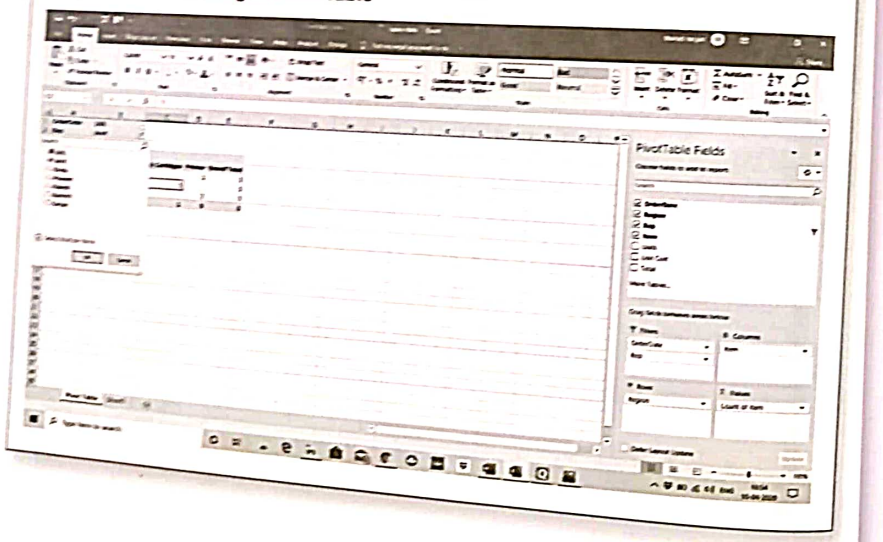


FIGURE 2.21

Step 2 of Data Filtering in Pivot Table



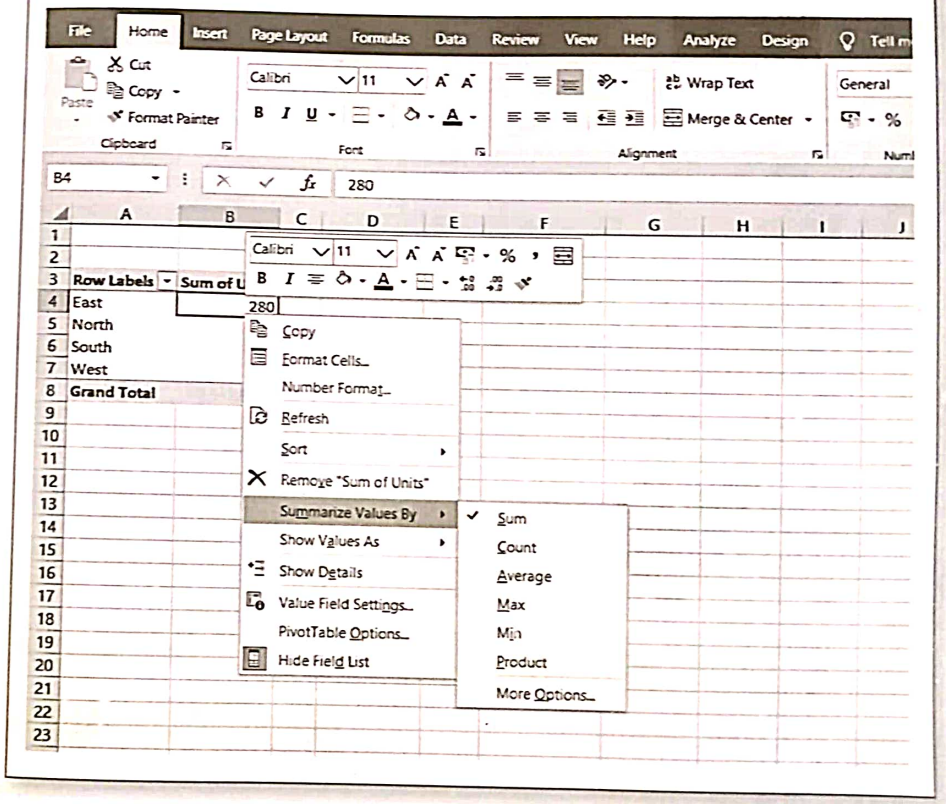
Likewise, dragging the representative field to the **Rows** area will show the region-wise sales by different representatives. The **Pivot Table** also shows both, the total of the sales and the individual sales by the respective representative.

We can also evoke arithmetic operations in the study of **Pivot Tables**. The numeric data allows us to do a lot of arithmetic calculations in the **Pivot Table** such as Sum, Count, Average, Product and many more.

In the ongoing example, to find the **Sum** of the total number of dispatched items region-wise, we drag the region **Field** to the **Row** box and units **Field** to the **Values** box. It then automatically shows the sum of the items dispatched region-wise (Figure 2.22). If we again wish to see the number of maximum units dispatched in the same function, we click anywhere on the **Sum of Units** column, right click and select **Summarize Values By** option and at last select Max. We can also choose Count, Average or any such option from the **Summarize value fields** by drop-down box as per our purpose.

FIGURE 2.22

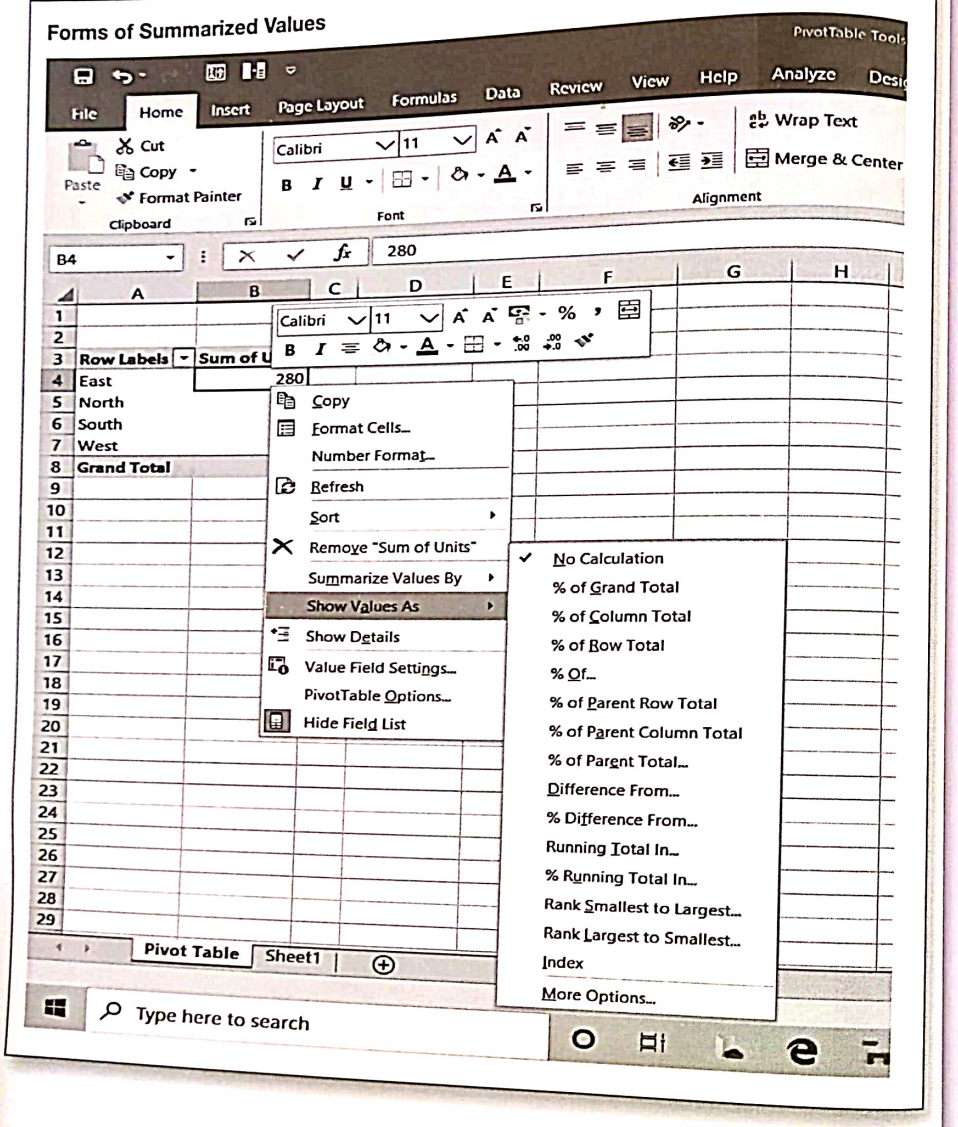
Summarizing of Values



There is a further option of **Show Values As**. Figure 2.23 shows that this option offers several features some of which are the percentage of Grand Total, percentage of Column Total and percentage of Row Total. For better visualization of data, we can **group** and **ungroup** dates in our data in months, quarters, years, as per our requirement.

FIGURE 2.23

Forms of Summarized Values



To recap, in this chapter, we have learned to perform basic arithmetic manipulations such as addition, subtraction, multiplication and division. Besides, we have gone through the following basic functions:

- AutoSum
  (To find aggregates of all numbers in our data set)
- SUMIF
  (To calculate sum of the data with one or more conditions attached)
- MAX and MIN
  (To find the maximum and the minimum value(s) in the data)
- COUNT and COUNTIF
  (To find the count of numerals in our data with or without conditions)
- Power Function
  (To get the result of a number raised to an exponential power)

Finally, we have been trained on how to summarize data using Pivot Tables. We have also explored the operation of arithmetic functions on the summarized data.

PRACTICE QUESTIONS

Following is the table showing the marks secured in Physics of Class Xth students. Read the information given and answer the questions given below.

		Physics		
Roll No.	*Name*	*Theory (out of 70)*	*Practical (out of 30)*	*Total (out of 100)*
1.	Neetu	65	27	
2.	Harshit	53	25	
3.	Kashish	45	29	
4.	Riya	55	26	
5.	Tannu	48	28	
6.	Khushi	55	25	
7.	Vaibhav	60	29	
8.	Archit		27	81
9.	Sonia	43		
10.	Shruti		28	78
11.	Mannu		22	57

(Continued)

(Continued)

		Physics		
Roll No.	*Name*	*Theory (out of 70)*	*Practical (out of 30)*	*Total (out of 100)*
12.	Aastha		27	85
13.	Shubham	57		79
14.	Amit	68	27	
15.	Atul	55	22	

1. Complete the table.
2. If the total marks are increased to 300, compute the new table showing the combined upgraded marks.
3. Using the AutoSum function, calculate the aggregate marks of all students in theory and practical.
4. What are the average marks obtained by the class in theory and practical?
5. Using SUMIF function, calculate the aggregate of marks obtained by students scoring more than 80.
6. How many students scored more than 80 marks (use COUNTIF function)?
7. Name the students who topped the class and who scored the least marks.
8. Summarize the class results by using the Pivot Table.

3

CHAPTER

Data: Input into Excel and Sources

LEARNING OBJECTIVES

After reading this chapter, the readers will be able to understand

- Qualitative and quantitative data
- Meaning of primary and secondary data
- Methods of entering data in Excel
- Importing data from other sources
- Sources of data—data sets available in the public domain such as RBI, NSE, BSE, MOSPI, World Bank and UNdata and how to download them in Excel

Data is one of the most essential and key aspects of any study. Data is a collection of numbers, facts, figures, texts, words or descriptions of things. A collection of data is known as data sets, usually organized in a tabular form. Data sets are then used for analysing and drawing conclusions to gain a better understanding of a phenomenon and/or to make informed decisions. In this chapter, we will learn data entry into an Excel sheet. This chapter also acquaints the readers with the freely available data resources on the various sectors of our economy that can be imported to an Excel spreadsheet.

3.1. UNDERSTANDING DATA

Data can be categorized into qualitative data and quantitative data. *Qualitative data* is non-numerical and based on categorical variables to describe, classify or categorize things based on characteristics, properties, quality or labels. For instance, information related to gender, marital status, suggestions and customer's feedback on services availed. On the other hand, *quantitative data* is expressed in numerical terms and can be counted or measured. Information about the height, weight, age, size, area, scores, income, revenues or prices illustrates quantitative data.

Based on the sources used to collect the data, the data can be classified into two types—*primary* and *secondary*. When the data is collected for the first time for a specific purpose of analysis, it is known as the primary data. Primary data, therefore, is the first-hand information collected. It is collected using surveys, questionnaires or personal interviews. The examples are data collected by an organization from its customers to conduct market research or the census data collected by the government.

Secondary data refers to the information already collected. Examples of sources on secondary data are either reports published by government departments and organizations on their websites. Since collecting primary data is usually time-consuming and expensive, the secondary data is used by many as it is less expensive, quick and easy to collect. The information collected from either of these data sources can be analysed using various techniques.

3.2. DATA ENTRY IN EXCEL

Excel has many features when it comes to data entry. We can enter both numbers and alphabet in Excel. Usually, the following are the methods adopted to do the same.

The *first method* to enter the data in Excel is to type the data directly into the cell by using a keyboard. Alternately, select the cell where the data is required to appear and use **Formula Bar** to enter the data. For example, if we want to enter 350 in cell B1, then select cell B1 and go to the **Formula Bar**, enter 350 and press **Enter** button. The same can be done by entering the data in cell itself. Nevertheless, this method to enter data manually, one by one, is difficult and not practical, if the data size is large. Moreover, manual entry of data is time-consuming and the probability of occurrence of error is also high.

The *second method* is by the use of *copy* and *paste* option. Select the cell that we would like to copy, right click and select **Copy** or press **Ctrl + C**. Then select the cell where we want to **Paste** the data, right click and select **Paste** or press **Ctrl + V**. In case, we want to copy the *same* data available in the cell to the subsequent cells, the method is to use the *Autofill* option. Select the data cell that we want to copy, bring the cursor to the bottom right corner of the cell to the + icon and drag up to the cell where we want the data to be copied. We can fill other types of data series as well in the same manner. Select the first two data cells, bring the cursor to the bottom right corner of the second cell to the same + icon, drag it up to the last cell where we want data to appear. The numbers in selected cells are filled orderly. Like numbers, one can also add months.

In case of any mistake in the data, like a spelling error or the number error, select the cell and double click into the cell and correct it or you can use **Formula Bar** and press **Enter**. Excel also provides *Auto Prompt*. If we type any name and want to repeat it in any other cell, select the cell where data is to be entered and type the first letter of the same name. Excel will auto prompt it

and press **Enter** to confirm. In addition to this, it is important to understand the character of fractions here. To illustrate, if 2 ¼ is typed in a cell and press **Enter**, the **Formula Bar** displays 2.25. But if we type fraction ¼, Excel automatically reads it in a date format. These types of mistakes can be avoided easily by just adding zero in front of the fraction (e.g., 0 ¼).

The **third method** is using the data entry form. In this utility, data can be entered up to 32 fields per record. This is a useful feature of Excel doing the process of data entry faster with fewer errors. Typically, the option of data is not available on **Quick Access Toolbar** but can be added in the following manner:

- Right click on any of the existing icons or click the down arrow in the **Quick Access Toolbar**
- Click on **Customize Quick Access Toolbar**
- From the drop-down list, select **More Commands**
- Select **All Commands** and choose **Forms** in the list shown
- Click **Add** and enter **OK.**
- The **Form** icon will be displayed on the **Quick Access Toolbar** as shown in Figure 3.1

To make a new form, we have to add headings in the column and adjust the width of each column. The form will use this width as the default size for all form fields. Select all the cells containing the headings and click **OK.** For example, if we want to create a form of consumer satisfaction with seven fields, then enter the headings such as name of the customer, quantity of apple purchased, income of the customer, price of apple, price of its substitutes, taste and preference, and number of family members. Verify that all the columns have headings. Highlight your data range including column headings by clicking the cell from where the data starts and press **Ctrl + Shift +** → in case the headings are in columns. Similarly, press **Ctrl + Shift +** ↓ in case the headings are in rows. Then select **Forms** from the **Quick Access Toolbar.** Click **OK.** A new window will appear as shown in Figure 3.2 with the heading of the column as fields.

Enter the data in the **Form** fields. Press **Enter** or **New** button to enter the record in the table and get a blank form for the next record. This data form can be created to enter the primary data.

A data entry form in Excel has many different buttons as shown in Figure 3.2.

- **New:** This will clear the existing data in the form and allows it to create a new record.
- **Delete:** This will allow to delete an existing record.
- **Restore:** If editing an existing entry, one can restore the previous data in the form.
- **Find Prev:** This will find the previous entry.
- **Find Next:** This will find the next entry.

FIGURE 3.1

Adding Form to Quick Access Toolbar

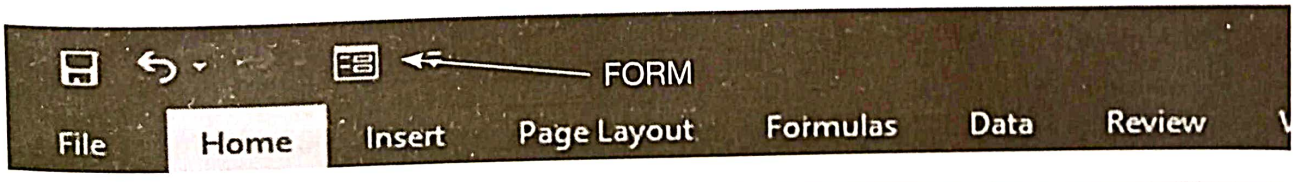
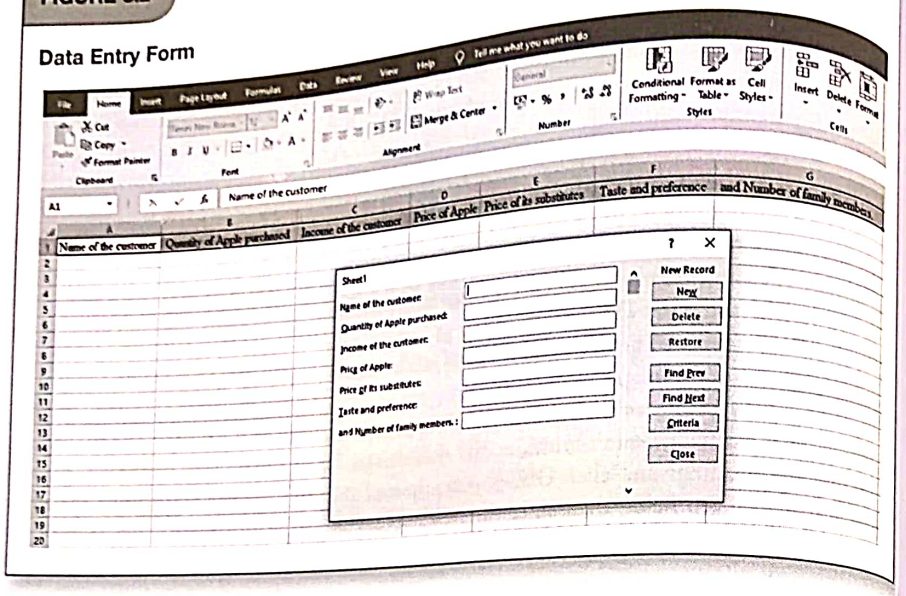


FIGURE 3.2

Data Entry Form



- **Criteria:** This allows us to find specific records
- **Close:** This will close the form.
- **Scroll Bar:** Use the **Scroll Bar** to go through the records

To navigate through existing records, select any cell in the Excel table and click on the **Form** icon in the **Quick Access Toolbar**. A new window will appear, to go to the next entry, click on the **Find Next** button. To go to the previous entry, click the **Find Prev** button. To edit an entry, simply make the change and press **Enter**.

3.3. IMPORTING DATA IN EXCEL

In Excel, data can be imported from text file, web query and comma-separated values (CSV) files. This is very useful and saves time. Through this feature, one can directly import the data from other sources and analyse it with the tools available in Excel. Steps involved in importing data from text are explained below:

- Select **Data** from the tabs.
- Select **From Text** from the Get external data group.
- From the new window that appears, select **Text Files** from the bottom right-hand corner if your file does not have CSV. Otherwise, you can directly choose the CSV file that you want to import to Excel and select **Import**.

- Select **Delimited** from Original Data Type of the newly appeared Text Import Wizard if your text data is in split form. Select **Next**.
- If your text is split with a comma, then select **Comma** command from the Delimiters list. Like this you can also select **Tab**, **Semicolon**, **Space** and others according to your data format and press **Next**.
- From the newly displayed window, select **General**, but there are cases when the data is starting with zero. If **General** is selected, then Excel will not consider the zero. So to avoid this, one has to be more vigilant and select/highlight the column, where zero starts and then select **Text** from the Column Data Format list.
- In the same window, we can also control decimal (.) and thousand separator (,) by selecting **Advanced Key** on the right-hand side. After all these steps, select **Finish**.

A new dialogue box will appear. Data can be imported in the existing sheet or a new sheet. To convert the imported text into a table, press **Ctrl + T**. A table with filters will be displayed.

Excel can also import data directly from the websites. This feature makes Excel user-friendly and can import any data from any website. For this, first copy the URL of the site which contains the data that you want to import, and then follow the steps:

- Select **Data** from tabs
- Select **From Web** from the get external data groups
- A new web query window appears, then paste the URL that is already copied in the Address box given and click **Go**
- From the website opened, you can select the arrow near the table you want to add and select **Import** and **OK**

3.4. DATA SETS AVAILABLE IN THE PUBLIC DOMAIN

Nowadays, the secondary data is easily accessible online or from public and university libraries. It is economical to collect such secondary data as a wide range of data sets can be downloaded either free of cost or by paying nominal charges. Since the data sets are usually available in Excel format, there is no need to prepare an Excel data sheet for given parameters. Here is the list of few important sources of publicly available data that are widely used.

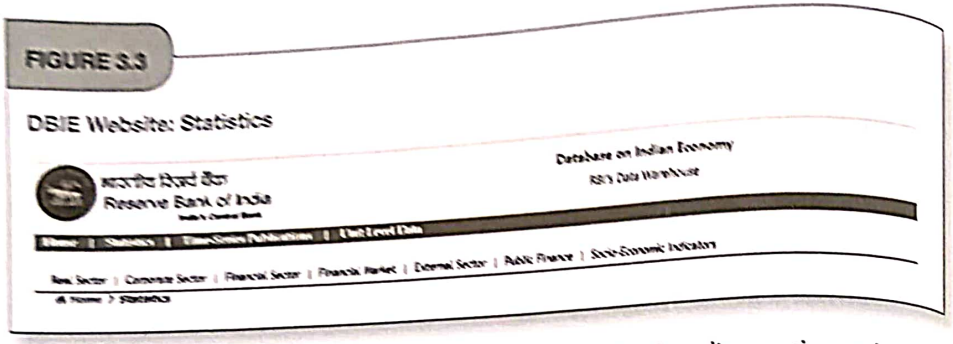
3.4.1. Database for Indian Economy

The Reserve Bank of India (RBI) is the central bank of India which prepares and furnishes a large volume of economic and financial data through the data publication and reports. Department of Statistics and Information Management of RBI has set up a public website (http://dbie.rbi.org.in) for database for the Indian economy referred to as the RBI's data warehouse. This website provides quick access to a wide range of real-time economic and financial data. Data on the Database on Indian Economy (DBIE) website can be downloaded and saved as Excel files for analysis. The website has four parts—*HOME*, *STATISTICS*, *TIME SERIES PUBLICATIONS* and *UNIT LEVEL DATA* (Figure 3.3).

The **Home page** has some important Indicators such as *Reserve Money*, *GDP*, *Money Market*, *G-Sec Turnover*, *Business of Scheduled Banks*, *Monthly WPI and IIP*, *RBI Balance Sheet* and *Foreign*

FIGURE 3.3

DBIE Website: Statistics



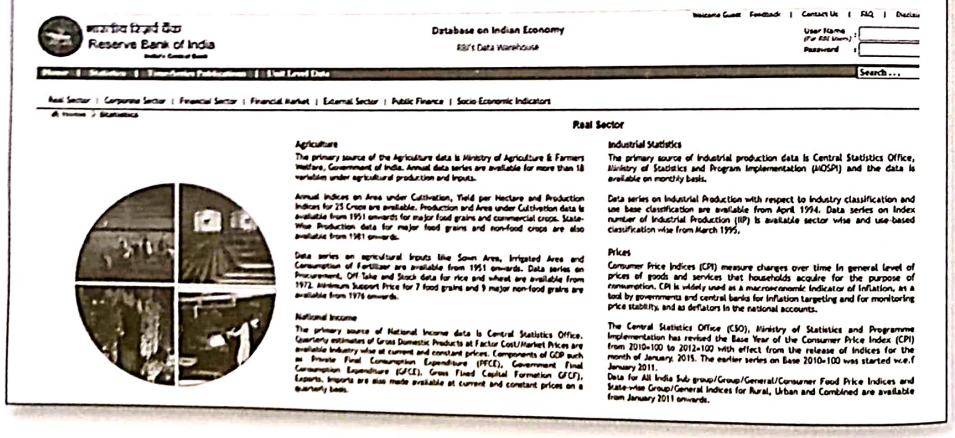
Exchange Reserves on the left-hand side. The SAARCFINANCE Database dispense time series data on macroeconomic and socio-economic variables of SAARC countries that can be downloaded and saved in different formats, including Excel. The Home page contains a dashboard that shows the economy at a glance. The new data or publication released on the website is notified below the dashboard. On the right-hand side, there are certain useful links such as *Brochure on DBIE (English)*, *Downloading Data into Excel*, *DBIE Data Release Calendar* and *RBI Main Site*.

The second part **Statistics** has been divided into *seven subject areas* (Figure 3.3). The data series related to the *subject areas* and their various *sub-sectors* are compiled periodically. The subject areas are as follows.

1. **Real sector** comprises agriculture, industry, national income, and price and wages consisting of information mentioned in Figure 3.4.
   - Agriculture such as production and area under cultivation data for major food grains and commercial crops, indices on area under cultivation, yield per hectare and state-wise production of food grains.

FIGURE 3.4

Data on Real Sector on DBIE Website

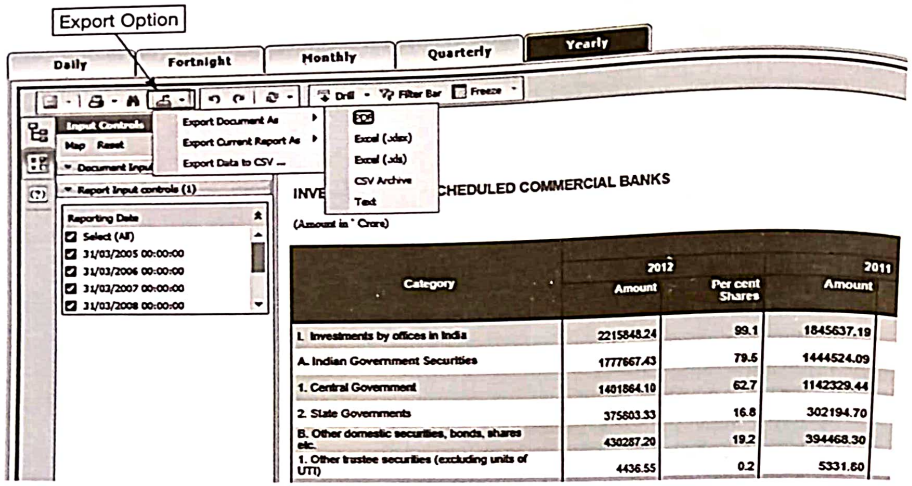


   - Industrial statistics such as production of select industries, Index of Industrial Production (IIP), and production and imports of crude oil and petroleum products.
   - National income such as quarterly estimates of gross domestic products, private final consumption expenditure, government final consumption expenditure, exports and imports.
   - Prices and wages like consumer price indices.
2. **Corporate sector** includes data on finances of FDI companies, listed non-government non-financial companies, non-government non-banking financial and investment companies, non-government non-financial private limited companies, non-government non-financial public limited companies.
3. **Financial sector** provides information on the following:
   - Banking—assets and liabilities such as capital, liabilities and assets of scheduled commercial banks
   - Banking—branch statistics such as number of banks, branches, newly opened branches and closed branches
   - Banking—performance indicators such as profitability, gross and net non-performing assets of scheduled commercial banks and so on
   - Banking—sectoral statistics, financial institutions such as assets and liabilities of Deposit Insurance and Credit Guarantee Corporation, Industrial Finance Corporation of India, EXIM Bank, National Bank for Agriculture and Rural Development, National Housing Bank, Small Industrial Development Bank of India and other financial institutions
   - Key rates such as bank rate, cash reserve ratio and statutory liquidity ratio
   - Monetary statistics, non-banking financial companies, payment systems
4. **Financial market** furnishes data on:
   - Equity and corporate debt market such as data series on Bombay Stock Exchange (BSE) and National Stock Exchange (NSE) indices
   - Forex market such as data series on exchange rates for four major currencies, that is, USD ($), Euro, GBP and Japanese Yen against the domestic currency (INR)
   - Government securities market and money market
5. **External sector** disseminates data on
   - External debt such as government borrowing
   - External sector indices, forex reserve
   - International finance such as foreign investment inflows and international trade like commodity-wise export and import
   - Components of balance of payments such as current and capital account balance, merchandise trade and service income
6. **Public finance** segment supplies data on central and state government finances (combined) such as direct and indirect tax revenues of central and state governments and market borrowings of central and state governments
7. **Socio-economic indicators** provide information on the employment situation in India and the population below the poverty line up to 2012.

To extract the data from the website, one can simply access the reports available in the STATISTICS tab. To do this, first, choose the sector and the sub-sector for which the data is required (Figure 3.4). The reports (daily, fortnightly, monthly, quarterly, yearly) will be displayed

FIGURE 3.5

Exporting Data (Sector-wise)



and then click on the required report. The period for which the data is available is mentioned next to each report. The data will appear as shown in Figure 3.5. Click on the **Export** option, then select **Export Document As** to export the report to Excel, PDF, Text and CSV.

The third part is **Timeseries Publications** that include numerous tables incorporated in various bank publications. They are organized under the respective publication tabs such as *Bank Branch Statistics, Data on International Banking Statistics of India, Handbook of Statistics on the Indian Economy, Monthly RBI Bulletin, Primary (Urban) Co-operative Banks' Outlook, Public Debt Statistics, Quarterly Statistics on Deposits and Credit of Scheduled Commercial Banks, Statistical Tables Relating to Banks in India* and *Weekly Statistical Supplement*. Reports under **Timeseries Publications** are real time as they are continuously updated without waiting for the annual printed versions of the publication giving access to the latest data to the users. It is important to note that the time series data available in these publications can be downloaded in Excel/Pdf/text format. The steps to export data remain the same as mentioned above.

The fourth and the last part **Unit Level Data** contains the reports related to the household surveys conducted by RBI like *Inflation Expectations Survey of Households* to record households' expectations of change in price level and inflation rate and *Consumer Confidence Survey* to collect data on the consumer current perception and future expectations on five parameters—economic condition, employment scenario, general price levels, income and spending of the households. The data can be downloaded in the same manner as the aforementioned.

3.4.2. Ministry of Statistics and Programme Implementation

Ministry of Statistics and Programme Implementation (MoSPI) was established as an independent ministry on 15 October 1999 after the merger of the Department of Statistics and the Department of Programme Implementation. The Ministry has two wings, one relating to statistics called the National Statistical Office and the other is Programme Implementation wing. The Ministry focuses and places a substantial amount of importance on the coverage and quality of the statistics released in the country. MoSPI plays a vital role in the preparation and maintenance of a comprehensive database of economics and social statistics. The statistics released are usually based on the administrative sources and the surveys conducted by the central and state governments. The Ministry releases the data in both printed form and on its website http://www.mospi.gov.in/. Figure 3.6 provides an overview of its website.

The Home page of the website contains all the latest Press Releases, Recent Reports and Announcements. A key performance indicator dashboard is also displayed. To download the reports issued by the ministry click on the Download Reports option on the Home page. The list of all the reports will be displayed. We can also search for the required reports in the Search Reports drop-down menu (Figure 3.7).

FIGURE 3.6

The MoSPI Website

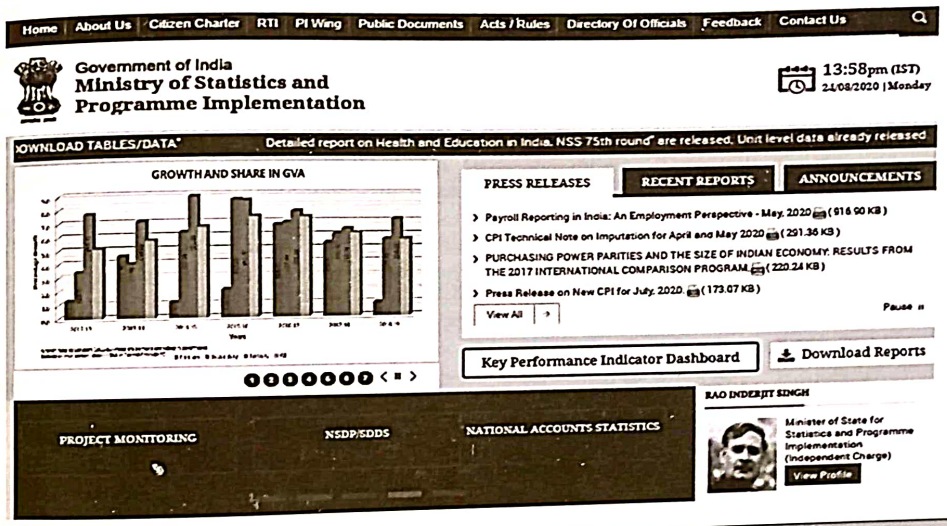
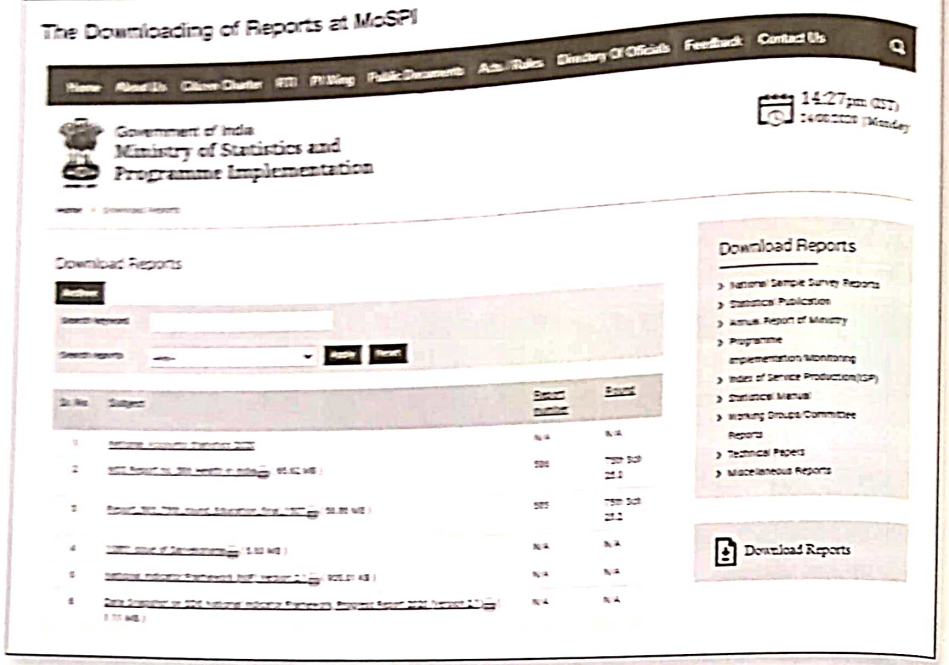


FIGURE 3.7

The Downloading of Reports at MoSPI



The statistics wing of ministry prepares *National Accounts Statistics* and publishes consolidated sets of accounts for various macroeconomic aggregates, non-financial corporations, government, households, rest of the world and industry-wise aggregates. These statements can be directly downloaded as Excel files for analysis. The latest national accounts statistics are available in the **Recent Reports** on the Home page.

To download the previous national accounts statistics and other statistical publications, click on the **Download Reports** on the Home page and choose Statistical Publication from the drop-down menu in **Search Reports** and then select the desired publication from the given options, that is, Social Statistics, Economic Statistical, National Accounts and Research and Publication. An example is shown in Figure 3.8.

The Ministry also furnishes *Quarterly GDP Data* and *Annual Estimates* of national products, government and private final consumption expenditure, savings and capital formation. The Ministry gathers and reports the *Monthly Estimates of Index of Industrial Production*, releases *Consumer Price Index* and annual inflations rates. The latest, as well as the previous estimates, can be downloaded from the list of the 'Press Release' by the Ministry (Figure 3.9). We can choose the desired estimate from the drop-down menu in the search button and can fill up the period (**Start date** and **End date**) for which we need to download the data.

FIGURE 3.8

Statistical Publications Provided by MoSPI

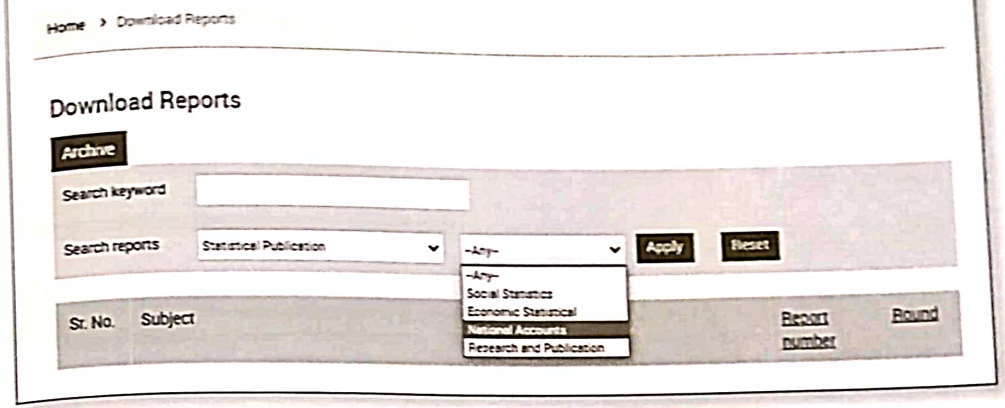


FIGURE 3.9

Press Release

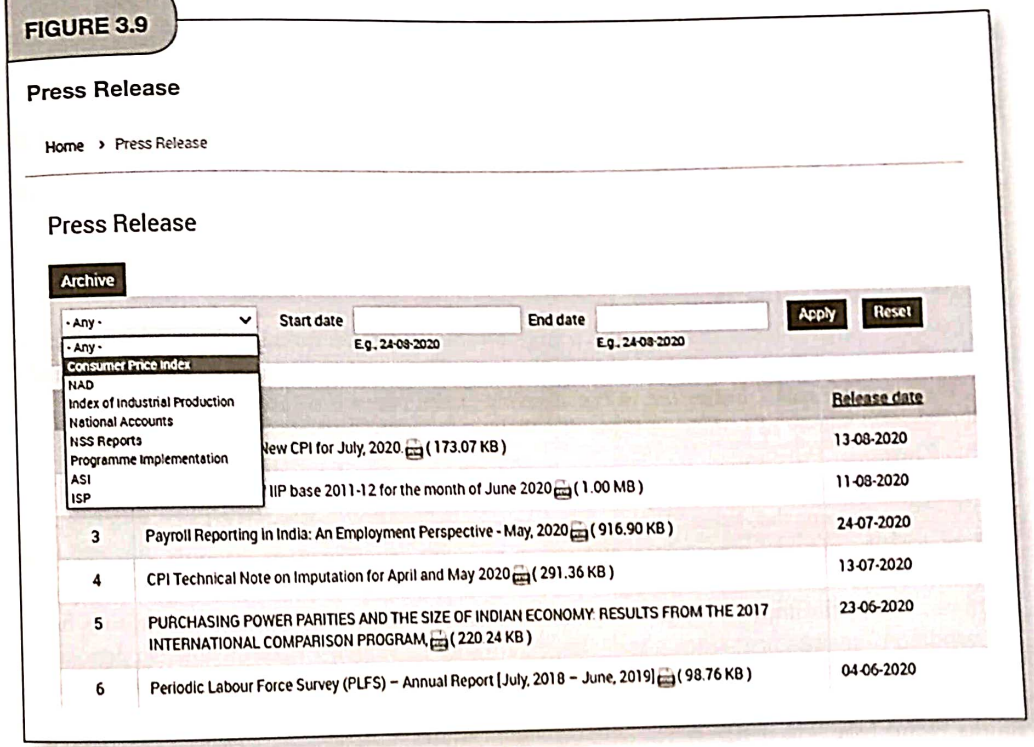
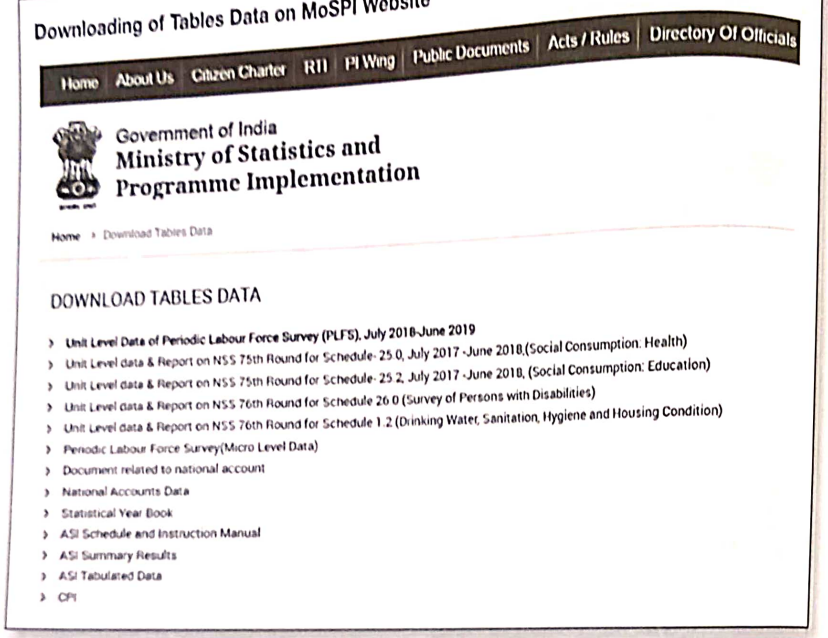


FIGURE 3.10

Downloading of Tables Data on MoSPI Website



To download a wide range of the tables data, select the options of **Download Tables Data** on the Home page. The list of tables data will be displayed as shown in Figure 3.10. It consists of

1. Unit Level Data for varied parameters
2. National Accounts Data
3. Statistical Year Book that contains a large volume of data on various subject matters such as area, population, agriculture and other significant primary sectors, industries, trade, education, banks, insurance, prices, exchange rates, railways and rainfall
4. Annual Survey of Industries Data
5. Consumer Price Index

MoSPI carries out the *Annual Survey of Industries* and publishes statistical information to estimate and analyse the growth, composition and structure of the manufacturing sector. It also conducts *Economic Censuses* periodically and *Large-scale All-India Sample Surveys*. National Sample Survey programme to study the impact of specific parameters such as employment, consumer expenditure, housing conditions, environment, literacy levels, health, nutrition and family welfare on different population groups. It also disseminates the microdata from these surveys and census. To browse, search and download relevant census or survey documents and metadata, click on **Micro Data Catalogue** (Figure 3.11) on the Home page. The person will be redirected to an external website powered by the National Data Archive that provides web-based access to the entire metadata (Figure 3.12).

FIGURE 3.11

Step 1 in Microdata Catalogue

FIGURE 3.12

Step 2 in Microdata Catalogue (National Data Archive)

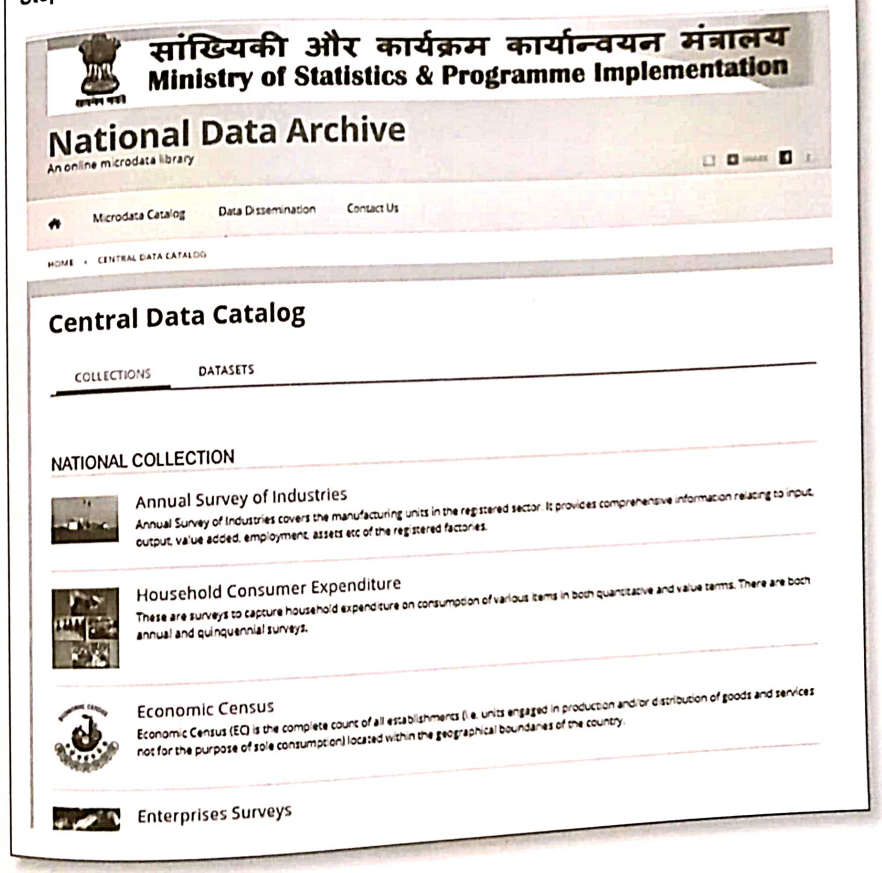
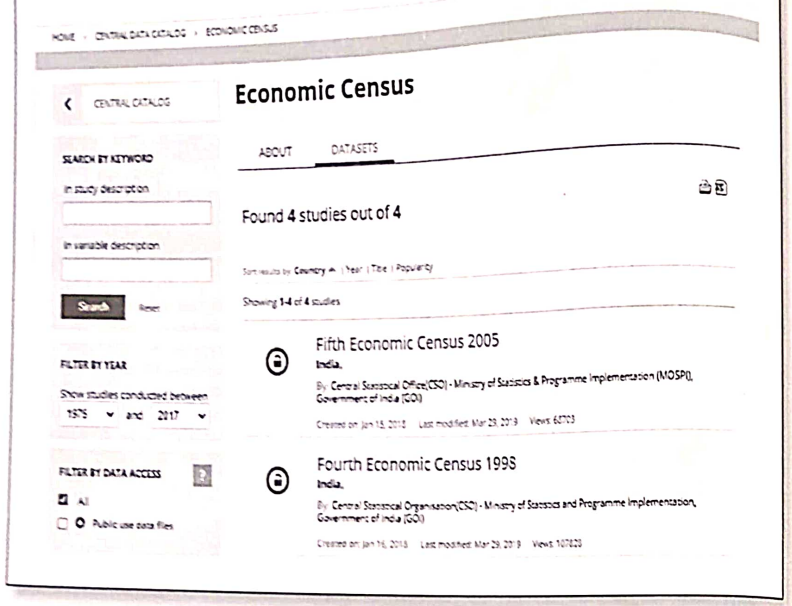


FIGURE 3.13

Step 3 in Microdata Catalogue (National Data Archive)



To view and download the data, click on the **Visit Catalogue** tab. The list of catalogues will be displayed (*free data available for public use as well as data not freely available to the users*), then choose the desired study (Figure 3.13). Once we choose the desired study, select the option Get Microdata. If the user is already registered, enter the username and password, otherwise, click on Register for the registration. After login, fill the application and click on Submit. Then download from the given data sets.

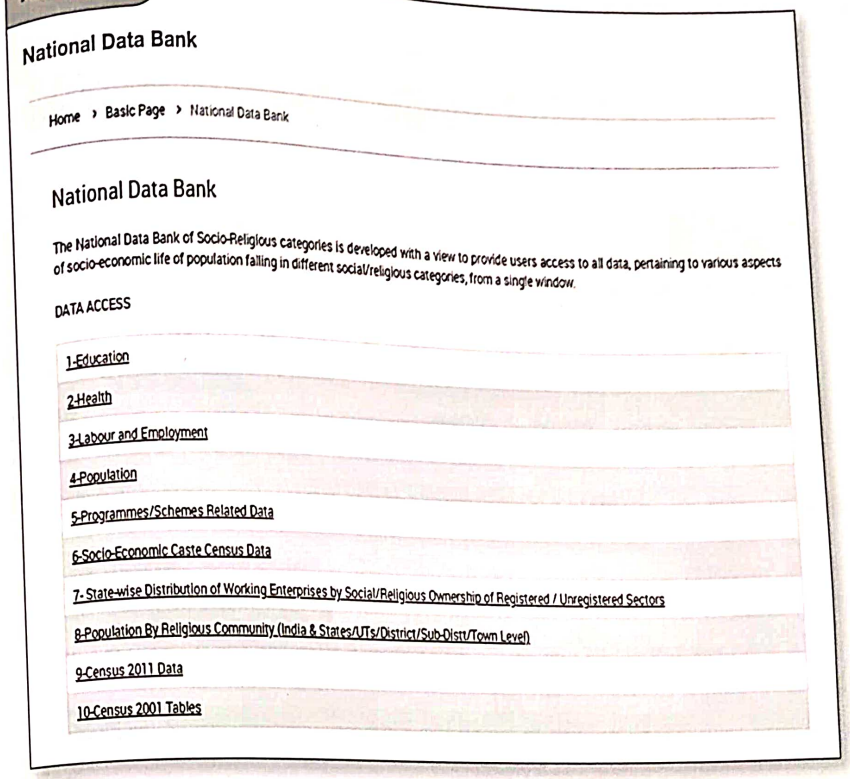
Data on various aspects of socio-economic life of population falling in different social and religious groups can also be downloaded from the National Data Bank. It is available as a link on the Home page of the website under Quick Links (Figure 3.14). The information sought can be found as per requirement.

3.4.3. Directorate of Economics and Statistics

The Directorate of Economics and Statistics, Ministry of Agriculture and Farmers Welfare, Government of India (https://eands.dacnet.nic.in/Default.htm) is an important data source of information pertaining to the agricultural sector. It provides links to various reports related to the agricultural sector such as *All India Crop Situation*, *Agricultural Statistics at a Glance* and *Pocket Book of Agricultural Statistics*. These reports provide information in a pdf format such as

FIGURE 3.14

National Data Bank



socio-economic indicators, *labour force*, *poverty and unemployment*, *land-use statistics* and *livestock statistics*, among many, mostly at the all-India level. It is important to note that every state has its own Directorate of Economics and Statistics which provides most of the economic data pertaining to the state and its districts.

3.4.4. National Stock Exchange

NSE is a leading stock exchange in India. It was incorporated in 1992 and commenced operations with the advanced and automated electronic-based trading in 1994. It provides a platform for the companies to list themselves and raise capital and also allows the investors to trade in various investment instruments and securities. A large volume of data and statistics related to equity, market, indices, mutual funds, equity derivatives, currency derivatives, commodity derivatives, interest rate derivatives, fixed income and debt and public issues is provided on the website of NSE. The Home page of the website includes all the important statistics, information, announcements and a dashboard that shows a chart demonstrating the

real-time value of Nifty 50. The new website can be accessed through URL https://www.nseindia.com/. The link to the old website is www1.nseindia.com. A glimpse of the new website is shown in Figure 3.15.

The process to download the historical data involves the following steps. First, select the **Resources** tab and choose the option **Historical Reports** from the Downloads menu (Figure 3.16). The historical reports of Capital, Derivative and Debt Market will appear (Figure 3.17). We can then

FIGURE 3.15

The New NSE Website

FIGURE 3.16

Step 1 in Downloading Historical Reports—NSE

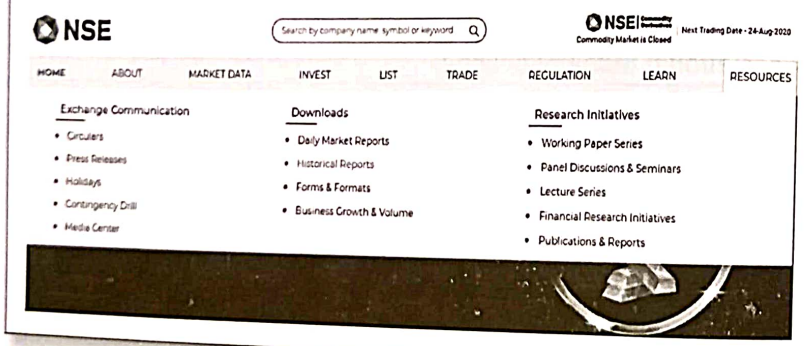


FIGURE 3.17

Step 2 in Downloading Historical Reports—NSE

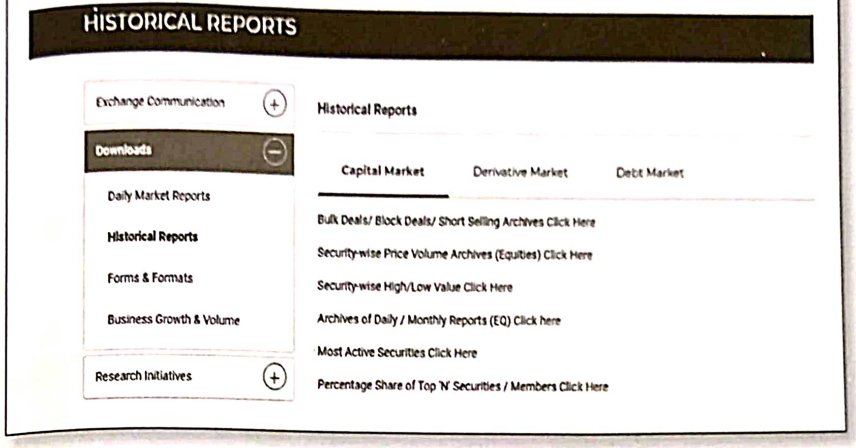
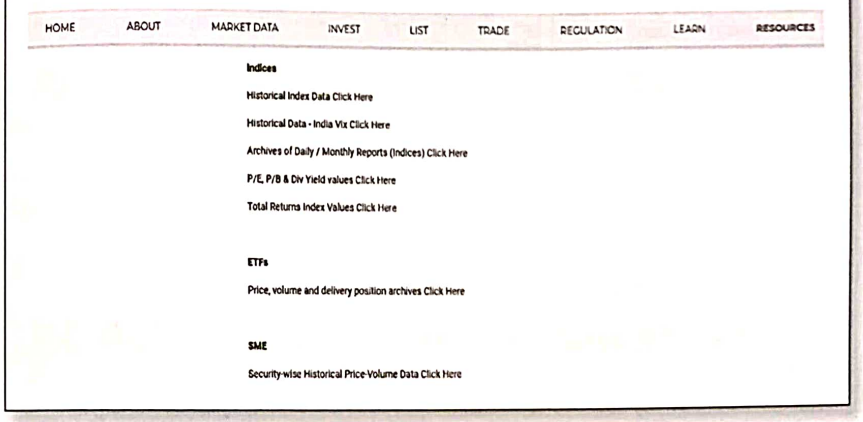


FIGURE 3.18

Step 3 in Downloading Historical Reports—NSE



select the required report from the numerous reports listed under different heads, for instance, Historical Index Data under the head Index in Capital Market reports (Figure 3.18). After filling up the relevant details like the Index Type and the period for which the data is to be extracted, that is, **start date** and **end date**, the data will be displayed. Finally, click on the **download option** on the right-hand side to export the data in Excel as a CSV file.

Almost all the notable companies are listed on the stock exchange. To download data for a specific company listed on NSE, we can directly search the name of the company in the Search option given on the top of the website (Figure 3.19). The information about the company will be displayed (Figure 3.20). Scroll down and select the option of **Historical Data** and the desired data will appear (Figure 3.21). We can fill up the dates for which the data is to be extracted, that is, **start date** and **end date** as per our requirements. Finally, click on the **download option** on the right-hand side to export the data in Excel as a CSV file.

FIGURE 3.19

Step 1 in Historical Data of a Company—NSE

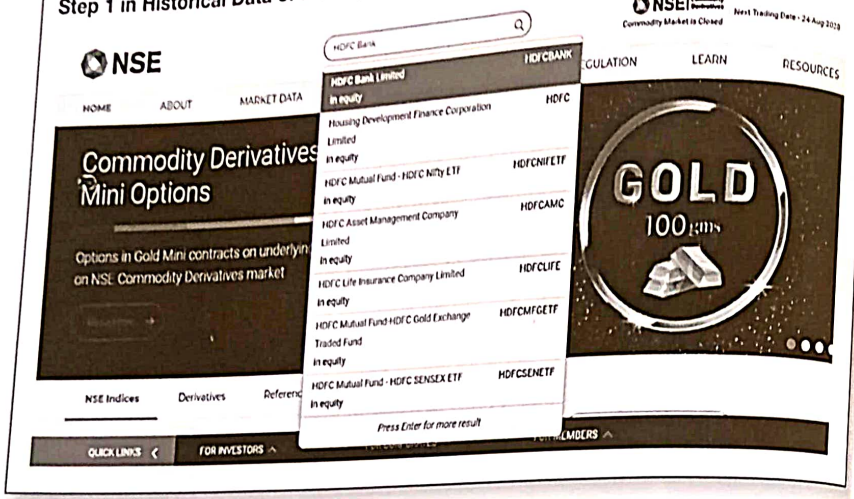


FIGURE 3.20

Step 2 in Historical Data of a Company—NSE

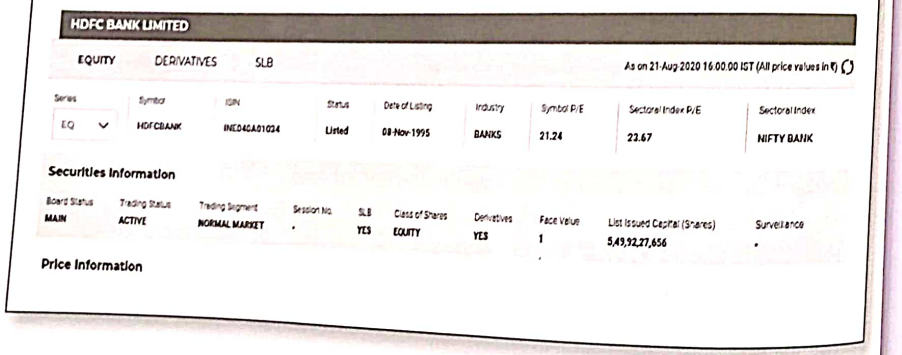
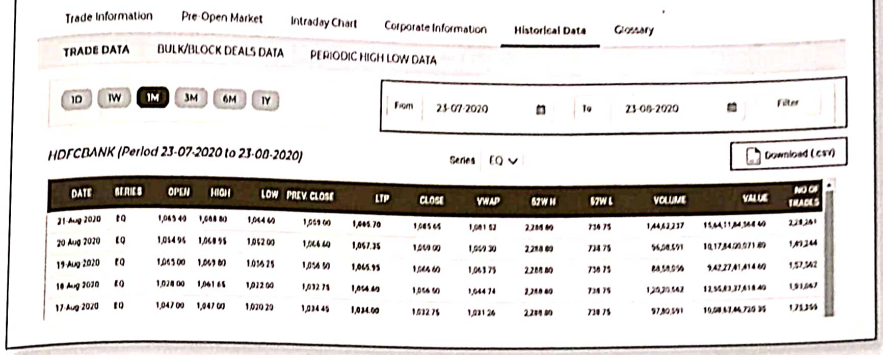


FIGURE 3.21

Step 3 in Historical Data of a Company—NSE



3.4.5. Bombay Stock Exchange

BSE is the first Indian stock exchange which was established in the year 1875. It is Asia's oldest stock exchange. Similar to NSE, the website of BSE equally provides access to quick and accurate market data (real-time and historical) about the equity, equity derivatives, currency derivatives (including interest rate derivatives), commodity derivatives and debt as well as corporate data to analyse and evaluate the performance of a company or the market as a whole and make informed decisions. The Home page of the website includes all the important statistics, information, announcements and a dashboard that shows a chart demonstrating the real-time value of BSE SENSEX. Its website URL is https: // www.bseindia.com/.

To export the market data available from the website in Excel, first select the Markets option from the menu as shown in Figure 3.22. Then select the Market Data option in the Markets menu (Figure 3.23).

The historical information concerning Equity, Indices, Public Issues/Offer for Sale (OFS)/ Buybacks, Listing, Derivatives, Currency Derivatives, Indian Depository Receipt (IRD), Debt,

FIGURE 3.22

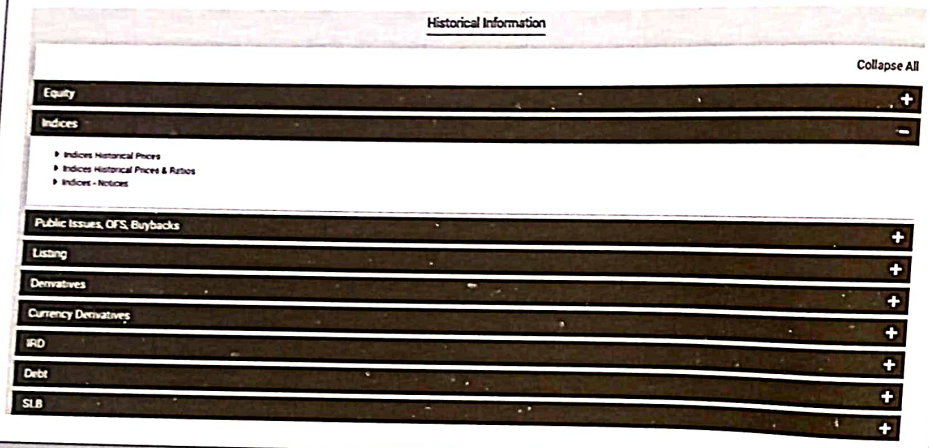
Step 1 in Downloading Historical Data—BSE

FIGURE 3.23

Step 2 in Downloading Historical Data—BSE

FIGURE 3.24

Step 3 in Downloading Historical Data—BSE



Securities Lending and Borrowing (SLB), Exchange Traded Funds (ETF), Mutual Funds, Commodity Derivatives, Foreign Institutional Investors (FII) and so on will be displayed. Choose the required data from the list. For instance, the Historical Prices of Indices under the head Indices (Figure 3.24).

Next, fill up the details for the required historical data such as the **Index Type** from the drop-down menu and the **period** for which the data is to be extracted, that is, *Daily, Monthly or Yearly* and the **start date**. Click on the **Submit button** (Figure 3.25). The desired data will be displayed. Finally, click on the **download option** below the data on the left-hand side to export the data in Excel as a CSV file (Figure 3.26).

FIGURE 3.25

Step 4 in Downloading Historical Data—BSE

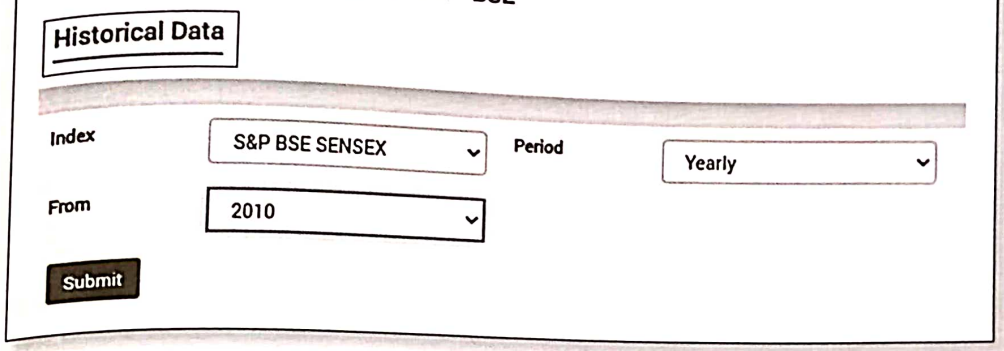
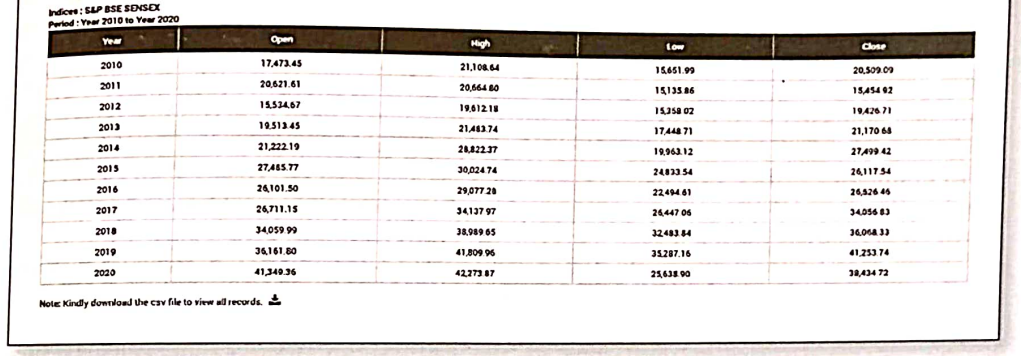


FIGURE 3.26

Step 5 in Downloading Historical Data—BSE



Indices : S&P BSE SENSEX
Period : Year 2010 to Year 2020

Year	Open	High	Low	Close
2010	17,473.45	21,108.64	15,651.99	20,509.09
2011	20,621.61	20,664.80	15,135.86	15,454.92
2012	15,534.67	19,612.18	15,358.02	19,426.71
2013	19,513.45	21,483.74	17,448.71	21,170.68
2014	21,222.19	28,822.37	19,963.12	27,499.42
2015	27,485.77	30,024.74	24,833.54	26,117.54
2016	26,101.50	29,077.28	22,494.61	26,826.46
2017	26,711.15	34,137.97	26,447.06	34,056.83
2018	34,059.99	38,989.65	32,483.84	36,068.33
2019	36,161.80	41,809.96	35,287.16	41,253.74
2020	41,349.36	42,273.87	25,638.90	38,434.72

Note: Kindly download the csv file to view all records.

3.4.6. World Bank

The World Bank is an international organization, formerly known as the International Bank for Reconstruction and Development, founded in 1944 out of the Bretton Woods Agreement. It offers financial and technical assistance such as loans, grants, policy advice, training, research and analysis to the developing countries with a motive to eradicate poverty, support economic advancement and ensure sustainable increase in the quality of lives of the people in these developing countries.

The Development Data Group at the World Bank coordinates statistical and data work. It collects and maintains several macroeconomic, financial and sectoral databases. Statistics and data can be easily accessed on the World Bank Open Data portal (https://data.worldbank.org) by all

FIGURE 3.27

World Bank Open Data

users. The World Bank provides free and open access to a comprehensive set of high-quality data about development in countries around the globe. Its site facilitates easy access, download and usage of the data. The data can be used free of cost with minimum restrictions. To obtain the data from the site, simply use the search box on the top of the page as shown in Figure 3.27. We can search for data for an Indicator or Country or any other topic (Figure 3.27).

As we enter the text, the search box shows a list of matching results. We can select one option and the results will be displayed using graphs and a data table. Click on the **Download button** on the right-hand side to download the data in Excel, CSV or XML format. If one selects a specific indicator, then the data downloaded will contain the data for *all countries* for *all years* with respect to the indicator selected (Figure 3.28). Similarly, if data is downloaded for a country, then it will provide *all data* for *all years* for that particular country (Figure 3.29).

To download the data for *combinations of indicators and countries* (Figure 3.30) or *multiple countries* (Figure 3.31), we can insert additional items in the search box. However, we can only search for one indicator at a time.

The site provides open access to the data on the World Development Indicators which dispense a variety of relevant, high quality and internationally comparable indicators such as population, Gross Domestic Product (GDP), CO_2 emissions, literacy rate, consumption, income, stock markets and tourism. The database contains 1,600 time series indicators for 217 economies and more than 40 country groups, with data for most of the indicators going back more than 50 years. These data sets can be found in several main themes: poverty and inequality, people, environment, economy, states and markets, and global links.

The site also covers diverse collections and databases including **data bank**—the data available on the World Bank Open Data site are a subset of those available in the World Bank's data bank, which contains substantial collections of time series data. One can access data bank as either a registered or an unregistered user and download the data to your system in either Excel, CSV or text format. **Data catalogue**—a complete list of data sets published by the World Bank is available in the data catalogue. It comprises all the data sets in data bank, along with many other useful data sets that cover data from surveys. **Microdata**—the World Bank's Microdata Library provides access to microdata collected through sample surveys of households, business enterprises and so on.

FIGURE 3.28

Downloading Data: Indicator-wise

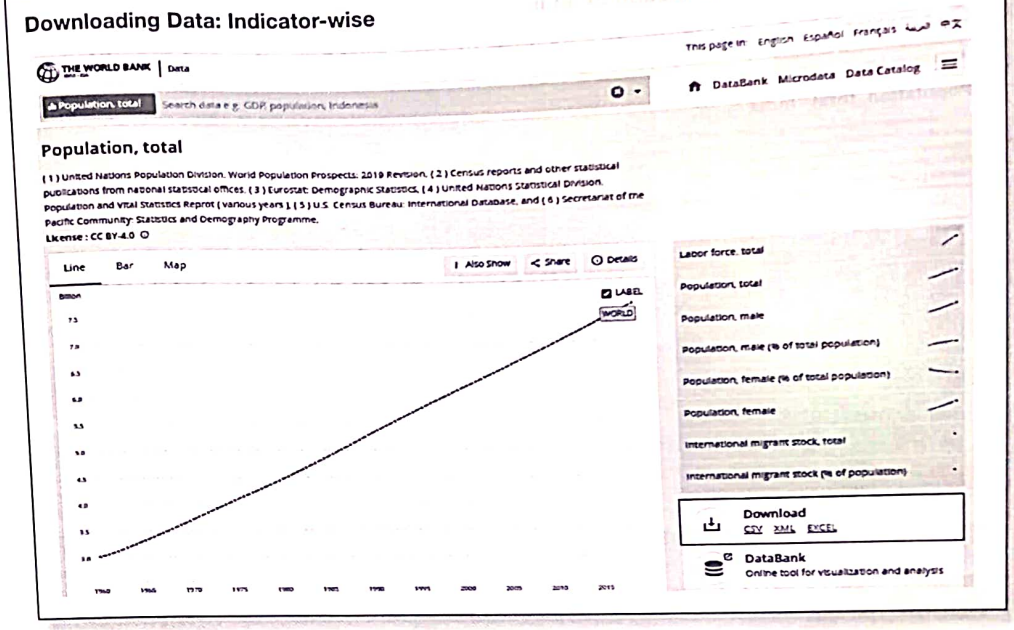


FIGURE 3.29

Downloading Data: Country-wise

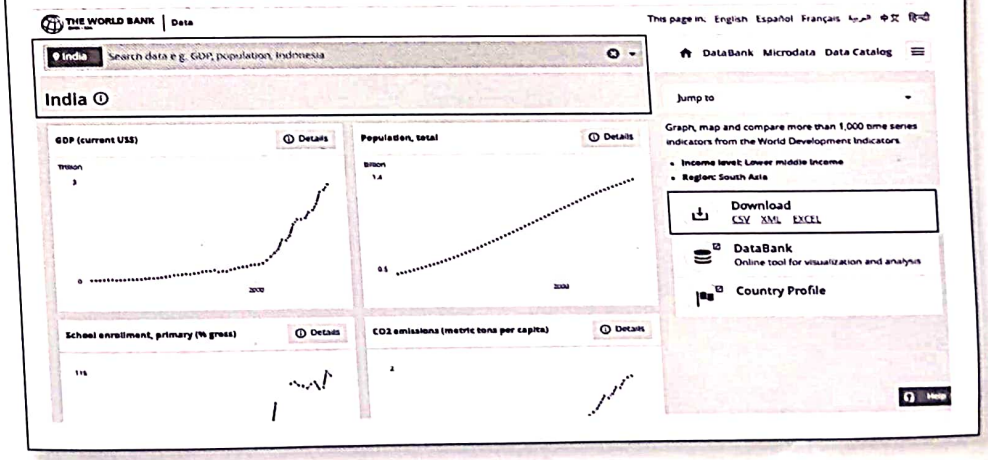


FIGURE 3.30

Downloading Data: Combination of an Indicator and Country

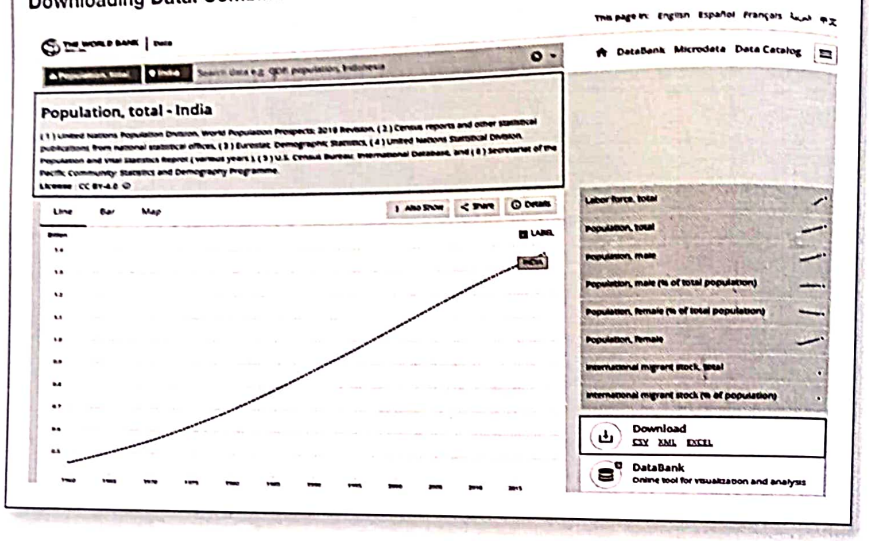
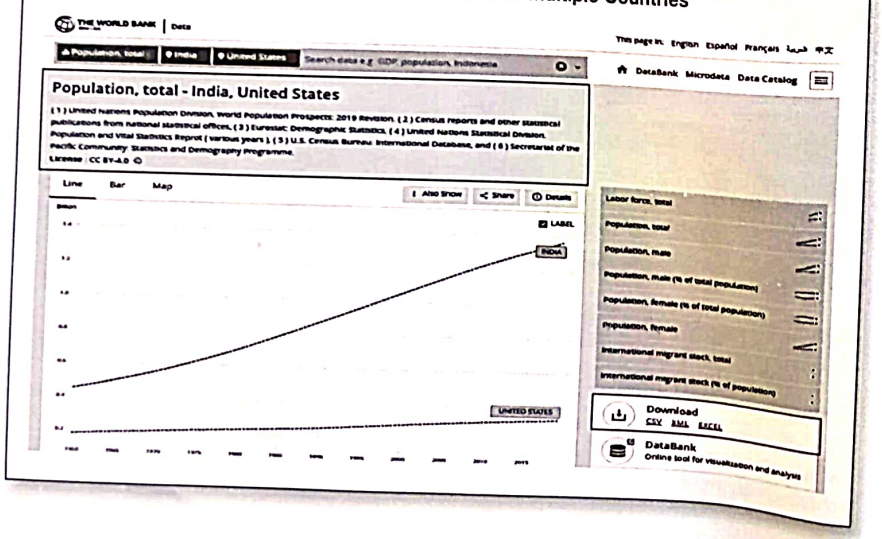


FIGURE 3.31

Downloading Data: Combination of an Indicator and Multiple Countries



3.4.7. UNdata

UNdata is a web-based data service for the worldwide users maintained by the Statistics Division of the Department of Economic and Social Affairs of the United Nations Secretariat. It was launched in 2005 as part of a project called 'Statistics as a Public Good'. The main objective of UNdata is to provide free access to global data and to educate users about the significance of statistics for the purpose of policy and decision-making. The users around the globe can search as well as download a large volume of statistical resources and data compiled by the United Nations and other international agencies. The website to download data is https://data.un.org/.

Plenty of databases and tables collectively known as *Datamarts* contain *millions of data points* and cover a broad range of subjects including population and migration, national accounts, education, labour, price and production indices, international trade, energy, crime, gender, nutrition and health, agriculture, science and technology, finance, environment, communication, transport and tourism, and development assistance (Figure 3.32).

To download the data, one needs to simply type the keywords in the search box on the top of the page as shown in Figure 3.33, and click on the **Search button**.

The search box recognizes countries, regions, subjects/indicators and years. For instance, data series of 'GDP per capita' matches the keywords shown in Figure 3.34. One can click on the desired series to view the entire record. As shown in Figure 3.35, the GDP per capita for all the countries and all years will be displayed. Click on the **Download button** above the data to

FIGURE 3.32

UNdata Website

FIGURE 3.33

Step 1 in Downloading of Data Series—UNdata

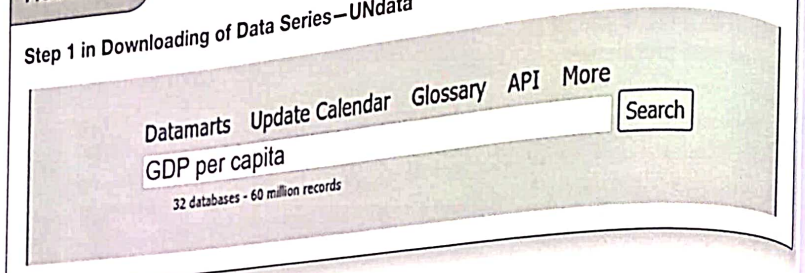
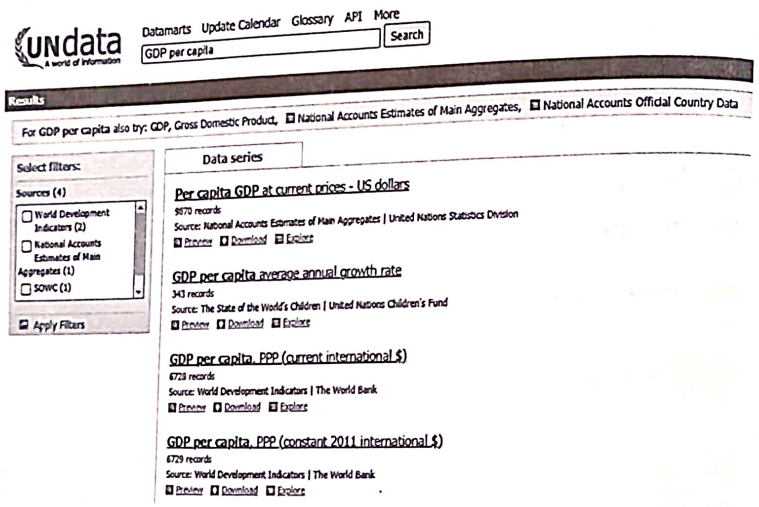


FIGURE 3.34

Step 2 in Downloading of Data Series—UNdata



download the data series in XML, Excel or text format. We can also apply filters to download the data series for one or more desired countries and years. However, we can only search for one indicator at a time (Figure 3.36).

The data from the sources discussed above are made available freely for its users with minimal restrictions on specific databases. However, some websites provide access to databases after paying a subscription charge. One such example is the EPWRF India Times Series. It provides access to reliable, up-to-date and high-quality time series data to assist in research across various sectors of the Indian economy. This online database comprises over 50,000 variables compressed in

FIGURE 3.35

Step 3 in Downloading of Data Series—UNdata

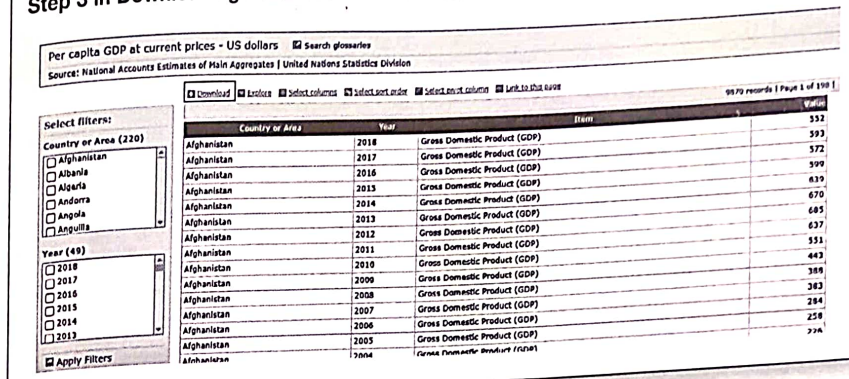
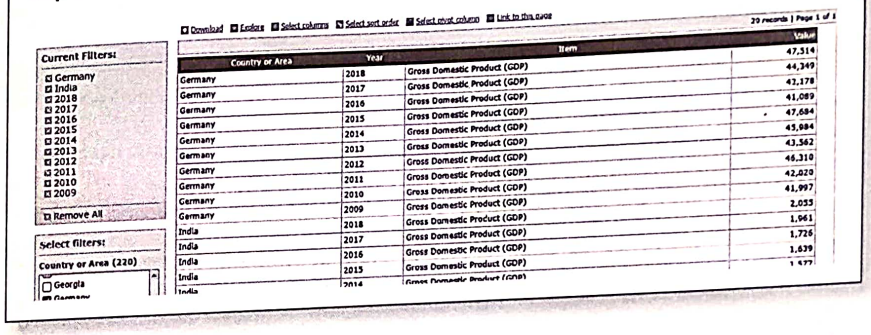


FIGURE 3.36

Step 4 in Downloading of Data Series—UNdata



20 modules and continuous time series from 1950 based on the availability. To access the data from a module, the user (individual, universities, institutes) can either pay annual subscription rates which enables unlimited download from the subscribed module or pay per use service (pay uniform rates each time they use the service) based on the number of rows and columns to be accessed. Likewise, Indiastat.com is a subscription-based website that provides a comprehensive compilation of socio-economic data at the national and state levels.

To sum up, this chapter provides a broad understanding of data, various ways to input data into an Excel spreadsheet and various data sources from where the data can be downloaded particularly in an Excel format. The downloaded data can be analysed and interpreted on the basis of different statistical techniques used in the book.

4

CHAPTER

Visualization of Data Using Charts in Excel

LEARNING OBJECTIVES

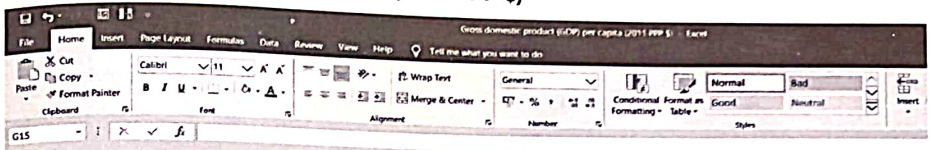
After reading this chapter, the readers will be able to understand

- Popular charts provided by Excel
- Plotting and interpreting the data using a line graph
- Plotting and interpreting the data using a column chart
- Plotting and interpreting the data using histogram
- Representing the data using a pie chart
- Plotting and interpreting the scatter plot

Charts (also known as graphs) provide us with the easiest method for visualization of data. Not only they display data at a glance, but they also facilitate comparison and unveil trends and relationships. Raw data, if numerous, does not convey any information per se. The data needs to be summarized, processed and presented for obtaining information. Communicating information quickly is, therefore, an essential purpose of drawing graphs. In this background, the present chapter explores the charting features of Excel. We will introduce here the popular charts provided by this software and learn how to plot them. Some of the popular charts we will study are line chart, column chart, pie chart and scatter plot. The data we have exercised for this task is per capita gross domestic product for SAARC countries for the period 2010–2018, as shown in Figure 4.1. The data is expressed in dollars using purchasing power parity rates in 2011.

FIGURE 4.1

Per Capita Gross Domestic Product (2011 PPP $)



	A	B	C	D	E	F	G	H	I
1		Country							
2	Year	Afghanistan	Bangladesh	Bhutan	India	Maldives	Nepal	Pakistan	Sri Lanka
3	2010	1672	2518	6814	4451	11965	1987	4072	8503
4	2011	1627	2651	7269	4625	12486	2053	4095	9156
5	2012	1773	2791	7548	4817	12261	2155	4149	9980
6	2013	1808	2925	7616	5065	12573	2250	4241	10239
7	2014	1796	3067	7954	5378	12892	2385	4348	10647
8	2015	1767	3232	8380	5743	12684	2455	4459	11078
9	2016	1757	3424	8944	6145	13019	2447	4609	11447
10	2017	1758	3634	9247	6516	13333	2606	4771	11706
11	2018	1735	3879	9348	6899	13611	2724	4928	11956

Source: UNDP. 2018. 2018 Statistical Update: Human Development Indices and Indicators. New York. Available at http://hdr.undp.org/en/content/human-development-indices-indicators-2018 (accessed 29 April 2020).

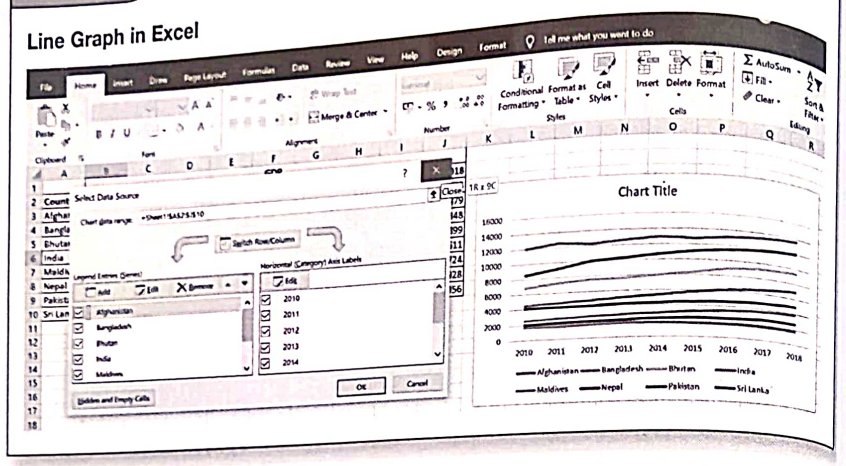
4.1. LINE GRAPH

The line graph consists of a horizontal X-axis (independent variable) and a vertical Y-axis (dependant variable). The independent axis is called so because the values of the variable on this axis do not depend on the other variable taken. On the other hand, the values of the variable on the Y-axis depend on the independent variable taken on X-axis. The line graphs are appropriate when the values on both the axes are *numerally ordered* (instead of qualitative variables). We generally use this graph when we want to measure the values of a variable over a time period. Conventionally, time (expressed in days, weeks, quarters, years and so on) is usually placed on the X-axis. We may now illustrate the preparation of the line graph using the data given in Figure 4.1 in Excel. To start with, note that the cursor is placed on an empty cell in Excel and follow the below-mentioned steps:

- In the **Insert** tab, select the **Insert Line or Area Chart** in the **Charts** option. Usually, the first option of **Line** is chosen, which is a more commonly used option among the options granted. Another similar way can be to select data first and then click on the **Insert** tab to choose the type of graph.
- The blank slate that appears is the Format Chart Area. Right click and choose **Select Data** option. In the pop-up window, named *Select Data Source*, choose the **Chart data range**, which in our case is **A2–I11** (note that the data range also includes an Excel name sheet along with it).

FIGURE 4.2

Line Graph in Excel



- The Legend Entries column (in the pop-up window itself) shows different data series for which we want the data to be presented as shown in Figure 4.2. In our example, different data series include data from different countries. Note that the Legend Entries column also shows 'Year' which is a *variable* and not a separate data series. Hence, the Year box may be unchecked.
- The Horizontal Axis Labels need to be edited because as mentioned before, the horizontal axis represents the time variable (which in our case are years). After clicking on **Edit**, select the data range, which is the year column (**A3–A11**). Click **OK** and another **OK**.
- The chart is ready, as shown in Figure 4.2.

4.1.1. How to Put Labels?

This chart is not self-explanatory. We, therefore, need to label this chart further. This can be done in the following two ways:

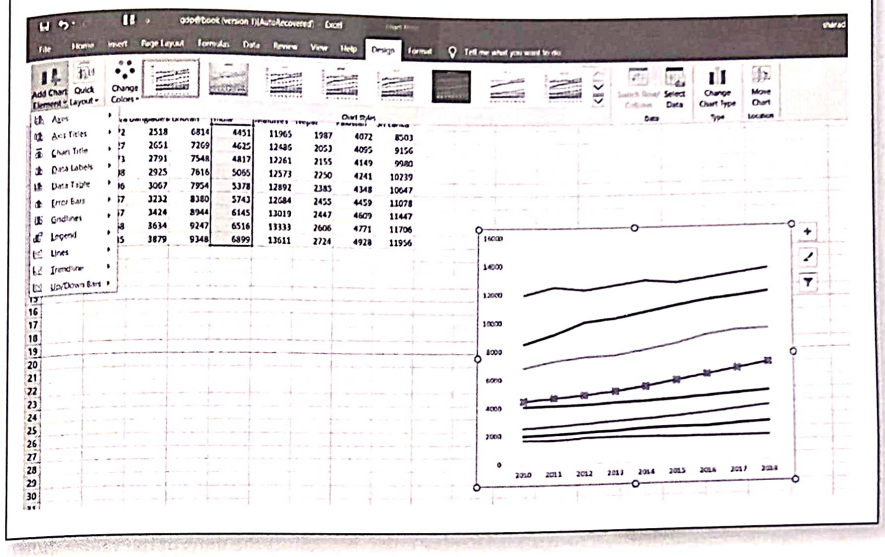
1. We may select the **Design** tab and then choose the **Add Chart Element** as shown in Figure 4.3.
2. We may click on the + sign given on the top right corner of the chart.

Different options displayed in the methods mentioned above have been elaborated here.

- **Axis:** It helps us to show the two primary axes (*horizontal* and *vertical*) on the graph. Generally, we show these two axes while preparing our chart.
- **Axis Title:** It enables us to name the axis according to the variables we choose or select. We may click on the **Axis Title** which appears next to both the axes and then rename it.
- **Chart Title:** We can name or give the title to our chart by using it.
- **Data Labels:** It displays the values of different data points in our graph. The values can be placed around the data points in different ways such as centre, left, right, above and below.

FIGURE 4.3

Labelling of Chart



- **Data Table:** In case we wish to show our entire data set right below the chart, we may use this option.
- **Error Bars:** It represents the variation graphically in data using standard error, percentage and standard deviation. This option is more often used by statisticians.
- **Gridlines:** These are the horizontal and vertical lines that get displayed in the graph for better readability of data.
- **Legend:** It labels the different series of data presented in our graph. These labels can be placed on either right, top left or bottom of the chart. It is an essential tool for differentiating between the different groups of data in the graph.
- **Trendline:** It shows the general pattern or direction of the data in a linear, exponential or linear forecast manner. At our level, it is not of much use.
- **Up/Down Bars:** It is used to connect the first and the last series of the data. Again, it is not of much importance at this stage.

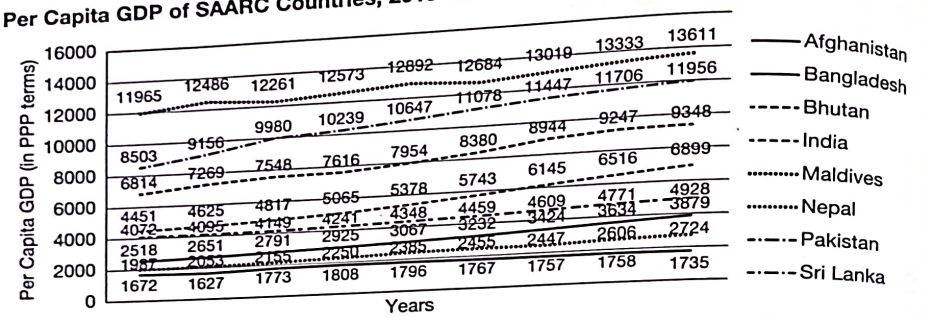
4.1.2. Formatting the Axes

We can format the axes as well. Let us format here the vertical value axis by right clicking on the values of the Y-axis.

- Select **Format Axis** and a format axis box will appear which offers various options.
- The first two options that are minimum and maximum indicate the starting and ending values on Y-axis, which can be adjusted according to one's requirement.

FIGURE 4.4

Per Capita GDP of SAARC Countries, 2010–2018



Source: UNDP. 2018. 2018 Statistical Update: Human Development Indices and Indicators. New York. Available at http://hdr.undp.org/en/content/human-development-indices-indicators-2018 (accessed 29 April 2020).

- Next, under the **Units**, the spacing between the values on the Y-axis can also be changed in the **Major** box.
- Another feature that can be used is **Display units** that adjusts the units of the values shown.

Using the features mentioned above, the chart shown in Figure 4.3 is labelled, which looks much more informative in Figure 4.4 now. We can say just by looking at the graph that per capita GDP is consistently the highest in Maldives and lowest in Afghanistan. It can also be observed that, for example, the per capita GDP for India increased from $4,451 in 2010 to $6,899 in 2018.

4.2. COLUMN CHART

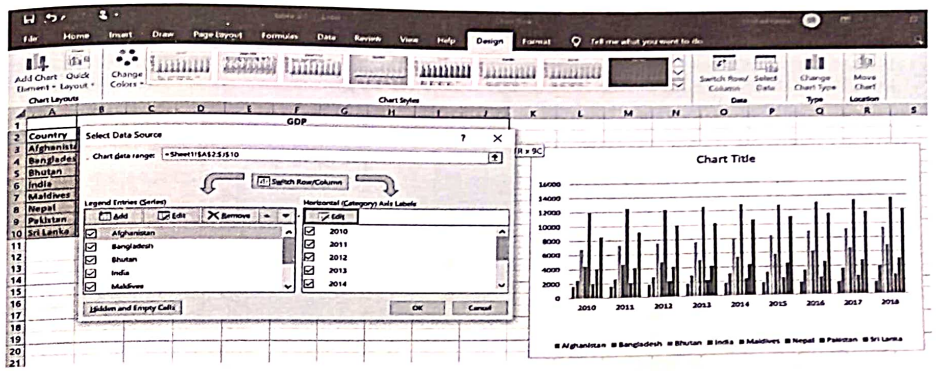
Unlike the line graph, which displays continuous data over a time period, column chart (graph) is used to compare values *across categories* such as years, countries and movie genres, where the data is discrete. The data is displayed in rectangular bars. The height of rectangular bars commensurate to data values. It provides a visual representation for the comparison of quantities leading to an immediate clear vision of the data. One may sometimes get confused between bar graph and column graph; hence, it may be important to create a line of difference. The major difference lies in their orientation. The bar graph is oriented *horizontally*, while the column graph is *vertically* oriented.

Let us now look at the preparation and interpretation of the column graph in Excel. We continue with the previous example of GDP of SAARC countries. Most of the steps for drawing column charts are similar to that of the line graph. In the Excel file, containing the data

- Click on **Insert** tab.
- Move the cursor on **Insert Column** or **Bar Chart**. Choose the first Graph option, which is usually used, among the various options available. We again get to see a blank slate.
- Right click in the slate area and choose **Select Data**.

FIGURE 4.5

Steps of Depicting Per Capita GDP of SAARC Countries, 2010–2018 in Column Chart



- In the dialogue box, click on **Add** under **Legend Entries**. Suppose we want to represent Afghanistan first. We select the cell named Afghanistan in the Series name. In Series values, we select the data values of Afghanistan, that is, **B3:B11**.
- Click **OK**.
- In **Horizontal (Category) Axis Labels**, edit the category and select the data values of years as it is the *category* we wish to present.
- Click **OK**, and we get our column graph ready for Afghanistan (Figure 4.5).

We can further add more series in our existing column chart. For example, if we click **Add** under Legend Entries and choose Bhutan as our second series, we will have to add the data values of Bhutan in the same way as we did for Afghanistan. Click **OK**, and we get the same column graph which now is a combined representation of both the countries. A similar practice can be followed for the remaining countries. If we want all the countries to be presented in our column graph, then following the aforesaid procedure of adding individual countries may be cumbersome. In that case, an alternative approach can be followed. In the dialogue box, we may select the entire data of the countries in the **Chart data range**, that is, **B2:I11**.

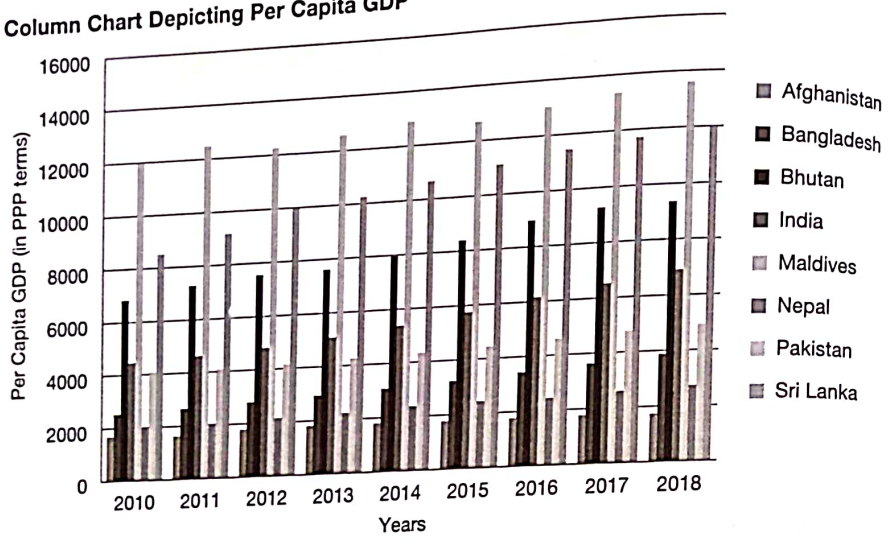
To make this chart more informative, we can make use of different chart element options, discussed in the Line graph. We get the following graph representing the same information as derived in the Line graph as shown in Figure 4.6.

4.3. HISTOGRAM

A histogram is essential as it helps in studying many statistical distributions which we will be discussing in the subsequent chapter. A histogram is a bar graph that represents a frequency distribution. The nature of data demarcates the main difference between a bar graph and a

FIGURE 4.6

Column Chart Depicting Per Capita GDP



Source: Human Development Report, 2018.

histogram. Whereas the bar chart plots *categorical data*, the histogram represents *quantitative data*, divided into equal parts, known as *bins* or *class intervals*. In other words, a histogram shows frequencies of data points that fall into different bins. Also, since there are no gaps between the data intervals, the bars of a histogram do not have spaces between them. The width of the bar represents the interval and the height corresponds to the frequency.

Let us say, the intervals are 0–20, 20–30, 30–40 and so on. The limits 0, 20, 30 become the lower limits and 20, 30, 40 become the upper limits of the aforementioned class intervals. The data is continuous in the sense that the lower limit is included in a given interval but the upper limit is included in the next range. For example, in the intervals mentioned above, 20 will be included in 20–30 class interval, but 30 will be taken in the next class interval, that is, 30–40.

We now take a look at the steps for constructing a histogram. Excel 2016 and higher versions provide a direct function for preparing a histogram of a given data. We take the hypothetical data shown in Table 4.1 as an example to learn how to construct it. This data pertains to the educational background of 35 workers.

We, here, take up a few ways of preparing it. The first method is the one, which is the most direct and easiest one. Select the data which we want to represent in the histogram, that is, column entitled 'Studied up to Which Class' in our example.

- Go to **Insert** tab
- Click the first graph option in the **Insert Chart Statistic**, which is next to **Insert Line or Area Chart**. We have the histogram ready, as shown in Figure 4.7.

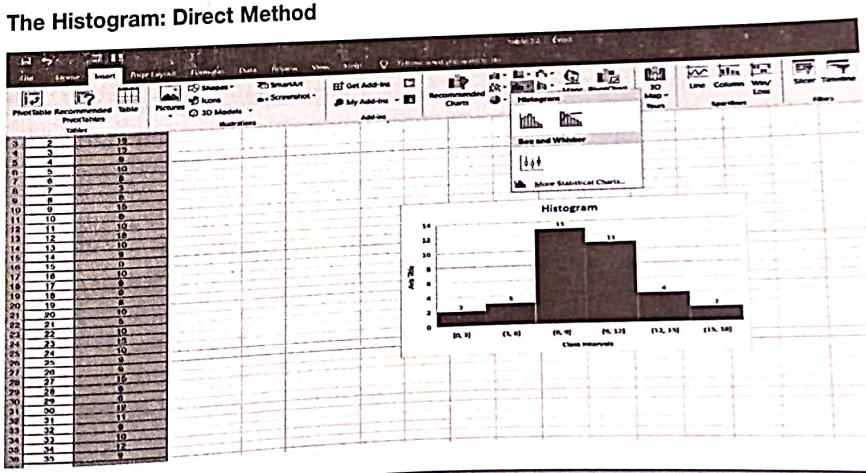
TABLE 4.1

Educational Qualification of Workers

Number	*Studied up to Which Class*	*Number*	*Studied up to Which Class*
1	9	18	9
2	18	19	8
3	15	20	10
4	9	21	5
5	10	22	10
6	8	23	15
7	3	24	10
8	8	25	9
9	15	26	9
10	8	27	15
11	10	28	6
12	18	29	6
13	10	30	12
14	9	31	11
15	0	32	8
16	10	33	10
17	9	34	12
		35	9

FIGURE 4.7

The Histogram: Direct Method



It may be noticed that Excel automatically selects the bin size but can be altered according to the requirement. To do this, right click on the horizontal axis, select the *format axis*. Various bin alterations can be made from the following options given in the format axis box.

- **Bin Width:** It can be used for changing the class interval size.
- **Number of Bins:** This option allows us to change the number of rectangular bars which may better represent the data.
- **Overflow Bin:** If there are many scattered observations above a particular bin, then this option comprises all those observations in a single bin.
- **Underflow Bin:** If there are many scattered observations below a particular bin, then this option comprises all those observations in a single bin.

It may be remarked that Excel is programmed in such a way that the class intervals are exclusive, which means that the upper limit (and not the lower limit) is included in the class intervals. For example, class interval 3–6 contains numbers between 4 and 6. On the preparation of histogram, like other charts, this chart can also be labelled by clicking the + sign on the top right corner of the graph, such as **Axis, Axis Titles, Chart Title, Data Labels** and **Legend** which serve the same functions as discussed earlier.

In the second method, we follow similar steps developed for the graphs prepared earlier.

- Click on the **Insert** tab, choose the first graph option from **Insert Statistic Chart.**
- Right click in the blank slate and choose **Select Data.**
- In the **Chart data range**, we select the data range and click **OK.**

The graph we get has gaps between the bars. To make it look like a histogram, right click on any bar and choose **Format Data Series.** A gap width option can be seen in **Series Options** with a three rectangular bar icon from where we can reduce the width to zero. All options discussed in the previous method can now be applied in the same way.

The method which we will discuss now is being used in the earlier versions of Excel as well. This method can be followed by clicking on the **Data** tab and choosing **Data Analysis** which can be found in the right top corner. It might be possible that this option may not be available in the **Data** tab. So first, we need to add this by clicking on the **File** option and choosing *Options*. Go to **Add-ins**, click **Go** next to **Manage** option and select **Analysis ToolPak** and **Analysis ToolPak-VBA.** Click **OK** and reopen the Excel file. The option is now added.

Before starting this method, we need to explicitly make an additional column for class intervals. In our ongoing example, we have created class intervals of 3.

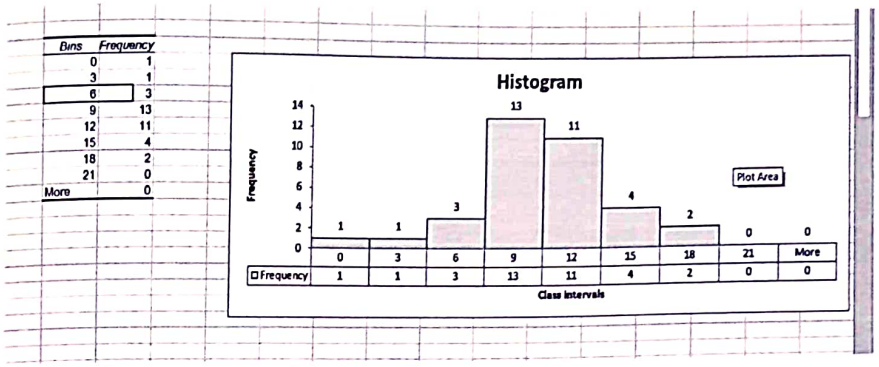
- In the *Data Analysis* dialogue box, select **Histogram** and click **OK.**
- In the next dialogue box created, select **Input Range** which is the data range of the variable that needs to be shown. The class interval column is then selected as the **Bin-Range.**

It is a good practice to click on the checkbox named **Labels** if the given data series have them. To get the desired graph in the existing or new sheet, we also need to check **Chart Output.**

- Click **OK** and Excel will create a chart along with frequency table.

FIGURE 4.8

The Histogram: Another Method



To convert it into a histogram, we need to reduce the space between the bars. To do this, right click in the chart window but not in the area where the chart is made. Select **Format Chart Area** and click on the arrow next to **Chart Options.** Select **Series Frequency,** then click on the icon with three bars and reduce the gap width to zero.

We have now got the desired histogram as in Figure 4.8. The bins here are not in the traditional form so they need to be interpreted according to the frequency table created. The bin column shows the upper limit of the group. For example, in the frequency table, the cell with an upper limit of 6 is representative of the frequency of numbers lying between 4 and 6. This means it excludes the lower limit and includes the upper limit.

However, as mentioned earlier, the *first method* is a favoured one as it is an improvement over the aforesaid methods as later methods go through a longer route.

4.4. PIE CHART

A pie chart is particularly used to show the data of a single series *only* in a pictorial form. It is a circular chart divided into different parts. Each part represents a proportion of the variable taken into account. The quantity represented in pie chart can either be expressed in absolute numbers or percentages. The total value of the pie sums up to 100 per cent when expressed in percentages. We may also note that these charts are useful only when there are few categories to display as a large number of categories may not be very comprehensive and easy to interpret. For example, in our previous example, we can only represent, say, data on Afghanistan using pie chart with years as our category. We cannot combine the data from other countries in the same pie chart. To illustrate it, let us move to a different example shown in Table 4.2. The data considered represents the category-wise production of motor vehicles in India in the years 2008–2009 and 2009–2010.

TABLE 4.2

Motor Vehicle Production during 2008–2009 and 2009–2010

Category	*2008–2009*	*2009–2010*
Total commercial vehicles	416,870	567,556
Cars	1,516,967	1,932,620
Multi-utility Vehicles	321,626	424,791
Total two-wheelers	8,419,792	10,512,903
Three-wheelers	497,020	619,194
Grand Total	**11,172,275**	**14,057,064**

Source: data.gov.in

First, we represent 2008–2009 data using a pie chart. Once the procedure of constructing the line and column chart explained earlier is clearly understood, then the formation of other charts can be similarly understood to a large extent. To begin with,

- Click on the **Insert** tab and select the pie chart icon right next to the **Recommended Charts**. We normally select the very first chart.
- We again see a blank slate in which we right click and choose the option **Select Data.** As usual, we **Add** the Legend Entries by selecting series name as **Category** (of motor vehicles) and series values as the data of 2008–2009, that is, **B2:B6**.
- Now, **Edit** the Horizontal Axis Labels by selecting the different category names, that is, **A2:A6** in our example.
- Click **OK** and the basic pie chart is ready to interpret (Figure 4.9). A similar pie chart can be drawn for the next data series of the year 2009–2010 as shown in Figure 4.10.

To explain the chart further, we can use some features by clicking on the + sign on the top right corner of the chart or click on **Add Chart Element** as done previously. The different options are elaborated as follows.

- **Chart Title:** We can give a title to our chart using a proper name (e.g., Production of Motor Vehicles 2008–2009).
- **Data Labels:** It enables us to label different slices of the chart and contains some handy features. When we click on the arrow in Data Labels, the first three options help in depicting the values, the different parts of the pie chart represent. These values can be placed in the centre, inside the circle or outside the circle depending on the structure of the chart. The last option, that is, **More Options**, contains the following Label Options.
  - **Series Name:** It displays the series which we are representing.
  - **Category Name:** It helps us to identify the different categories shown by the different parts of the chart.

FIGURE 4.9

Steps in Labelling Pie Chart, Motor Vehicle Production 2008–2009

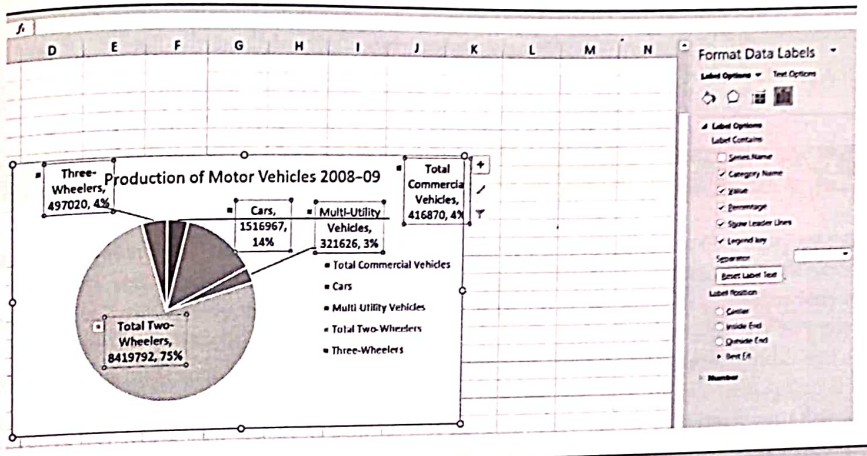
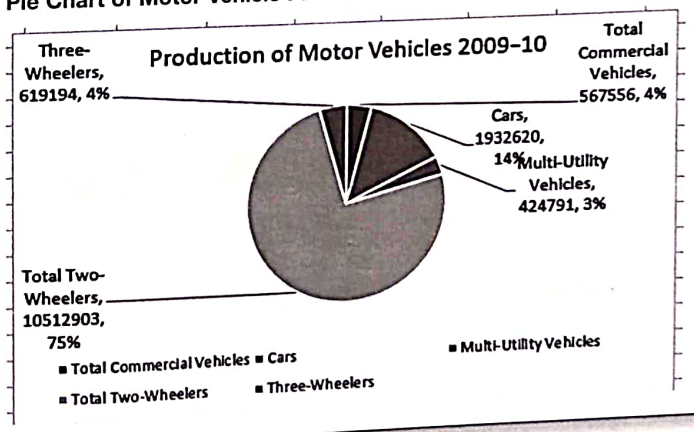


FIGURE 4.10

Pie Chart of Motor Vehicle Production 2009–2010



- **Value:** It depicts the absolute values of different categories of the data.
- **Percentage:** It shows how much of the proportion a particular slice constitutes of the total proportion, expressed in percentage.
- **Legend Key:** Along with the values, the represented key of different categories is shown.

Based on the chart drawn, we can make inferences by looking at the graph, which is one of the primary purposes to draw a pie chart. We can say that 75 per cent of the total production of motor vehicles is of two-wheeler categories. Multi-utility vehicles contributed the least in the production in 2008–2009. We can also draw a comparison between the two series, that is, 2008–2009 and 2009–2010 using separate pie charts. By looking at both the graphs, we observe that the trend has not changed much as the two-wheeler category continues to contribute the largest proportion of the total production in 2009–2010.

4.5. SCATTER PLOT

The scatter plot, also known as the XY plot, helps us in identifying the relationship between two variables. This relationship may either be positive, negative or null. Both X-axis and Y-axis in a scatter plot represent the two variables of interest. The coordinates of every point on the chart are the values of the two variables. For example, a coordinate (3,2) implies that the variable on the X-axis has value 3, while 2 represents the value of the variable on the Y-axis. The x coordinate of a point is called *abscissa* and y coordinate of a point is called *ordinate*.

We will now learn how to plot and interpret scatter plots in Excel. We start with a hypothetical example of a firm detailing its items sold in different months and the advertising cost incurred. The data is presented in Table 4.3. We are interested in finding out how the advertising cost and sales of the firm are related. We may keep in mind that if there is a priori relationship between the variables, or if one variable is dependent on the other variable, then the data may be organized by keeping the data of the independent variable first followed by the dependent variable data.

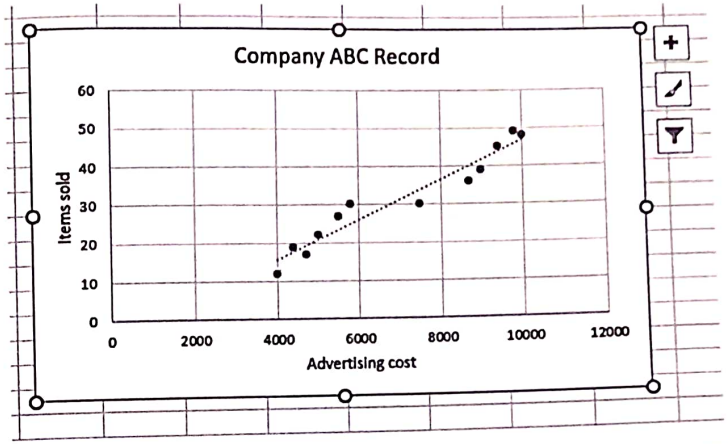
TABLE 4.3

Sales and Advertising Cost of Firm ABC

Month	*Advertising Cost (in ₹)*	*Items Sold*
January	4,000	12
February	5,500	27
March	4,700	17
April	7,500	30
May	9,000	39
June	10,000	48
July	9,800	49
August	9,400	45
September	8,700	36
October	5,000	22
November	4,400	19
December	5,800	30

FIGURE 4.11

The Scatter Plot of Company ABC



Otherwise, we will have to correct the graph by clicking **Switch Row/Column** from the **Select Data** option and then removing all the series except the series of interest. However, if the two variables taken are not related, then the series can be arranged in either way.

In Excel 2019, the process of creating a scatter graph is quite simple. To begin with, we select the data of the two variables between which we want to explore the relationship, that is, advertising cost and items sold in our example. Click on the **Insert** tab, in the charts area where we had earlier plotted charts such as line and pie, we also have **Scatter** option, choose the first graph icon from the **Insert Scatter** option next to **Pie Chart** option. The scatter plot is ready, as presented in Figure 4.11. To get a better visualization of data, we can use chart elements explained earlier. One additional feature which may be applicable in this chart is the Trendline.

- **Trendline:** It shows the nature of the trend that exists between the two variables by a line that best fits the data points.

The trendline we get in our example is upward sloping, indicating a positive relationship between advertising cost and items sold. With an increase in advertising costs, the sales of the company also go up. However, if we look at this, trendline does not tell us about the strength of the relationship. It is the subject matter of correlation to determine how strong a relationship is, which will be taken up in a later chapter. One critical practical application of scatter plot lies in simple linear regression which again will be discussed in the following chapter.

To recap, this lesson explored some essential tools of Excel, which helped us in the preparation of charts such as line, column, pie and scatter plots. We learned that during the process of data analysis, how we can identify and make the correct choice of the graph. Deciding on the right type of chart is a very crucial step in data analysis.

PRACTICE QUESTIONS

1. The following table gives the data of age and number of persons, corresponding to that age, in a particular locality.

Age	*Number of Persons*
18	5
21	25
24	15
27	40
31	46
35	92
37	70
39	64
42	33
45	42
49	24
51	46
55	28
57	14
60	7

Identify the type of graph suitable for given data and prepare it in Excel. Also, interpret the results.
Ans. Histogram.

2. The following table presents the annual data of Sunshine Company about its shoe sales.

Year	*Pair of Shoes Sold*
1990–1991	20
1991–1992	45
1992–1993	110
1993–1994	245
1994–1995	290
1995–1996	250
1996–1997	503
1997–1998	483
1998–1999	565
1999–2000	610

Choose the type of graph to be drawn for this firm and draw it in Excel.
Ans. Line graph.

3. The following data is about the household expenditure of Families A and B per month (in '000).

Items of Expenditure	*Family A*	*Family B*
Food	50	70
Clothing	30	50
Rent	25	35
Education	10	25
Medical	9	20
Others	15	30

Draw a suitable graph. On which item did Family A and Family B spend the most and the least?
Ans. Pie chart.

4. Prepare a suitable graph from the following data of a firm.

Year	*Exports (in ₹)*	*Imports (in ₹)*
2010–2011	459	305
2011–2012	320	290
2012–2013	350	340
2013–2014	380	400
2014–2015	300	280
2015–2016	280	315
2016–2017	350	370

Ans. Scatter graph.

5. The following table provides the birth rate (per thousand) of different countries.

Country	*Birth Rate*
India	35
Bangladesh	40
France	15
Australia	18
Brazil	25
Nigeria	42

Suggest and draw a suitable graph that can be drawn.
Ans. Column graph.

5

CHAPTER

Measures of Central Tendency

LEARNING OBJECTIVES

After reading this chapter, the readers will be able to understand

- Concept of descriptive statistics
- Measures of central tendency
- Meaning, calculation and interpretation of arithmetic mean using Excel
- Meaning, calculation and interpretation of median using Excel
- Meaning, calculation and interpretation of mode using Excel
- Calculation and application of weighted mean, geometric mean and harmonic mean using Excel

Descriptive statistics is the starting point of any data analysis. When we are working with a large volume of data, it becomes difficult to make sense of it. Therefore, it is important to identify a unique value that provides useful information capable of describing the data. Descriptive statistics *describe* the nature of the data set. It helps us to understand the characteristics of a large set of data in a more meaningful manner. Descriptive statistics include the two following categories of measurement of data:

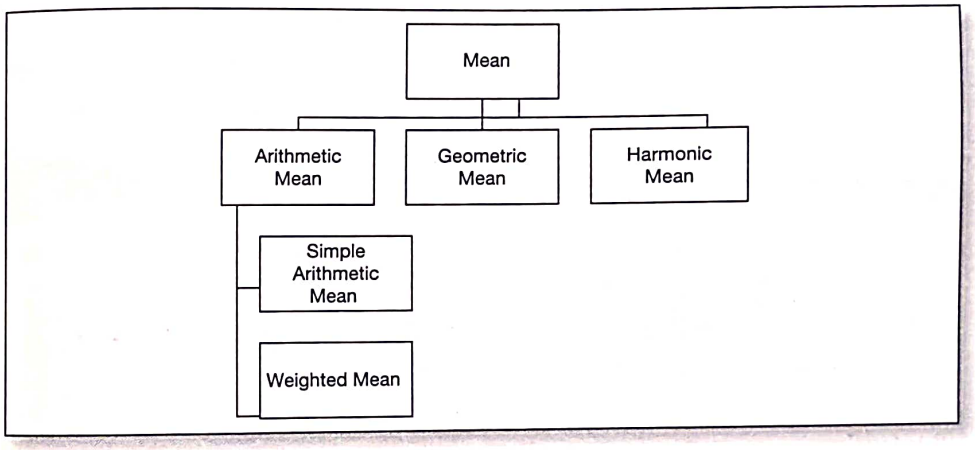
1. *Measures of central tendency or averages*
2. *Measures of dispersion or variability*

This chapter introduces the first measure of descriptive statistics, that is, measures of central tendency—their concept, their relevance, their application and how to calculate them. The most commonly used measures of central tendency are mean, median and mode.

A measure of central tendency is a *statistical measure* that summarizes the whole data set by identifying the central position of the data or the distribution. It is a single value that represents the centre or middle of a group of data where most of the values tend to lie. Three widely used measures of central tendency are as follows:

1. Mean
2. Median
3. Mode

Mean is further classified into the following categories based on the specific type of average required in different problems. We will explain these specific averages after we get through with the basic averages.



5.1. ARITHMETIC MEAN

The arithmetic mean, popularly known as the *average* or the *expected value*, is one of the most common and frequently used measures of central tendency. It reflects the central location of the distribution and represents the data as a whole. For example, the mean height of a group of 50 students in a class is *1.62 metres* or the average time spent watching television per viewer in India is *4 hours a day*.

The application of mean may include a company analysing its performance or position by using averages like *average net profits* earned over a specific period by it. Such data is available on the website of the respective company or the website of NSE (https://www1.nseindia.com/). Similarly, the arithmetic mean can also be used in finding out the average rating of apps on Google Play

Store. For example, 11,200 people rated an app between 1 and 5. Out of them, 3,000 rated 5, 2,500 rated 4 and so on. In this example, the average rating is 3.5625.

Rate		*No. of People*	*Rate × No. of People*
Good	5	3,000	15,000
	4	2,500	10,000
	3	4,000	12,000
	2	1,200	2,400
Bad	1	500	500
Total		11,200	39,900
Mean		39,900/11,200	3.5625

The mean is calculated in the following two simple steps.

Step 1. We add up all the values in our data set called observations.
Step 2. We divide the sum by the number of observations.

The formula then becomes the following:

$$\bar{X} = \frac{\text{Sum of all the observations}}{\text{Number of observations}}$$

$$\text{or } \bar{X} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where $\bar{X}$ = symbol or notation for mean, x_i = ith observation in the distribution, Σ = Sigma/operator for summation and n = number of observations.

Illustration

The distribution of income of 20 employees working in a multinational corporation (MNC) is given in the following table.

Employee ID	*Monthly Income*	*Employee ID*	*Monthly Income*
1	₹20,000	11	₹60,000
2	₹25,000	12	₹38,000
3	₹15,000	13	₹40,000
4	₹18,000	14	₹40,000
5	₹30,000	15	₹24,000
6	₹24,000	16	₹36,000
7	₹12,000	17	₹25,000
8	₹36,000	18	₹20,000
9	₹22,000	19	₹30,000
10	₹45,000	20	₹40,000

First Method (Direct Method)

Now, we need to calculate the mean monthly income of these 20 employees.

Mean monthly income ($\bar{X}$) = ₹(20,000 + 25,000 + 15,000 + 18,000 + 30,000 + 24,000 + 12,000 + 36,000 + 22,000 + 45,000 + 60,000 + 38,000 + 40,000 + 40,000 + 24,000 + 36,000 + 25,000 + 20,000 + 30,000 + 40,000) / 20
= ₹6,00,000 / 20
= ₹30,000

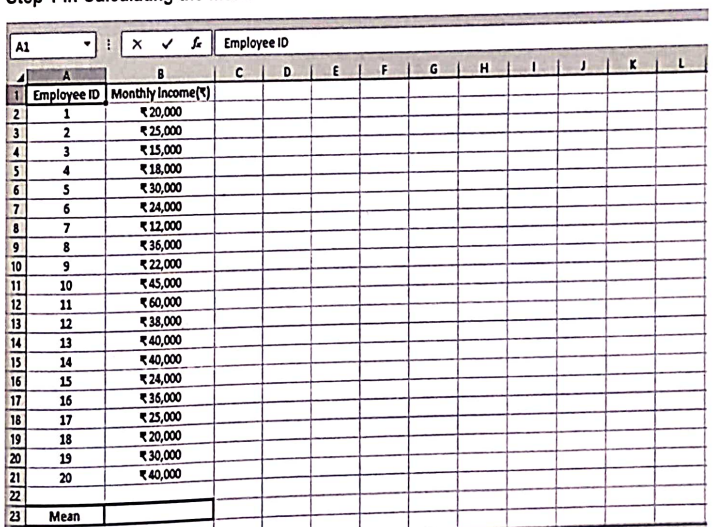
Here, the mean monthly income of ₹30,000 suggests that on average an employee working in this MNC earns ₹30,000 per month.

Second Method (Using Excel)

Step 1. Enter the employee ID and monthly income in an Excel sheet (Figure 5.1).

FIGURE 5.1

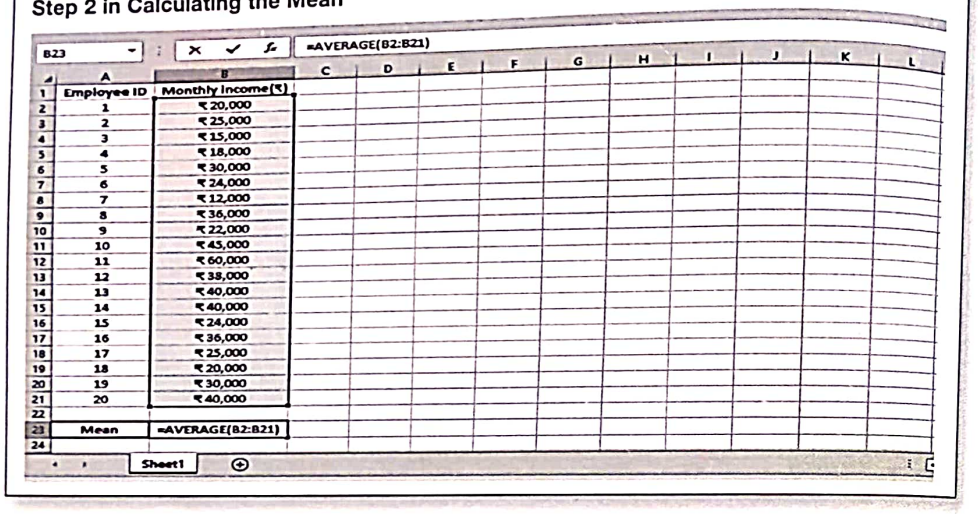
Step 1 in Calculating the Mean



Step 2. Use the **AVERAGE** function to calculate the mean. Type the '=' (equal to) sign in the cell where the result is desired, say 'B23', and enter the function name **AVERAGE** and then select the range of numbers for which average is to be calculated, that is, **B2:B21** (Figure 5.2).

FIGURE 5.2

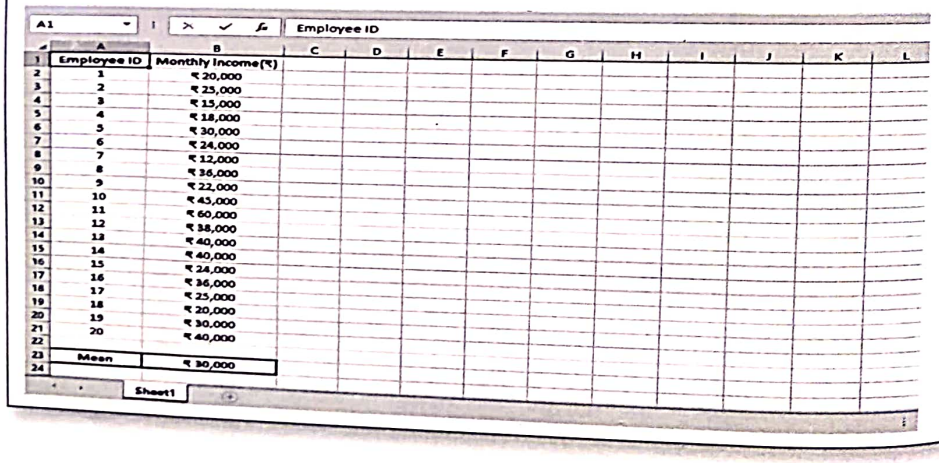
Step 2 in Calculating the Mean



Step 3. Press **Enter**. The result will be displayed in Cell B23 (Figure 5.3).

FIGURE 5.3

Step 3 in Calculating the Mean



5.1.1. Merits of Mean

1. It is easy to calculate and simple to understand.
2. It is reliable as it takes into account all the values in the distribution.
3. It is widely used in various fields such as mathematics, statistics, economics and finance.

5.1.2. Demerits of Mean

1. It is not possible to calculate mean if any observation is missing.
2. It cannot be calculated for qualitative or categorical data or data measured on a nominal and ordinal scale. For example,
   a. The mean honesty of employees working in a department cannot be calculated.
   b. Similarly, the mean colour of balloons in a room cannot be calculated.
3. Sometimes, it may give an impossible or inconsistent result.
   Let us see this with the help of an example, the following table shows the number of students in different years in a college.

Year	*Number of Students Enrolled*
First year	560
Second year	620
Third year	550
Fourth year	480
Total	**2,210**

   The average number of students in all years is
   = 2,210 / 4
   = 552.5
   which is impossible as the number of people cannot be in terms of decimals or fractions.
4. If there are one or more extremely high values or low values in the data set, the results may be misleading as mean is affected by those extreme values or outliers and does not represent a true picture.
   To understand this, let us consider the following example of the expenditure of Mr A for the year 2019.
   He spends ₹50,000 per month.
   In addition to this expense, he purchased a house worth ₹5,000,000 in August.

Now the average monthly expenditure of Mr A for the year 2019 can be calculated as follows.

Months	*Expenditure*
January	₹50,000
February	₹50,000
March	₹50,000
April	₹50,000
May	₹50,000

June	₹50,000
July	₹50,000
August	₹5,050,000
September	₹50,000
October	₹50,000
November	₹50,000
December	₹50,000
Total	₹5,600,000

Total expenditure for the year = ₹5,600,000
Number of months in a year = 12
Mean expenditure = ₹5,600,000 / 12
= ₹466,666.67 per month

- In this example, the result leads to a misleading estimation of the amount spent by Mr A on average in a month.
- Mr A spends exceptionally less than the mean expenditure in 11 out of the 12 months. It is the unusually high expenditure in August that leads to a high mean value.

Thus, the mean may prove to be a less effective measure of central tendency in the presence of extremely high or low values or outliers.

5.2. MEDIAN

The median is generally calculated to overcome the drawback of the mean. As noted in the preceding example, if there are one or more extremely high values or low values in the data set, the results may be misleading as mean is affected by those extreme values or outliers and does not represent a true picture.

Median refers to the middlemost number or the central value of the distribution when **sorted** in ascending or descending order. The median divides the data set into two halves, one which contains values more than the median and the other containing values less than the median, that is, 50 per cent of the values lie above the median and 50 per cent of the values lie below the median. Median is generally used to estimate the average property prices representing the general property market or the average per capita income. In India, the Ministry of Statistics and Program Implementation (http://mospi.nic.in/data) releases such summary statistics of GDP using the median.

Median is calculated with the help of the following steps:

Step 1. Arrange the series of numbers in ascending or descending order.
Step 2. Locate the middlemost number of the series.

- If there are *odd number of observations*, then there will be one value in the middle, that middlemost value is the median.
- If there are *even number of observations*, then there are two values in the middle and median is the average or the midpoint of those two values.

5.2.1. Formula

- If there are odd number of observations,

  then Median = $\left(\frac{n+1}{2}\right)$th term.

- If there are even number of observations,

  then Median = Average of $\left(\frac{n}{2}\right)$th term and $\left(\frac{n+1}{2}\right)$th term

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

where n = number of observations in the data set.

Illustration

The daily sales figures (in ₹) of a clothing store in Delhi, Mumbai and Bengaluru for the week are given in the following table. The stores in Delhi and Mumbai are open seven days a week, whereas the store in Bengaluru remains closed on Sunday. Calculate the median weekly sale for each city.

City	*Monday*	*Tuesday*	*Wednesday*	*Thursday*	*Friday*	*Saturday*	*Sunday*
Delhi	6,500	8,600	7,800	9,700	5,900	10,500	6,750
Mumbai	12,000	8,250	9,000	7,400	9,450	6,800	5,500
Bengaluru	4,500	9,650	8,700	8,650	5900	9200	–

Solution
First Method (Direct Method)

Step 1. We arrange the series in ascending order.

City							
Delhi	5,900	6,500	6,750	7,800	8,600	9,700	10,500
Mumbai	5,500	6,800	7,400	8,250	9,000	9,450	12,000
Bengaluru	4,500	5,900	8,650	8,700	9,200	9,650	–

Step 2. Find the middlemost value.

City	*Number of Observations*	*Median*
Delhi	7	$\left(\frac{n+1}{2}\right)$th term = $\left(\frac{7+1}{2}\right)$th term = 4th term = **₹7,800**

Mumbai	7	$\left(\frac{n+1}{2}\right)$th term = $\left(\frac{7+1}{2}\right)$th term = 4th term = ₹8,250
Bengaluru	6	$\frac{\left(\frac{n}{2}\right)\text{th term} + \left(\frac{n}{2}+1\right)\text{th term}}{2}$ $= \frac{\text{3rd term} + \text{4th term}}{2}$ = (₹8,650 + ₹8,700) / 2 = ₹8,675

The median sale for Delhi is ₹7,800, that is, half of the sale values are less than ₹7,800 and another half are more than ₹7,800.

Similarly, the median sales for Mumbai and Bengaluru are ₹8,250 and ₹8,675, respectively.

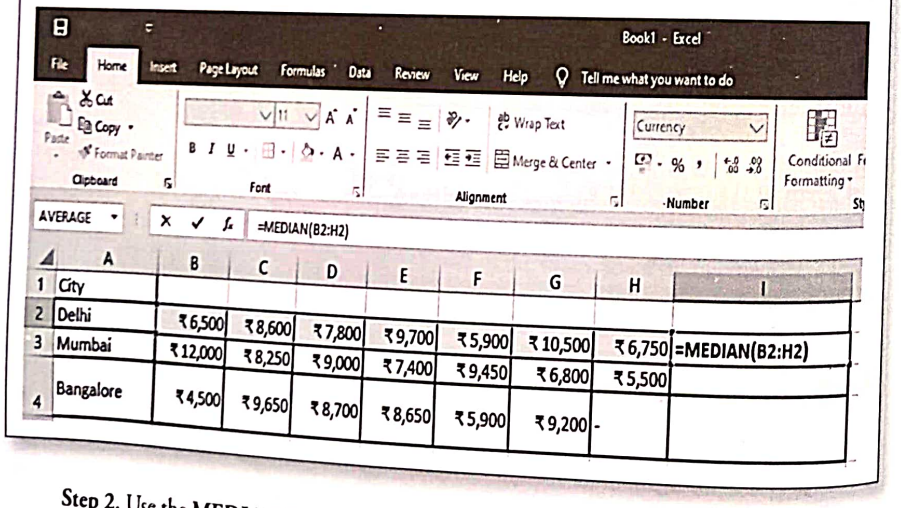
Second Method (Using Excel)

It is very easy to calculate the median using Excel.

Step 1. Enter the sales figures for each city in an Excel sheet. It is not necessary to sort the data in ascending order when calculating the median using Excel (Figure 5.4).

FIGURE 5.4

Step 1 in Calculating the Median

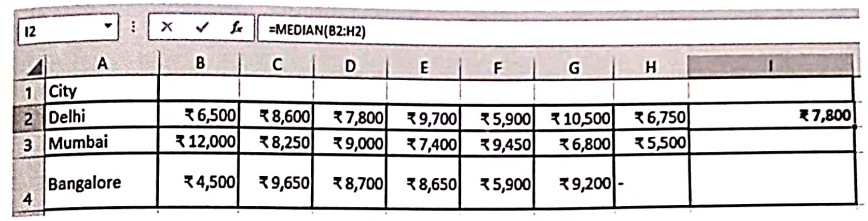


Step 2. Use the **MEDIAN** function to calculate the median. Type the '=' (equal to) sign in the cell where the result is desired, say 'I2' for Delhi and enter the function name **MEDIAN**

and then select the range of numbers for which median is to be calculated, that is, **B2:H2** for Delhi (Figure 5.5).

FIGURE 5.5

Step 2 in Calculating the Median

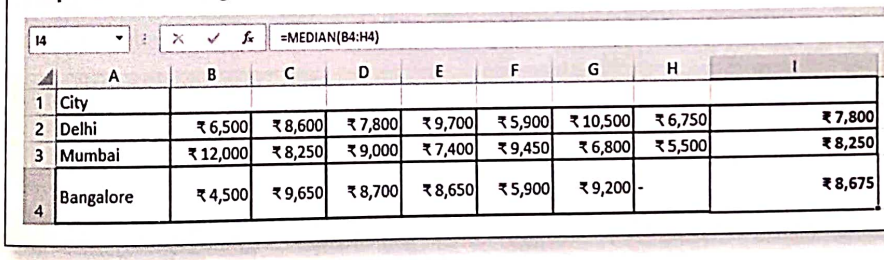


Step 3. Press **Enter**. The result will appear in Cell I2 (Figure 5.6).

Then click on the fill handle (the small black square in the bottom right corner when you select the cell) and drag it vertically to automatically calculate the median for the next two cities.

FIGURE 5.6

Step 3 in Calculating the Median



5.2.2. Merits of Median

1. It is simple to calculate and easy to understand.
2. Unlike the mean, the median is not affected by the extreme values in the data set. Thus, it is a better measure of central tendency in case of skewed distributions.
3. Median can be calculated for data measured on both interval and ordinal scale.
4. It can be determined graphically.

5.2.3. Demerits of Median

1. It is not based on all the values in the distribution as it does not utilize all the information in the data set.
2. To calculate the median, the data always needs to be arranged or sorted in order, that is, from least to greatest or vice versa.
3. Sometimes, it may give results that are not practically possible.

5.3. MODE

A mode is a measure of central tendency which helps to identify the value having the highest frequency. It represents the most common and frequently occurring number in the data set. The best thing about the mode is that it can be calculated for qualitative or categorical data.

The mode is generally useful and applied in situations that demand numbers. For example, we may be interested in popular YouTube videos—the one with the highest views/most downloaded. Similarly, in elections, the candidate seeking the maximum votes. The mode is also helpful for companies manufacturing popular products. The mode is calculated by counting the frequencies of all the values given in the dataset, that is, the number of times the value appears in the data set. The value(s) having the highest frequency, or which is repeated the maximum number of times in the data set is the mode. Thus, there is no mathematical formula to calculate the mode in case of individual or discrete series.

Illustration

Compute the modal age from the following data of age (in years) for 30 people who attended the economics conference.

20	27	35	41	50
22	27	36	45	52
25	30	38	47	56
25	32	38	47	56
25	32	40	48	56
25	32	40	49	57

Solution

First Method (Direct Method)

As we know, the mode is the value which repeats the maximum number of times in the data set.

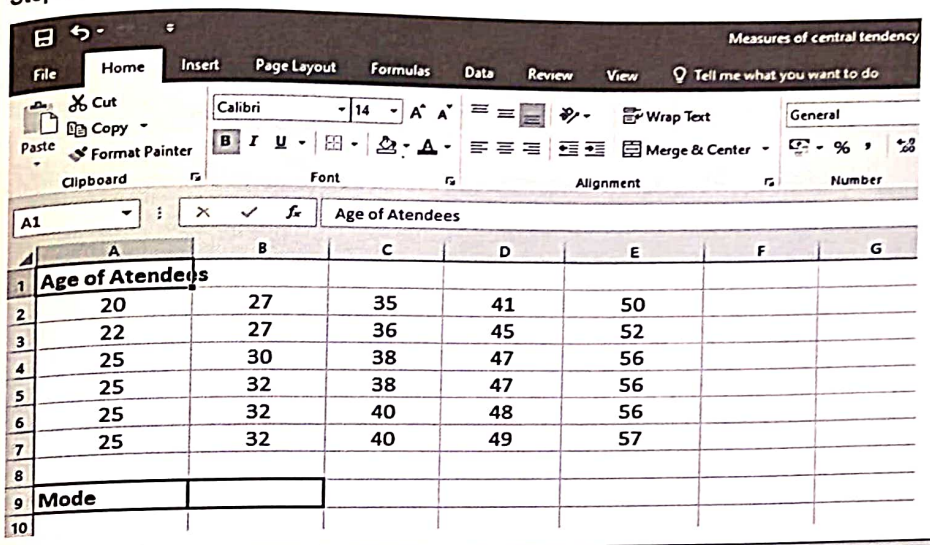
The age of 25 years appears 4 times in the data set which is the highest. Hence, the modal value is 25 years.

Second Method (Using Excel)

Step 1. Enter the data in the Excel sheet (Figure 5.7).

FIGURE 5.7

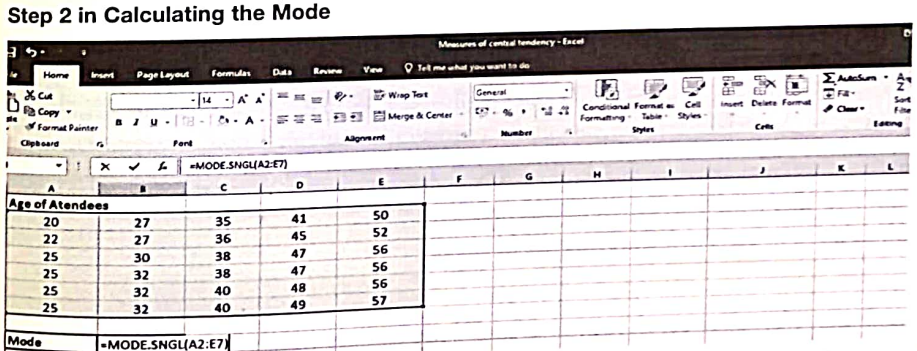
Step 1 in Calculating the Mode



Step 2. Use the **MODE.SNGL** function to calculate the mode. Type the '=' (equal to) sign in the cell where the result is desired, say 'B9', and enter the function name **MODE.SNGL** and then select the range of numbers for which mode is to be calculated, that is, **A2:E7** as shown in Figure 5.8.

FIGURE 5.8

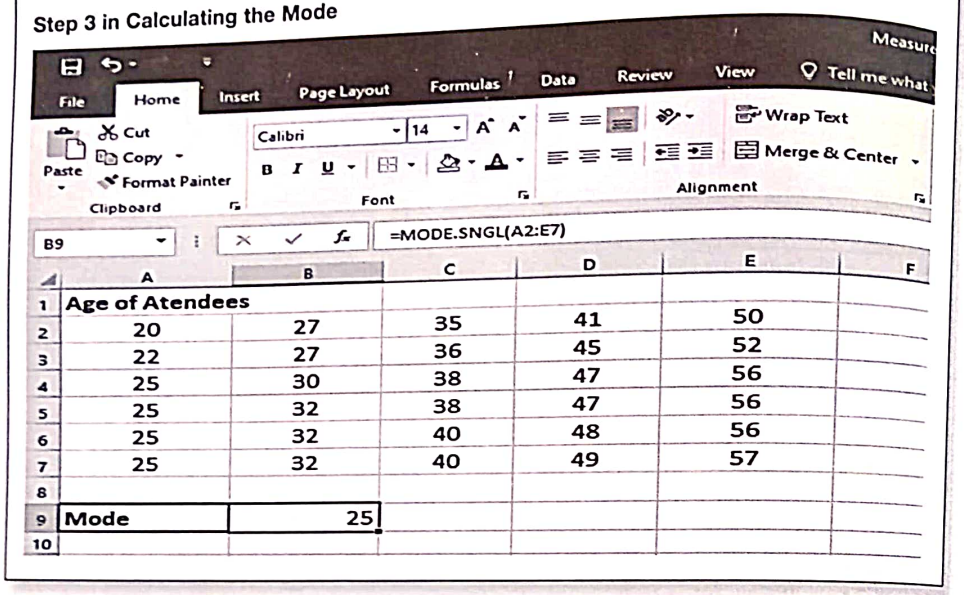
Step 2 in Calculating the Mode



Step 3. Press **Enter**. The result will be displayed in Cell B9 (Figure 5.9).

FIGURE 5.9

Step 3 in Calculating the Mode



It may be noted that the function **MODE.MULT** was introduced in Excel 2010. This function may replace **MODE.SNGL** function. The limitation of **MODE.SNGL** function is that it can be used only when there is a single mode present in the data set. However, the function **MODE.MULT** can be used both in the case of single and multiple modes. In case, there is no mode, it will return #N/A.

5.3.1. Using Mode If the Data Is Given in Discrete Series (Frequency Distribution)

The marketing head of a multi-brand store is interested in finding out the most popular shoe brand among the customers. The data of the purchases by customers from the store for the last week is summarized in the table as follows.

Brand	*Frequency*
Nike	67
Adidas	50
Reebok	36
Vans	46
Sketchers	58
New Balance	35
Puma	48
Total	**340**

Solution

First Method (Direct Method)

- To determine the most popular shoe brand, we calculate the mode.
- In the table, the brand 'Nike' has the highest frequency, that is, 67.
- This means that the brand 'Nike' appears the maximum number of times as the most purchased brand in the data set.
- So the marketing head can conclude that 'Nike' is the most popular shoe brand as it is the mode of the distribution.

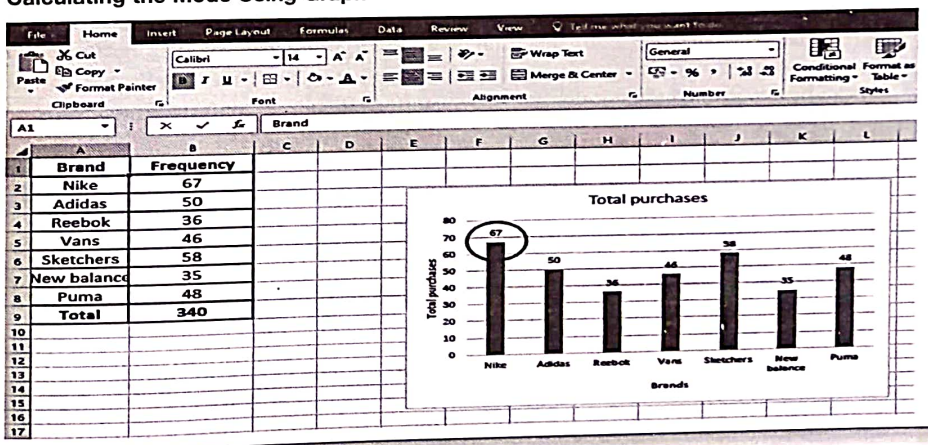
Second Method (Graphical Method)

One of the best things about the mode is that it can be determined graphically. In Chapter 4, we discussed in detail the various types of charts, their application and how to insert them in Excel. To compute mode, we can insert a bar graph depicting the frequencies of the variables in the distribution.

Step 1. Enter the data in an Excel sheet.
Step 2. Select the entire distribution, that is, **A1:B8**.
Step 3. Click on **Insert** tab and select **Insert Column** or **Bar Chart** and choose the first graph. The graph is displayed. Add chart elements (Figure 5.10).
Step 4. Identify the bar with the longest height depicting the highest frequency, that is, 67 for 'Nike'.

FIGURE 5.10

Calculating the Mode Using Graph



Third Method (Using Excel)

We can compute mode in case of discrete series (frequency distribution of the variables is known) using Excel function **MAX** which returns the largest number in the data set. This is useful in case of a huge volume of data.

Step 1. Enter the data in an Excel sheet.
Step 2. Type '**=MAX(B2:B8)**' in the cell where the result is desired, say B11 (Figure 5.11).
Step 3. Press **Enter**. The maximum number will be displayed in the cell as shown in Figure 5.12.

FIGURE 5.11

Step 1 in Calculating the Mode Using MAX

SUM | =MAX(B2:B8)

	A	B	C	D
1	Brand	Frequency		
2	Nike	67		
3	Adidas	50		
4	Reebok	36		
5	Vans	46		
6	Sketchers	58		
7	New balance	35		
8	Puma	48		
9	Total	340		
10				
11	Maximum Number(Mode)	=MAX(B2:B8)		
12				

FIGURE 5.12

Step 2 in Calculating the Mode Using MAX

B11 | =MAX(B2:B8)

	A	B	C	D
1	Brand	Frequency		
2	Nike	67		
3	Adidas	50		
4	Reebok	36		
5	Vans	46		
6	Sketchers	58		
7	New balance	35		
8	Puma	48		
9	Total	340		
10				
11	Maximum Number(Mode)	67		
12				

5.3.2. Merits of Mode

1. It is simple to understand and sometimes can be computed easily with a single glimpse of data.
2. It is not affected by the presence of extreme values in the data set.
3. It can be calculated for both quantitative and categorical data or data measured on a nominal scale.
4. It can be determined graphically.

5.3.3. Demerits of Mode

1. It is not based on all the values in the data set as it only takes into account the highest frequency.
2. Results may be unclear or ill-defined as the data set may have more than one mode in case of two or more values having the same highest frequencies. Similarly, a distribution may have no mode if none of the values repeat.

Now, we take up the special types of averages which were referred to at the beginning of the chapter.

5.4. WEIGHTED ARITHMETIC MEAN

In scenarios when the observations do not contribute equally, the application of arithmetic mean in averaging is erroneous. The contribution in the calculation of the average, that is, value or importance of observation as compared to others is given the term of *weights*. The weight given to observation is known as the *weighting factor*. Usually, the sum of all the weights equals to 1 or 100 per cent.

Let us understand the argument with the help of an example of portfolio investment where weighted mean is widely used. Suppose an investor invests ₹100,000 in 5 different investment options. He/she invests ₹20,000, ₹15,000, ₹25,000, ₹16,000 and ₹24,000 in investment options A, B, C, D and E respectively. The returns are 5 per cent, 7.5 per cent, 6 per cent, 10 per cent and 10 per cent from these investment options. In such cases, we cannot simply calculate the arithmetic mean to find the average return from this portfolio investment because we have invested different amounts in each investment category which has different returns.

To arrive at an appropriate average, we first determine the proportion of funds invested in each type of investment. These are weights, that is, the proportion of funds invested by an investor in each type of investment. In this example, weight for investment A is 20,000/100,000 = 0.2 or 20 per cent (funds invested in investment Option A / Total funds invested). Similarly, weights are 0.15, 0.25, 0.16 and 0.24 in the rest of the schemes, respectively. The following steps are taken to proceed with the calculation.

Step 1. Multiply the observations with their corresponding weights.
Step 2. Add up all the products calculated in Step 1.
Step 3. Divide the sum calculated in Step 2 with the sum of all the weights.

$$\textbf{Formula} = \bar{X}_w = \frac{\text{Sum of all the products of the values and their weight factor}}{\text{Sum of all the weights}}$$

or

$$\bar{X}_w = \frac{w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_nx_n}{\sum w_i} = \frac{\sum_{i=1}^{n} w_ix_i}{\sum w_i}$$

where $\bar{X}_w$ = Symbol or notation for weighted mean,
x_i = ith observation in the distribution,
w_i = Individual weight factor of ith observation,
Σ = Sigma/operator for summation and n = Number of observations.

Putting values in the formula

$$= \frac{0.2\times5\% + 0.15\times7.5\% + 0.25\times6\% + 0.16\times10\% + 0.24\times10\%}{1}$$

$$= 7.63\%$$

The average return earned by this investor is 7.63 per cent, given his/her portfolio investment returns.

Illustration

Let us consider another example in detail to further clarify the concept and calculation along with the application of Excel. The college decided to grade the students for a competition based on the following criteria.

Criteria	*Weightage*
Round 1 (easy)	20%
Round 2 (medium)	30%
Round 3 (difficult)	50%

Marks obtained by 4 groups out of 100 marks are as follows.

Groups	*A*	*B*	*C*	*D*	*Weights*
Round 1 (easy)	95	75	85	60	**0.2**
Round 2 (medium)	60	65	70	55	**0.3**
Round 3 (difficult)	50	80	70	95	**0.5**

Calculate the weighted score and compare it with the simple arithmetic mean of group.

First Method

The difficulty levels of the rounds are different, Round 1 is easy (weightage 20%), so the score of Round 1 contributes least to the overall performance of the group. Similarly, Round 3 is difficult (weightage 50%), so the score of Round 3 contributes the most to the overall performance of the group. So we calculate the weighted mean.

We use the formula

$$\bar{X}_w = \frac{w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_nx_n}{\sum w_i} = \frac{\sum_{i=1}^{n} w_ix_i}{\sum w_i}$$

Groups	*Sum of Products/Sum of Weights*	*Weighted Average*	*Ranking*	*Simple Average*	*Ranking*
A	= (95 × 0.2 + 60 × 0.3 + 50 × 0.5) / 1	62	IV	68.33	IV
B	= (75 × 0.2 + 65 × 0.3 + 80 × 0.5) / 1	74.5	II	73.33	II
C	= (85 × 0.2 + 70 × 0.3 + 70 × 0.5) / 1	73	III	75	I
D	= (60 × 0.2 + 55 × 0.3 + 95 × 0.5) / 1	76	I	70	III

The calculated weighted mean for Groups A, B, C and D is 62, 74.5, 73 and 76, respectively. Compared to it, the simple averages are 68.33, 73.33, 75 and 70. It may be noted that in this example, according to the simple average, the first rank is given to Group C as it has the highest mean score of 75. Comparing it with the weighted average, the first rank is given to Group D as it has the highest score of 76.

Second Method

Excel calculates weighted mean easily using the following steps.

Step 1. Enter the scores of all the groups for each round. Also, enter the corresponding weights for each round.

Step 2. Use the **SUMPRODUCT** function to calculate the weighted mean. Type the '=' (equal to) sign in the cell where the result is desired, say 'B7', and enter the function name

SUMPRODUCT and then select the range for **Array 1**, that is, the scores of Group A for each round, that is, **B2:B4**. For **Array 2**, select the range of the weights for each round, that is, **$F2:$F4**[a] (Figure 5.13).

Step 3. Press **Enter**. The result will appear in cell B7. Then click on **the fill handle** (the small black square in the bottom right corner when you select the cell) and drag it horizontally to automatically calculate the weighted mean for Groups B, C and D too (Figure 5.14).

FIGURE 5.13

Step 1 in Calculating the Weighted Arithmetic Mean

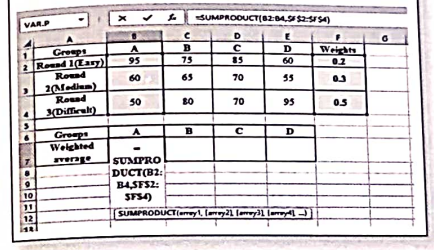
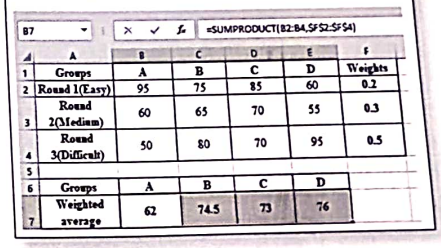


FIGURE 5.14

Step 2 in Calculating the Weighted Arithmetic Mean



5.5. GEOMETRIC MEAN

It is a special type of mean and is commonly used to take into account the *effect of compounding*. Therefore, the geometric mean is particularly used in the field of finance and demographic growth. In finance, it is used to calculate the average annual returns on investments over a specific period or average growth rate. Similarly, the population growth rate is calculated using this type of mean only.

Illustration

Let us understand its application in finance. For instance, the value of a firm increases from ₹100,000 to ₹110,000 in the first year and ₹165,000 in the second year. The growth rate in the first year will be 10 per cent and 50 per cent in the second year. The simple average of the growth rates will be 30 per cent. Using simple average growth, the firm's value after second year will be

$$= 100{,}000 \times (130\% \times 130\%)$$
$$= ₹169{,}000$$

We can notice that this value is *more than the actual* value since mean is additive. In contrast, the geometric mean is multiplicative and calculated as the nth root of the product of 'n' numbers. That is, to calculate the geometric mean for 'n' numbers, we multiply all the numbers and take the nth root of the product. In our example, calculating $\sqrt[2]{110 \times 150}$ yields 128.45

[a] We need to use absolute cell reference **$F2:$F4** in our formula so that it will not change when the formula is copied to other cells as the range **F2:F4** (weights for each round) remains constant for all the groups.

(equivalent to 28.45% compound annual growth). Using geometric mean, the firm's value after second year will be

$$= 100{,}000 \times (128\% \times 128\%)$$
$$= 165{,}000$$

This is approximately equal to the actual value. Hence, geometric mean is a better measure as it smooths the growth rates better. A generalized formula to calculate the geometric mean for n values is

$$= (\text{Product of all the numbers})^{1/n}$$
$$= \sqrt[n]{x_1 \times x_2 \times x_3 \times \ldots \times x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

where
Π = Notation for the product,
x_i = ith observation in the data set and
n = Number of observations.

Calculating Geometric Mean Using Excel

It must be realized by now that it is difficult to calculate geometric mean manually, at least if the size of the data is large, due to extensive calculations involved. It can, however, be calculated using the Excel function **GEOMEAN** in a few simple steps.

Step 1. Enter the data in an Excel spreadsheet.
Step 2. Use the **GEOMEAN** function to calculate the geometric mean (Figure 5.15). Type the '=' (equal to) sign in the cell where the result is desired, say 'C7', and enter the function name **GEOMEAN** and then select the range for which geometric mean is to be calculated, that is, C3:C4.
Step 3. Press **Enter**. The result will appear in cell C7 (Figure 5.16).

FIGURE 5.15

Step 1 in Calculating the Geometric Mean

	A	B	C	D
1	Year	Value of a firm	Growth rate	
2	0	1,00,000	-	
3	1	1,10,000	110.00%	
4	2	1,65,000	150.00%	
5				
6				
7	Geometric Mean		=GEOMEAN(C3:C4)	
8	Simple Mean			

FIGURE 5.16

Step 2 in Calculating the Geometric Mean

Function Library

C7 | =GEOMEAN(C3:C4)

	A	B	C	D
1	Year	Value of a firm	Growth rate	
2	0	1,00,000	-	
3	1	1,10,000	110.00%	
4	2	1,65,000	150.00%	
5				
6				
7	Geometric Mean		128%	
8	Simple Mean		130%	

A geometric mean of 128 per cent can be interpreted as the mean rate of growth of the firm value for two years which means if a firm value grows by 28.45 per cent uniformly over the two years, then starting with 100,000 it would reach 165,000. The deduction of 100 is due to the reason that the base value taken at the beginning of the period is 100. To derive the compound annual growth rate, the value of 100 taken initially is to be deducted from the value arrived at the end of the period. Therefore, whenever we talk of growth over a period, it is a geometric mean and not the arithmetic mean.

5.6. HARMONIC MEAN

Harmonic mean is another type of average which refers to the reciprocal of the simple average of the reciprocals of the observations. This type of mean gives large weightage to smaller values and smaller weightage to larger values. This mean is usually used with rates, ratios (fractions) like average speed or price earnings ratio. For example, a car travels at a speed of 30 km/hr from A to B and 45 km/hr from B to A. Then the *average speed is not the arithmetic mean*, that is (30 + 45) / 2 km/hr = 37.5km/hr.

The *average speed is harmonic mean*, that is $\dfrac{2}{1/30 + 1/45} = 36.36$ km/hr

Let us try to understand it by the following explanation.

- Speed = Distance/Time
- Time = Distance/Speed
  Time (from A to B) = D/30
  Time (from B to A) = D/45
  Let us say the distance is D
  Average speed = Total distance/Total time

$$= \frac{\text{Distance (from A to B)} + \text{Distance (from B to A)}}{\text{Time (from A to B)} + \text{Time (from B to A)}}$$

$$= \frac{2D}{\frac{D}{30} + \frac{D}{45}} = \frac{2D}{D\left(\frac{1}{30} + \frac{1}{45}\right)} = \frac{2}{1/30 + 1/45} = 36.36 \text{ km / hr}$$

A generalized formula of the harmonic mean is as follows:

$$\text{Harmonic mean} = \frac{\text{Number of observations}}{\text{Sum of reciprocals of given values}}$$

$$= \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \ldots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_n}}$$

where
x_i = ith observation in the distribution,
Σ = Sigma/operator for summation and
n = Number of observations.

The steps involved are as follows.

Step 1. Calculate the reciprocals of all the values in the data set.
Step 2. Compute the simple arithmetic mean of the values calculated in Step 1.
Step 3. Calculate the reciprocal of the value obtained in Step 2.

5.7. RELATIONSHIP BETWEEN ARITHMETIC MEAN, GEOMETRIC MEAN AND HARMONIC MEAN

1. The value of geometric mean is always less than or equal to arithmetic mean.
   The value of harmonic mean is always less than or equal to geometric mean.
   Therefore, we can conclude that
   Arithmetic mean ≥ *Geometric mean* ≥ *Harmonic mean*
2. Arithmetic mean, geometric mean and harmonic mean are Pythagorean means, that is $(\textit{Geometric mean})^2 = \textit{Arithmetic mean} \times \textit{Harmonic mean}$.

To conclude, this chapter focused on the significance of the measures of central tendency to reduce a large amount of data to a single value that represents and summarizes the data as a whole by identifying its central position. Further, we learned how to compute mean (arithmetic, geometric and harmonic), median and mode and apply them under different conditions. The question arises whether it is adequate or do we need to compute additional statistics to make the data more meaningful. We will find out the answer to this in our next chapter where we will learn the computation and application of measures of dispersion or variability.

PRACTICE QUESTIONS

1. The sales targets achieved by 20 sales representatives of P Ltd are summarized as follows. Calculate mean and median for the data.

Sales Representative	*Sales*	*Sales Representative*	*Sales*
1	213,000	11	401,000
2	489,000	12	270,000
3	280,000	13	228,000
4	396,000	14	452000
5	111,000	15	471,000
6	379,000	16	464,000
7	297,000	17	181,000
8	306,000	18	207,000
9	447,000	19	390,000
10	383,000	20	372,000

Ans. Mean = 336,850; Median = 375,500.

2. The maximum temperature recorded for November in Delhi is stated as follows. Calculate the median temperature and mode.

Day	*Temperature (Celsius)*	*Day*	*Temperature (Celsius)*	*Day*	*Temperature (Celsius)*
1	33	11	31	21	31
2	33	12	28	22	32
3	36	13	26	23	34
4	31	14	30	24	32
5	33	15	33	25	33
6	32	16	35	26	28
7	32	17	32	27	30
8	32	18	30	28	36
9	33	19	31	29	34
10	34	20	32	30	32

Ans. Median = 32; Mode = 32.

3. XYZ Ltd earned net profit (million US$) as follows in the last 10 years. Calculate the geometric mean.

Year	*Net Profit (Million US$)*
2019	67,160
2018	64,660
2017	63,500
2016	62,700
2015	63,050
2014	66,600
2013	66,400
2012	65,400
2011	57,830
2010	51,250

Ans. 62,669.18

4. Ten cars travelled from Delhi to Chandigarh.
   The table shows the petrol consumption for each car. Calculate the harmonic mean to estimate the average consumption of petrol per car.

Car	*Petrol Consumption (Litres per Km)*
1	20
2	18
3	15
4	21
5	18
6	17
7	19
8	12
9	19
10	24

Ans. 17.71531.

Illustration

The test scores of 15 students in an examination are given as follows. Calculate the harmonic mean for scores.

Student	*Scores*
1	90
2	65
3	48
4	76
5	84
6	62
7	94
8	33
9	59
10	86
11	92
12	80
13	71
14	55
15	78

Solution

First Method (Direct Method)

$$\text{Using Harmonic mean} = \frac{\text{Number of observations}}{\text{Sum of reciprocals of given values}}$$

n = 15

Scores
1/90
1/65
1/48
1/76
1/84
1/62
1/94
1/33
1/59
1/86
1/92
1/80

1/71
1/55
1/78
Sum = 0.2264955

$$\sum_{i=1}^{n} \frac{1}{x_n} = 0.2264955$$

$$\text{Harmonic mean} = \frac{15}{0.2264955}$$

$$= 66.2264$$

Second Method (Using Excel)

It is extremely difficult to calculate harmonic mean manually for large data. It can be calculated using the Excel function **HARMEAN** in the following steps.

Step 1. Enter the data in an Excel spreadsheet.
Step 2. Use the **HARMEAN** function to calculate the harmonic mean.
Type the '=' (equal to) sign in the cell where the result is desired, say 'B18' and enter the function name **HARMEAN** and then select the range for which harmonic mean is to be calculated, that is, B2:B16 (Figure 5.17).
Step 3. Press **Enter**. The result will appear in cell B18 (Figure 5.18).

FIGURE 5.17

Step 1 in Calculating the Harmonic Mean

SUM | =HARMEAN(B2:B16)

	A	B	C	D
1	Student	Scores		
2	1	90		
3	2	65		
4	3	48		
5	4	76		
6	5	84		
7	6	62		
8	7	94		
9	8	33		
10	9	59		
11	10	86		
12	11	92		
13	12	80		
14	13	71		
15	14	55		
16	15	78		
17				
18	Harmonic Mean	=HARMEAN(B2:B16)		
19				

Sheet1

FIGURE 5.18

Step 2 in Calculating the Harmonic Mean

B18 | =HARMEAN(B2:B16)

	A	B	C	D
1	Student	Scores		
2	1	90		
3	2	65		
4	3	48		
5	4	76		
6	5	84		
7	6	62		
8	7	94		
9	8	33		
10	9	59		
11	10	86		
12	11	92		
13	12	80		
14	13	71		
15	14	55		
16	15	78		
17				
18	Harmonic Mean	66.22646972		
19				

Sheet1

5. A college decided to grade the students for a semester based on the following criteria.

Criteria	*Weightage*
Assignment	10%
Test	20%
Exam	60%
Viva	10%

Marks obtained by Shipra and Shraiya are as follows.

Groups	*Shipra*	*Shraiya*	*Weights*
Assignment	80	94	10%
Test	85	65	20%
Exam	92	70	60%
Viva	90	86	10%

Calculate the weighted score.
Ans. Shipra = 89.2; Shraiya = 73.

6

CHAPTER

Measures of Dispersion

LEARNING OBJECTIVES

After reading this chapter, the readers will be able to understand

- Concept of measure of dispersion
- Meaning, calculation and interpretation of range using Excel
- Meaning, calculation and interpretation of standard deviation using Excel
- Meaning, calculation and interpretation of variance using Excel
- Calculation and application of coefficient of variation
- Concept of symmetrical and asymmetrical distribution
- Computation of coefficient of skewness using Excel and its interpretation
- Concept and calculation of kurtosis using Excel
- Using data analysis tools to compute descriptive statistics

In the previous chapter, we discussed in detail the first measurement tool of descriptive statistics, that is, measures of central tendency. This chapter will introduce the second category of descriptive statistics, that is, measures of dispersion or variability. As the name suggests, it describes how spread out the data set is. Commonly used statistical measures of dispersion are range, standard deviation, variance and coefficient of variation. The extent to which the individual items vary or differ from the central value of the data set is called *dispersion*. Besides, this chapter will also introduce the *skewness of distribution and kurtosis*.

The need to calculate the measures of dispersion can be understood with the help of the following table which summarizes the test scores in two different sections of 10 students each.

											Mean	Median
Class A	70	56	35	99	96	100	80	88	59	42	72.5	75
Class B	76	82	61	60	74	90	78	62	65	77	72.5	75

Here, both the data sets have an exactly equal arithmetic mean of 72.5 and a median of 75. But we cannot say that the performance of both sections is the same. Looking closely, we find that in the first set, the numbers are *farther away from the mean score of 72.5* (highest score being 100 and lowest is 35), whereas the numbers in the second set are much *closer to the mean score* (90 being the highest score and lowest is 60). It implies that the values in the *first set of numbers are more disperse* as the differences in scores are comparatively higher.

It is obvious then that the measures of central tendency are not adequate summaries of the data and we need additional measures related to the amount of variation in the set of data. Plotting graphs of the above data provides a much clear understanding of variability in the scores. The first set of Class A scores is largely scattered (Figure 6.1), whereas the scores in the second set have greater uniformity/concentrated (Figure 6.2).

In the light of the above, we explain below the aforesaid measures of degree of variability present in the data set. Any good measure of it should meet the following criteria:

FIGURE 6.1

Scatter Plot for Scores of Class A

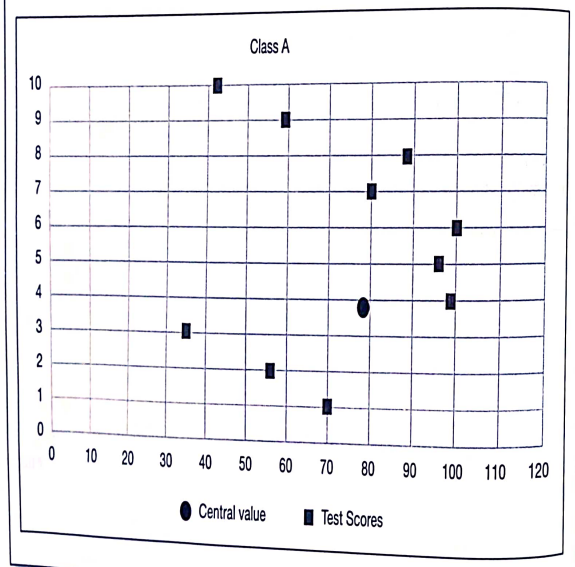
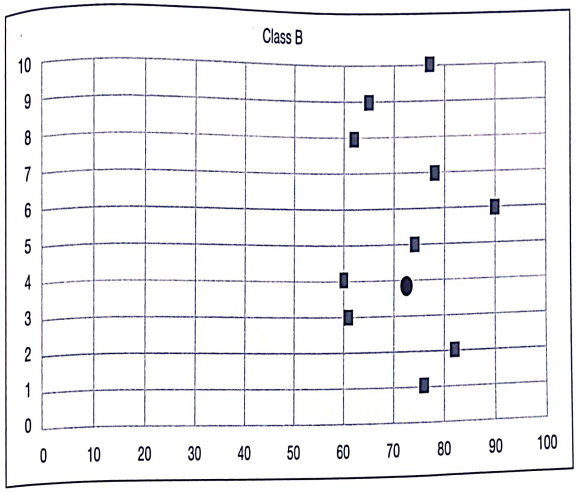


FIGURE 6.2

Scatter Plot for Scores of Class B



1. It should be simple to calculate and easy to understand.
2. It should be based on all the values in the data set which will make it more reliable.
3. It should not be affected by the presence of extreme values or outliers.
4. It should facilitate further statistical analysis.
5. It should be able to depict true and accurate variability in the data.

6.1. RANGE

The range is the simplest measure of dispersion. A higher range implies higher variation or spread in the data. It is the difference between the highest and the lowest values in the data set. It is, therefore, calculated by subtracting the minimum value from the maximum value in the data set.

Therefore, Range = Maximum value of data – Minimum value of data

In the example cited above, the range of Class A score will be
= Maximum score – Minimum score
= 100 – 35 = 65

Correspondingly, range of Class B score will be
= Maximum score – Minimum score
= 90 – 60 = 30

Illustration

Let us consider the historical data on exchange rate for US Dollar (USD) to Indian Rupee (INR) for 12 months starting from January 2019 to December 2019.

Calculate the range for the data given as follows.

Date	*USD/INR*
December	71.311
November	71.722
October	70.99
September	70.4987
August	71.6802
July	69.0775
June	69.325
May	69.98
April	69.8118
March	69.43
February	71.19
January	71.12

Source: Yahoo Finance. https://finance.yahoo.com/quote/INR%3DX/history? period1=1563608203&period2 =1595230603&interval =1mo&filter=history&frequency=1mo

First Method (Direct Method)

Range of exchange rates
= Maximum exchange rate – Minimum exchange rate
= 71.722 – 69.0775 which is equal to 2.6445

Second Method (Excel)

To calculate the range using Excel, follow these steps.

Step 1. Enter the exchange rates in an Excel sheet.

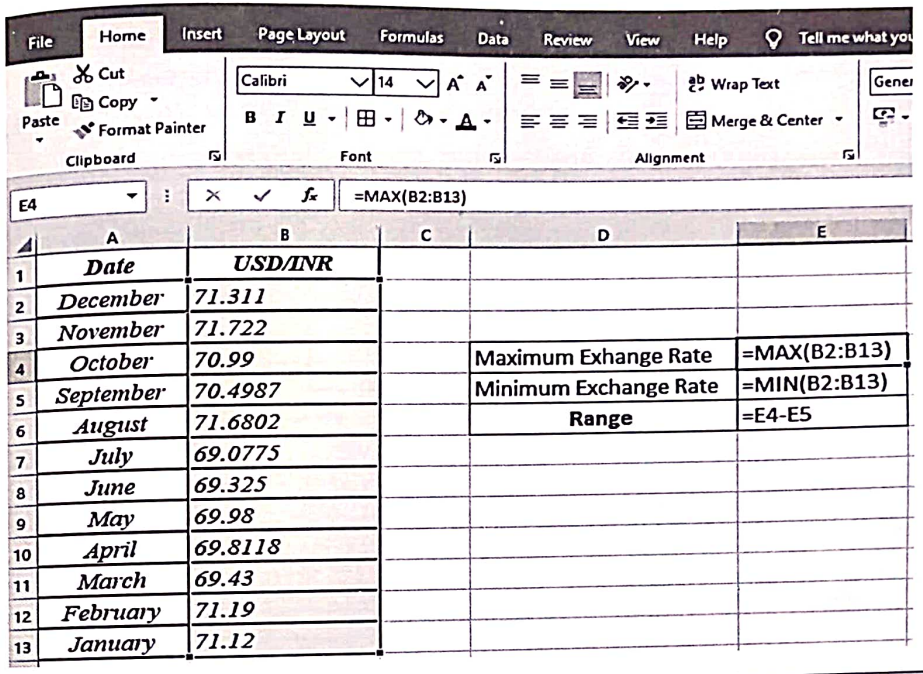
Step 2. Use the **MAX** and **MIN** function to calculate the highest exchange rate and lowest exchange rate. Range will be the difference between the highest and lowest values.
Type the '=' (equal to) sign in the cell where the result is desired, say 'E4' and enter the function name **MAX** and then select the range of numbers to calculate the highest exchange rate, that is, **B2:B13**. Press **Enter.**

Step 3. Type the '=' (equal to) sign in the cell where the result is desired, say 'E5' and enter the function name **MIN** and then select the same range of numbers to calculate the lowest exchange rate, that is, **B2:B13**. Press **Enter.**

Step 4. To calculate **Range**, type the '=' (equal to) sign in the cell where the result is desired, say 'E6' and then subtract the result in E5 from E4. Press **Enter** (Figures 6.3 and 6.4).

FIGURE 6.3

Step 1 in Calculation of Range



6.1.1. Merits of Range

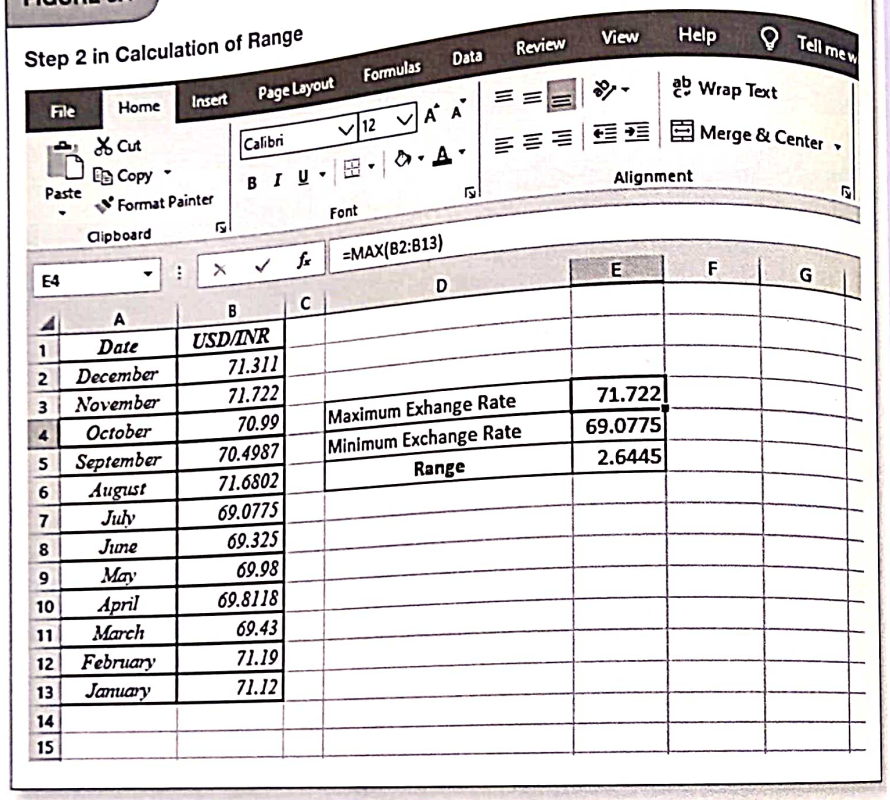
1. It is very easy to compute.
2. It is simple to understand even by a layman.
3. It is less time-consuming.

6.1.2. Demerits of Range

1. It is not based on all the values in a data set.
2. It does not facilitate further analysis.
3. In the case of open-end continuous series, the upper limit of the highest class and lower limit of the lowest class have to be estimated.

FIGURE 6.4

Step 2 in Calculation of Range



	A	B	C	D	E	F	G
1	Date	USD/INR					
2	December	71.311					
3	November	71.722					
4	October	70.99		Maximum Exhange Rate	71.722		
5	September	70.4987		Minimum Exchange Rate	69.0775		
6	August	71.6802		Range	2.6445		
7	July	69.0775					
8	June	69.325					
9	May	69.98					
10	April	69.8118					
11	March	69.43					
12	February	71.19					
13	January	71.12					
14							
15							

6.2. STANDARD DEVIATION

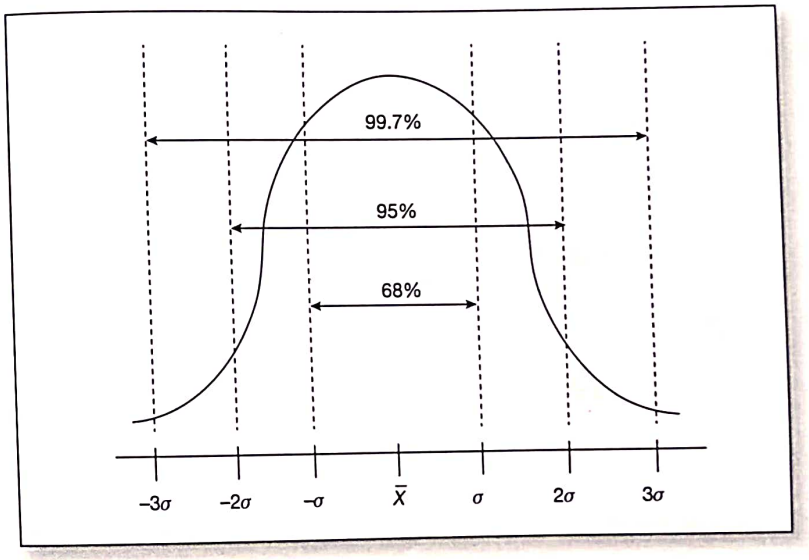
Standard deviation (denoted by the Greek letter sigma) is the *square root of the sum of all the squared deviations of the observations from their mean divided by the number of observations in the data set.* Or in other words, the square root of the average of the squared deviations from the mean is the standard deviation. A large standard deviation indicates that the data has high variation or is widely spread. On the other hand, a low standard deviation means lower variation in the data set. It is calculated with the help of the following steps.

Step 1. Calculate the arithmetic mean for a set of observations.
Step 2. Subtract the arithmetic mean calculated in Step 1 from each observation. The difference between each observation and the arithmetic mean is called a deviation from the mean.

Step 3. Square the deviations calculated above.
Step 4. Calculate the arithmetic mean of the squared deviations.
Step 5. Calculate the square root of the result obtained in Step 4.

Rule of Thumb (Assumes Normal Distribution)

- We expect 68 per cent of our data to fall within the range + or – 1 standard deviation of the mean (± 1 × std. dev.)
- We expect 95 per cent of our data to fall within the range + or – 2 standard deviation of the mean (± 2 × std. dev.)
- We expect 99.7 per cent of our data to fall within the range + or – 3 standard deviation of the mean (± 3 × std. dev.)
- 0.3 per cent values falling beyond the third range are considered to be rare and unusual



The aforesaid steps of standard deviation for a series of numbers $x_1, x_2, x_3, \ldots, x_n$ can be formalized as

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2}{n}}$$

where σ = Symbol for standard deviation,
Σ = Sigma/operator for summation,
$\bar{X}$ = Arithmetic mean,
$(x_i - \bar{X})^2$ = Squared deviation of ith observation from the arithmetic mean and
n = Number of observations.

Important note: When we calculate the standard deviation for a sample to infer about the population as a whole from which it is drawn, we use $n-1$ instead of n in the denominator, as it gives a better measure of standard deviation.

Population refers to the aggregate of all the items under study. Whereas the sample is the set of items taken out of the population as a representative of it.

Therefore, the population standard deviation is = $\sqrt{\dfrac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$

and the sample standard deviation is = $\sqrt{\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$.

Here, μ = population mean, $\bar{x}$ = sample mean, N = number of observations in population and n = number of observations in a sample.

Illustration

The following table presents the monthly salaries of CEOs of 10 MNCs. Calculate the standard deviation of salaries.

Company	*Salary ('000)*
A	100
B	120
C	400
D	280
E	90
F	250
G	340
H	170
I	300
J	600

First Method (Direct Method)

Company	*Salary ('000) (X)*	*X-Mean*	*(X-Mean)²*
A	100	–165	27,225
B	120	–145	21,025
C	400	135	18,225
D	280	15	225
E	90	–175	30,625
F	250	–15	225
G	340	75	5,625
H	170	–95	9,025
I	300	35	1,225
J	600	335	112,225
Total	2,650		225,650

Mean of salaries ('000) = $\dfrac{\sum_{i=1}^{n} x_i}{n}$ = 2650 / 10 = ₹265

Standard deviation ('000) = $\sqrt{\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}} = \sqrt{\dfrac{225,650}{10}}$ = ₹150.21651

- According to the rule of thumb, this means that the majority of the salaries (68%) lie within the range of ₹265,000 – (1) × ₹150,216.5 and 265,000 + (1) × 150,216.5, that is, within ₹114,784 and ₹415,217.

Second Method (Using Excel)

Use the **STDEV.P** function to calculate the standard deviation of a population. If using the entire population is not possible, we can calculate the standard deviation of a sample drawn from the population using the Excel function **STDEV.S**.

To calculate the standard deviation using Excel, follow these steps as displayed in Figures 6.5 and 6.6.

FIGURE 6.5

Step 1 in Calculation of Standard Deviation

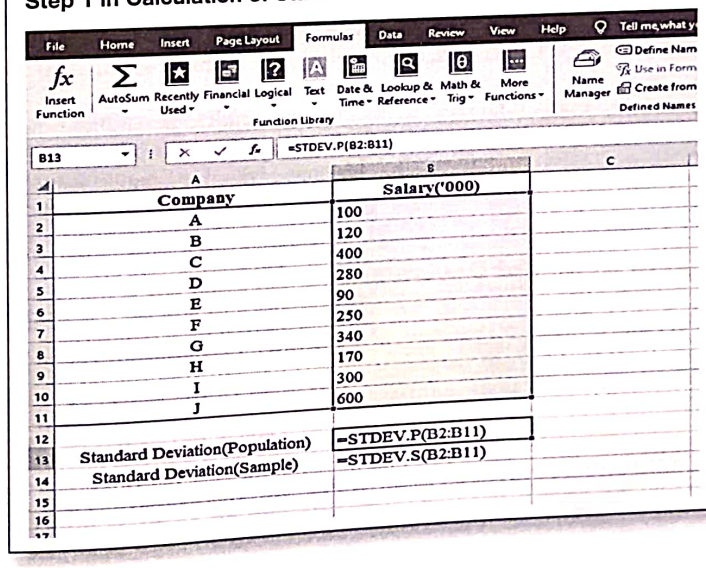
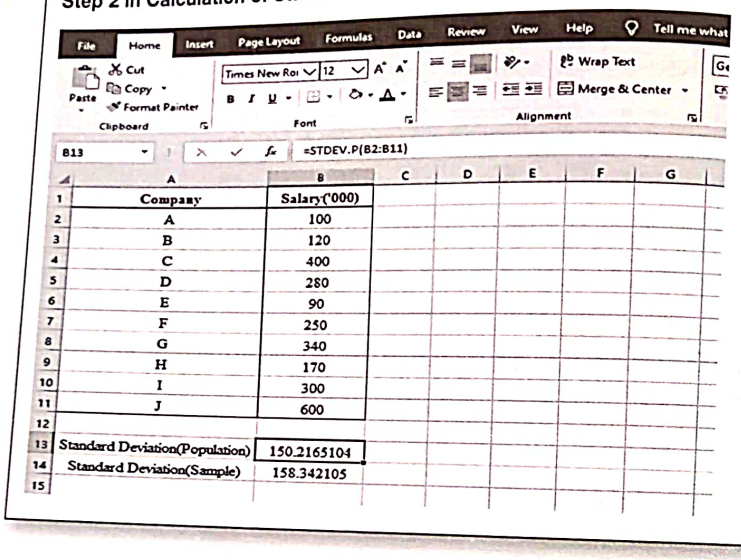


FIGURE 6.6

Step 2 in Calculation of Standard Deviation



	A	B
1	Company	Salary('000)
2	A	100
3	B	120
4	C	400
5	D	280
6	E	90
7	F	250
8	G	340
9	H	170
10	I	300
11	J	600
12		
13	Standard Deviation(Population)	150.2165104
14	Standard Deviation(Sample)	158.342105
15		

Step 1. Enter the salaries in an Excel sheet.

Step 2. Use the **STDEV.P** function to calculate the standard deviation of the population. Type the '=' (equal to) sign in the cell where the result is desired, say 'B13' and enter the function name **STDEV.P** and then select the range of numbers, that is, **B2:B11**. Press **Enter**.

Step 3. Use the **STDEV.S** function to calculate the standard deviation of the sample. Type the '=' (equal to) sign in the cell where the result is desired, say 'B14' and enter the function name **STDEV.S** and then select the range of numbers, that is, **B2:B11**. Press **Enter** (Figures 6.5 and 6.6).

6.2.1. Merits of Standard Deviation

1. It describes the spread or the variation of the data set from their average value.
2. It is based on all the values of the data set.
3. It indicates the reliability of the measures of central tendency.
4. It enables a comparative study of two or more data sets with respect to their consistency, stability or volatility.
5. It is generally taken as an estimation of risk in an investment and helps to measure the market volatility.

6. It facilitates further statistical analysis such as skewness, correlation, regression and hypothesis testing.
7. It is expressed in the same units as the mean of the distribution, hence easier to interpret and more accurate.

6.2.2. Demerits of Standard Deviation

1. The calculation of standard deviation is difficult as it involves many steps.
2. It cannot be easily understood.
3. The standard deviation of two series cannot be added as it is a square root.
4. It is affected by extreme values in the data set because the deviations from mean are squared due to which the larger deviation becomes proportionately bigger than the smaller deviation.
5. Units of measurement do influence the value of the standard deviation. For example, if we measure some length in terms of inches and measure the same in terms of centimetres, then the standard deviation in the latter will be 2.53 times the former.

6.3. VARIANCE

Now let us look at another important measure of dispersion called variance. Variance refers to the average of the squared deviations from the mean. In simple words, it is the square of the standard deviation. It is denoted by (σ^2) referred to as sigma square. Mathematically, the variance for a series of numbers $x_1, x_2, x_3, \ldots, x_n$ is expressed as

$$\sigma^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2}{n}$$

where

σ = Symbol for standard deviation,
Σ = Sigma/operator for summation,
$\bar{X}$ = Arithmetic mean,
$(x_i - \bar{X})^2$ = Squared deviation of ith observation from the arithmetic mean and
n = Number of observations.

Illustration

Two teams played 10 matches for the national-level cricket tournament in Delhi. The scores of the two teams for all the matches in the tournament are given as follows.

Match	*Team A*	*Team B*
1	150	260
2	236	110
3	280	140

Match	Team A	Team B
4	175	265
5	210	289
6	190	300
7	185	95
8	146	70
9	250	145
10	235	290

Calculate

1. Standard deviation
2. Variance
3. Which team has consistent performance?

First Method

Match	Team A	Team B	Team A (X-Mean)	Team B (X-Mean)	Team A (X-Mean)²	Team B (X-Mean)²
1	150	260	–55.7	63.6	3,102.49	4,044.96
2	236	110	30.3	–86.4	918.09	7,464.96
3	280	140	74.3	–56.4	5,520.49	3,180.96
4	175	265	–30.7	68.6	942.49	4,705.96
5	210	289	4.3	92.6	18.49	8,574.76
6	190	300	–15.7	103.6	246.49	10,732.96
7	185	95	–20.7	–101.4	428.49	10,281.96
8	146	70	–59.7	–126.4	3,564.09	15,976.96
9	250	145	44.3	–51.4	1,962.49	2,641.96
10	235	290	29.3	93.6	858.49	8,760.96
Total	2,057	1,964			1,7562.1	76,366.4

For Team A

Mean of scores = $\frac{\sum_{i=1}^{n} x_i}{n}$ = 2,057 / 10 = 205.7

a. Standard deviation = $\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}} = \sqrt{\frac{1,7562.1}{10}}$ 41.90715929

b. Variance = (Standard deviation)² = 1,756.21

For Team B

Mean of scores = $\frac{\sum_{i=1}^{n} x_i}{n}$ = 1,964 / 10 = 196.4

a. Standard deviation = $\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}} = \sqrt{\frac{76,366.4}{10}} =$ = 87.38787

b. Variance = (Standard deviation)² = 7,636.64

Based on the results, Team A performed consistently as their standard deviation, that is, total variation is less than that of Team B. However, variance describes the spread in the data set, but no conclusions can be drawn from it. It is the standard deviation that gives us a better picture of the variation in the data set with the help of the rule of thumb (68–95–99.7 rule). For instance, in the above illustration

- The majority of the scores (68%) of Team A lie within the range of 205.7–41.90 and 205.7 + 41.90, that is, *within 163.8–247.6.*
- The majority of the scores (68%) of Team B lie within the range of 196.4–87.38 and 196.4 + 87.38, that is, *within 109.02–283.78.*

Second Method (Using Excel)

Use the **VAR.P** function to calculate the variance of a population. If using the entire population is not possible, we can calculate the variance of a sample drawn from the population using the Excel function **VAR.S**.

To calculate the variance using Excel, follow the steps as shown in Figures 6.7 and 6.8.

FIGURE 6.7

Step 1 in Calculation of Variance

B13 | =VAR.P(B2:B11)

	A	B	C	D	E	F	G
1	Match	Team A	Team B				
2	1	150	260				
3	2	236	110				
4	3	280	140				
5	4	175	265				
6	5	210	289				
7	6	190	300				
8	7	185	95				
9	8	146	70				
10	9	250	145				
11	10	235	290				
12							
13	Variance for population	=VAR.P(B2:B11)	=VAR.P(C2:C11)				
14	Variance for sample	=VAR.S(B2:B11)	=VAR.S(C2:C11)				
15							
16							

FIGURE 6.8

Step 2 in Calculation of Variance

B13 | =VAR.P(B2:B11)

	A	B	C	D	E	F	G
1	Match	Team A	Team B				
2	1	150	260				
3	2	236	110				
4	3	280	140				
5	4	175	265				
6	5	210	289				
7	6	190	300				
8	7	185	95				
9	8	146	70				
10	9	250	145				
11	10	235	290				
12							
13	Variance for population	1756.21	7636.64				
14	Variance for sample	1951.34444	8485.155556				
15							
16							

Step1. Enter the scores in an Excel sheet.
Step 2. Use the **VAR.P** function to calculate the variance of the population.
Type the '=' (equal to) sign in the cell where the result is desired, say 'B13' and enter the function name **VAR.P** and then select the range of numbers, that is, **B2:B11**. Press **Enter.**
Step 3. Use the **VAR.S** function to calculate the variance of the sample.
Type the '=' (equal to) sign in the cell where the result is desired, say 'B14' and enter the function name **VAR.S** and then select the range of numbers, that is, **B2:B11**. Press **Enter.**

6.3.1. Merits of Variance

1. It describes the spread or the variation in the data.
2. It is a useful statistic used to draw insights from surveys.
3. It is also used in accounting to determine the gap between the budgeted and actual amount of costs and revenues.

6.3.2. Demerit of Variance

1. It is difficult to calculate.
2. It is not easy to interpret variance as it is expressed in square units and not in the same units as mean.

6.4. COEFFICIENT OF VARIATION

This statistical measure of dispersion measures relative variability. It is defined as the ratio of the standard deviation to the mean and is independent of the unit of measurement for which it is calculated. The coefficient of variation is expressed as:

$$= \frac{\text{Standard deviation}}{\text{Mean}}$$

$$= \frac{\sigma}{\bar{X}}$$

It can be represented in terms of ratio or percentage. For percentage, we multiply the above formula by 100, that is, $\frac{\sigma}{\bar{X}} \times 100$. Lower coefficient of variation suggests lesser variation than mean. Similarly, a higher coefficient of variation signifies greater variation than mean. That is, if Set A has a coefficient of variation of 15 per cent and Set B has a coefficient of variation of 20 per cent, then we can say that Set B has more variation relative to its mean.

The coefficient of variation is a useful statistic to compare two or more data sets having varied standard deviation and mean or different units of measurement. One should prefer the coefficient of variation for comparison instead of standard deviation. However, one problem with using the coefficient of variation is that when the mean is zero or negative, then the coefficient of variation can be misleading.

For example, let us consider two investment portfolios. We are given the mean of the returns earned and the standard deviation for the returns for each investment portfolio.

	Investment A	*Investment B*
Mean	0.48	0.53
Standard deviation	0.110	0.149

Now by looking at the results, one might think that Investment B provides comparatively higher average returns than A and that the risk is quite similar. But let us calculate the coefficient of variation to make a meaningful decision by using the formula of the coefficient of variation

$$= \frac{\sigma}{\bar{X}} \times 100.$$

	Investment A	*Investment B*
Mean	0.48	0.53
Standard deviation	0.110	0.149
Coefficient of variation	22.92%	28.11%

The calculation returns the coefficient of variance of 22.92 per cent in the case of Investment A and 28.11 per cent in the case of Investment B. A higher degree of variability in Investment B makes it riskier compared to the alternative. In this case, Investment A is safer than Investment B.

Calculating the Coefficient of Variation Using Excel

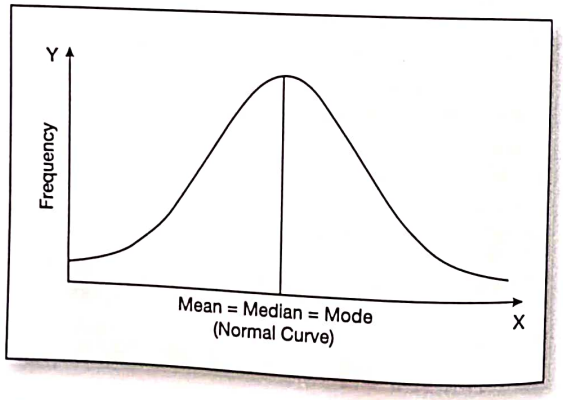
There is no direct formula to calculate the coefficient of variation in Excel. Therefore,

1. We first calculate the simple arithmetic mean of the data set using the **AVERAGE** function.
2. Then we calculate the standard deviation using the **STDEV.P** function.
3. Finally, for calculating the coefficient of variation, we divide the cell containing the standard deviation by the cell containing the mean.

6.5. SKEWNESS

A distribution of the data set may be symmetrical or asymmetrical. Whether distribution is symmetrical or asymmetrical can be determined by the shape of the distribution when plotted.

Symmetrical distribution is the one that looks the same on the left side as on the right side, that is, the left side is identical to the right side. A *normal distribution* is a symmetrical distribution in which the value of the mean = median = mode and the skewness is zero. The frequency curve depicting such symmetrical distribution is known as a normal curve or perfectly bell-shaped curve as it looks similar to a bell. For symmetrical distributions, the mean = median = mode is represented by the middle of the distribution.



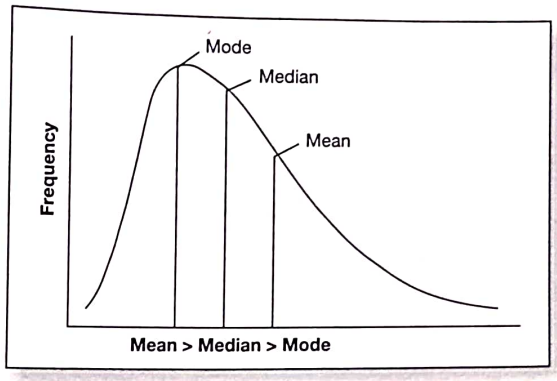
Asymmetrical distribution refers to a distribution whose left side is not identical to its right side. It has either large number of extremely high values and fewer number of smaller values or large number of smaller values and fewer number of high values. Skewness is a descriptive statistic that *measures the lack of symmetry*, that is, *asymmetry*. For an asymmetrical distribution, the mean ≠ median ≠ mode. The greater the differences between mean, median and mode, the higher is the skewness of data. An asymmetrical distribution may be *positively skewed* or

negatively skewed. A positively skewed distribution is skewed to the right and there is a long tail towards the right. Also, mean > median > mode.

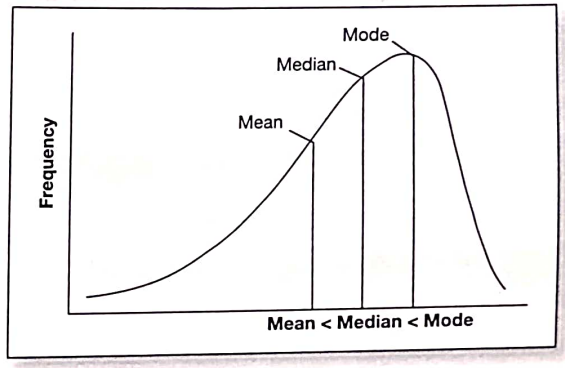
Whereas a negatively skewed distribution is skewed to left. There is a long tail towards the left. Mean < Median < Mode.

The measure to estimate the skewness as given by Karl Pearson is:

$$\text{Coefficient of skewness (Skp)} = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}} = \frac{\bar{X} - M_o}{\sigma}.$$



How to interpret the measure of skewness?



- If the coefficient of skewness =0, then it is a *symmetric distribution or normal distribution.*
- If the coefficient of skewness >0, then it is a *positively skewed distribution.*
- If the coefficient of skewness <0, then it is a *negatively skewed distribution.*
- If 0.5 > | coefficient of skewness | > 0, then the distribution is *comparatively symmetrical.*
- If 1 > | coefficient of skewness | > 0.5, then the distribution is *moderately skewed.*
- If | coefficient of skewness | > 1, then the distribution is *highly skewed.*

As we have noted earlier, the mode of a data set may be ill-defined. A data set may have more than one mode or may have no mode at all. In that case, we can use the median to compute the mode.

The relationship between the mean, median and mode is known as the *empirical relationship*. According to which, for a skewed distribution, the difference between mean and mode is three times the difference between mean and median.

Mean – Mode = 3 (Mean – Median)
or Mode = 3Median – 2Mean

Illustration

Rice production (in metric tonnes) by 30 farmers in India is summarized in the table as follows. Calculate the Karl Pearson coefficient of skewness.

Rice Production (in Metric Tonnes)

Farmer	*Rice Production*	*Farmer*	*Rice Production*
1	29	16	35
2	20	17	15
3	24	18	25
4	30	19	35
5	34	20	33
6	26	21	26
7	19	22	25
8	10	23	40
9	40	24	50
10	40	25	45
11	5	26	15
12	20	27	28
13	24	28	26
14	10	29	35
15	26	30	23

Farmer	*Rice Production*	*(X-Mean)²*	*Farmer*	*Rice Production*	*(X-Mean)²*
1	29	3.61	16	35	62.41
2	20	50.41	17	15	146.41
3	24	9.61	18	25	4.41
4	30	8.41	19	35	62.41
5	34	47.61	20	33	34.81
6	26	1.21	21	26	1.21
7	19	65.61	22	25	4.41
8	10	292.41	23	40	166.41
9	40	166.41	24	50	524.41
10	40	166.41	25	45	320.41
11	5	488.41	26	15	146.41
12	20	50.41	27	28	0.81
13	24	9.61	28	26	1.21
14	10	292.41	29	35	62.41
15	26	1.21	30	23	16.81
Total	**357**	**1,653.75**	**Total**	**456**	**1,554.95**

$$\text{Mean} = \frac{\sum_{i=1}^{n} x_i}{n} = 27.1$$

$$\text{Median} = \frac{\left(\frac{n}{2}\right)\text{th term} + \left(\frac{n}{2}+1\right)\text{th term}}{2} = 26$$

Mode = Highest frequency in the data set, that is, 26.

$$\text{Standard deviation} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} = \sqrt{\frac{1{,}653.75 + 1{,}554.95}{30}} = 10.342$$

$$\text{Karl Pearson coefficient of skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}} - \frac{\bar{X} - M_o}{\sigma}$$

$$= \frac{27.1 - 26}{10.342} = 0.106$$

Answer and interpretation of the result:

0.5 > | coefficient of skewness | > 0, the data is comparatively *symmetrical distribution.*

Second Method (Using Excel)

To calculate the skewness using Excel, follow these steps.

Step 1. Enter the data in an Excel sheet.
Step 2. Use the **SKEW function** to calculate the skewness.
Type the '=' (equal to) sign in the cell where the result is desired, say 'E6' and enter the function name **SKEW** and then select the range of numbers, that is, **B2:B31**.
Step 3. Press **Enter**, the result will be displayed in the cell (Figures 6.9 and 6.10).

The *histogram* is an effective graphical representation of skewness. Let us plot a histogram for the data related to rice production. Looking at the graph, it can be concluded that the distribution is approximately symmetrical (Figure 6.11).

6.6. KURTOSIS

Another measure of descriptive statistics, 'kurtosis', defines the shape of a distribution's tail in relation or in comparison with the normal distribution.[1] It is a measure of 'tailedness' of a distribution. It is denoted by β_2. The value of kurtosis is always compared to that of a normal distribution. It helps to identify whether the tails of the distribution include extreme values or outliers.

[1] Originally, the kurtosis has been defined as a measure of peakedness or flatness. This interpretation was settled to be incorrect and proved wrong by Dr Peter Westfall in his article 'Kurtosis as Peakedness, 1905–2014. R.I.P', *The American Statistician*, Vol. 68, No. 3, 2014. In his article, Westfall argued that kurtosis tells virtually nothing about the shape of the peak—it's only unambiguous interpretation is in terms of tail extremity, that is, either existing outliers (for the sample kurtosis) or propensity to produce outliers (for the kurtosis of a probability distribution). He also laid down numerous examples proving why kurtosis should never be defined in terms of peakedness. The new definition states that kurtosis is the measure of the combined sizes of the tails of the distribution relative to the normal distribution.

FIGURE 6.9

Step 1 in Calculation of Skewness

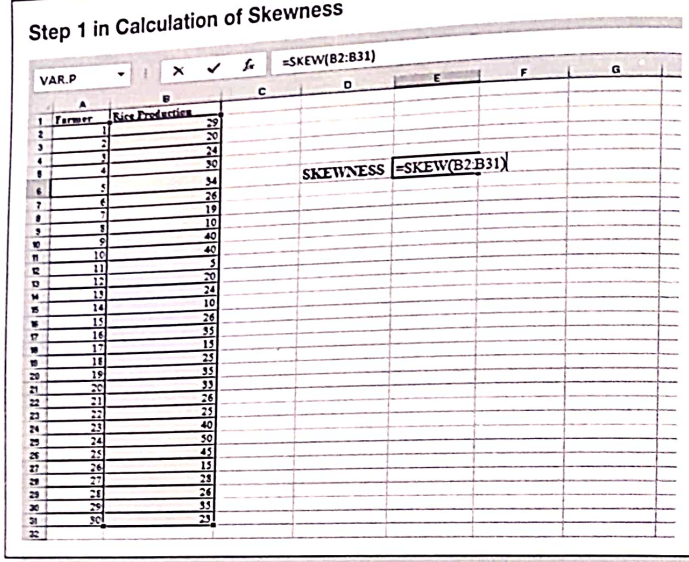
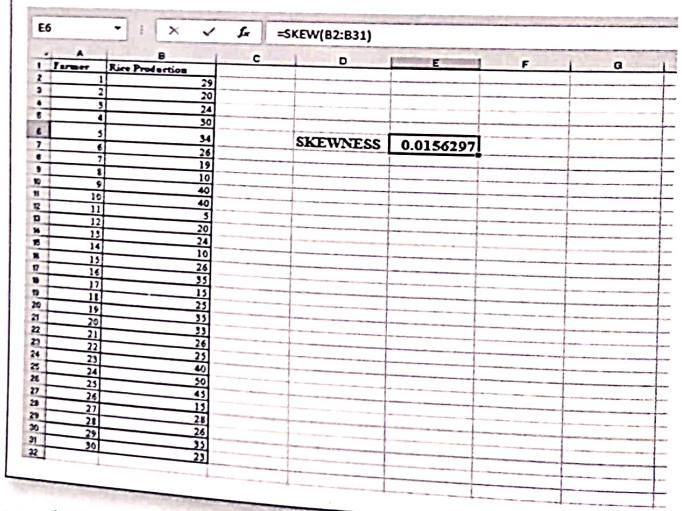


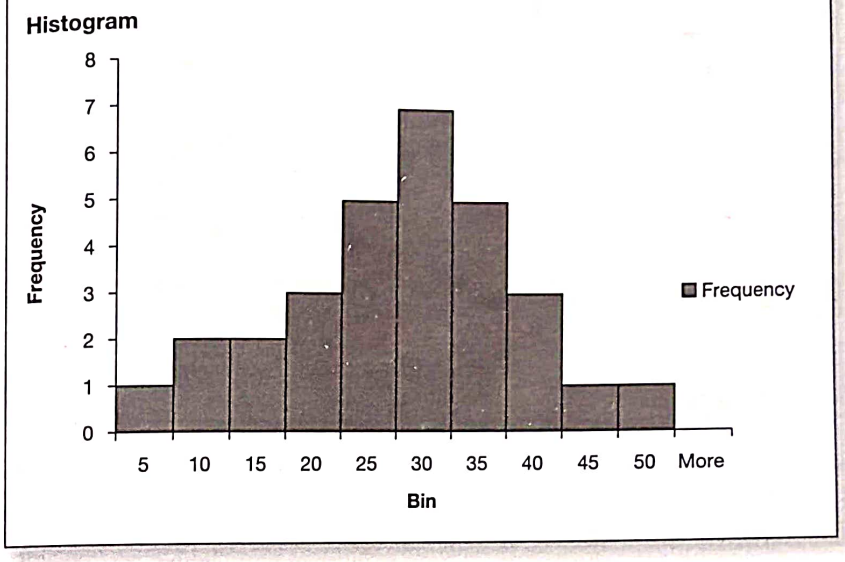
FIGURE 6.10

Step 2 in Calculation of Skewness



Note: Excel uses a different formula to calculate the skewness due to which there can be a slight difference in the results as compared to the results using the Karl Pearson method.

FIGURE 6.11



A heavy-tailed distribution has greater extreme values and a high kurtosis or in other words, heavier the tails, higher is the kurtosis. On the other hand, a light-tailed distribution has fewer extreme values and a low kurtosis or in other words, lighter the tails, lower is the kurtosis. There are the following three categories of Kurtosis:

1. Mesokurtic
2. Leptokurtic
3. Platykurtic

6.6.1. Mesokurtic Distribution

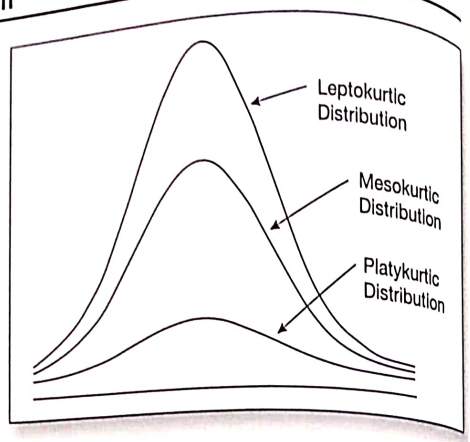
A mesokurtic distribution refers to the distribution having kurtosis similar to that of normal distribution. The tails of such distribution are normal. It has a kurtosis equal to 3. There is equal distribution of items around the mean.

6.6.2. Leptokurtic Distribution

A leptokurtic distribution refers to the distribution having kurtosis greater than a mesokurtic distribution. 'Lepto' means thin or narrow which describes the long and narrow looking peak of the distribution. Such distributions have heavy tails on either side indicating the presence of more extreme values than that of normal distribution. It has a kurtosis of greater than 3. There is too much concentration of items near the mean.

6.6.3. Platykurtic Distribution

A platykurtic distribution refers to the distribution having kurtosis lesser than a mesokurtic distribution. 'Platy' means flat or wide which describes the short and broad looking peak of the distribution. Such distributions have light tails on either side indicating the fewer or less extreme values than that of a normal distribution. It has a kurtosis of less than 3. There is less concentration of items near the mean.



6.6.4. Excess Kurtosis

Kurtosis is sometimes described as excess kurtosis. Excess kurtosis is calculated by subtracting 3 from the kurtosis.

Excess kurtosis = Kurtosis – 3

The following table summarizes excess kurtosis in different distributions.

Distribution	*Kurtosis* β_2	*Excess Kurtosis*	*Tails*	*Extreme Value/Outliers*
Mesokurtic/Normal distribution	3	3 – 3 = 0	Normal	Normal
Leptokurtic distribution	>3	More than 0/positive	Heavy	Greater than the normal distribution
Platykurtic distribution	<3	Less than 0/ negative	Light	Less than the normal distribution

Point to remember: Excel reports 'excess kurtosis'. Hence, kurtosis of a normal distribution computed in Excel will be equal to 0.

Illustration

Fifteen contestants were asked to solve a puzzle. Time taken by each contestant is given as follows. Calculate the kurtosis and comment.

Contestants	*Time Taken (in Minutes)*
1	10
2	40
3	25
4	20
5	44
6	10
7	36
8	26
9	11
10	50
11	27
12	42
13	43
14	32
15	24

To calculate kurtosis using Excel, follow these steps.

Step 1. Enter the data in an Excel sheet.
Step 2. Use the **KURT** function to calculate kurtosis.
Type the '=' (equal to) sign in the cell where the result is desired, say 'E5' and enter the function name **KURT** and then select the range of numbers, that is, **B2:B16** (Figure 6.12).
Step 3. Press **Enter**, the result will be displayed in the cell (Figure 6.13).

FIGURE 6.12

Step 1 in Calculation of Kurtosis

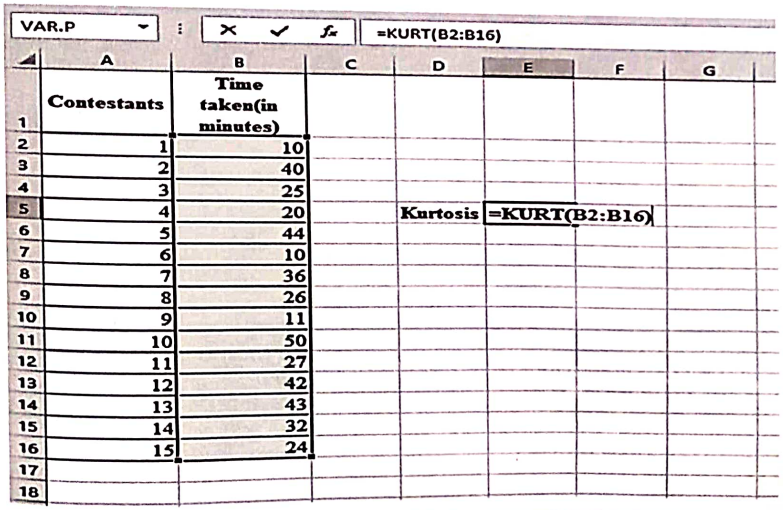


FIGURE 6.13

Step 2 in Calculation of Kurtosis

E5 =KURT(B2:B16)

	A	B	C	D	E	F	G
1	Contestants	Time taken(in minutes)					
2	1	10					
3	2	40					
4	3	25					
5	4	20		Kurtosis	-1.1072		
6	5	44					
7	6	10					
8	7	36					
9	8	26					
10	9	11					
11	10	50					
12	11	27					
13	12	42					
14	13	43					
15	14	32					
16	15	24					
17							
18							

Comment: Since excess kurtosis is less than 0, it is a platykurtic distribution. Such a distribution has light tails on either side indicating the fewer or less extreme values than that of a normal distribution.

6.7. DATA ANALYSIS USING EXCEL

Another method to compute descriptive statistics using Excel is to use the option of data analysis. It is one of the important and widely used tools to perform various financial, statistical, engineering and mathematical functions to analyse the data.

Illustration

The ratings for a movie given by 50 people between 1 and 5 (1 being the lowest rating and 5 being the highest) is given as follows. Calculate the descriptive statistics for the data.

Viewer	*Rating*	*Viewer*	*Rating*
1	5	26	5
2	3	27	1
3	5	28	5
4	4	29	1
5	2	30	3

Viewer	*Rating*	*Viewer*	*Rating*
6	2	31	3
7	2	32	5
8	3	33	4
9	1	34	2
10	3	35	3
11	3	36	1
12	5	37	4
13	3	38	4
14	1	39	1
15	3	40	1
16	3	41	4
17	1	42	1
18	1	43	4
19	4	44	3
20	4	45	3
21	1	46	1
22	3	47	5
23	3	48	2
24	2	49	4
25	2	50	4

To calculate the descriptive statistics using Excel, follow the steps shown in Figures 6.14 and 6.15.

Step 1. Enter the data in an Excel sheet.
Step 2. Go to **Data** tab, click on the option of **Data Analysis** in the analysis group on the right side (if you cannot find the **Data Analysis** button, then follow the steps discussed in Chapter 4 to load *Analysis ToolPak* in Excel).
Step 3. A dialogue box will appear. Select the option of **Descriptive Statistics** from the list and click **OK.**
Step 4. A descriptive statistics dialogue box will appear.

- Fill the following details:
  Input range: The range of numbers for which the statistics are to be calculated, that is, B2:B51.
  Output range: The cell reference where the result is desired. Say, D1.
- Check on **Summary statistics.**
- Click **OK.** The result will be displayed in the cell (Figure 6.16).

To recapitulate, this chapter delved into the common measures of dispersion which quantifies the variability in the data set. The common measures covered were range, standard deviation, variance and coefficient of variation. Besides, it also discussed the concept of skewness that

FIGURE 6.14

Step 1 in Data Analysis (Descriptive Statistics)

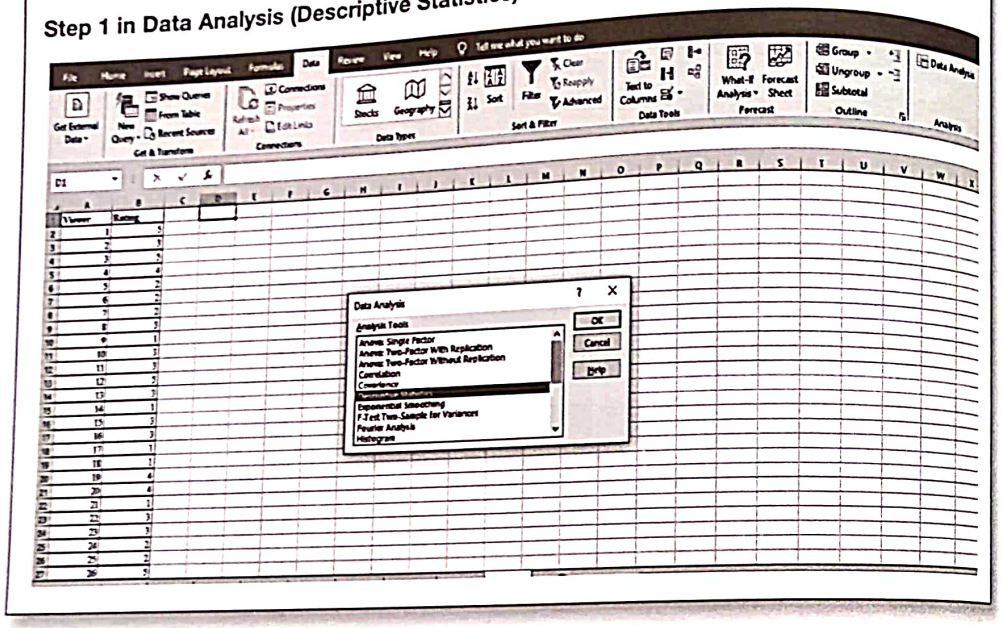


FIGURE 6.15

Step 2 in Data Analysis (Descriptive Statistics)

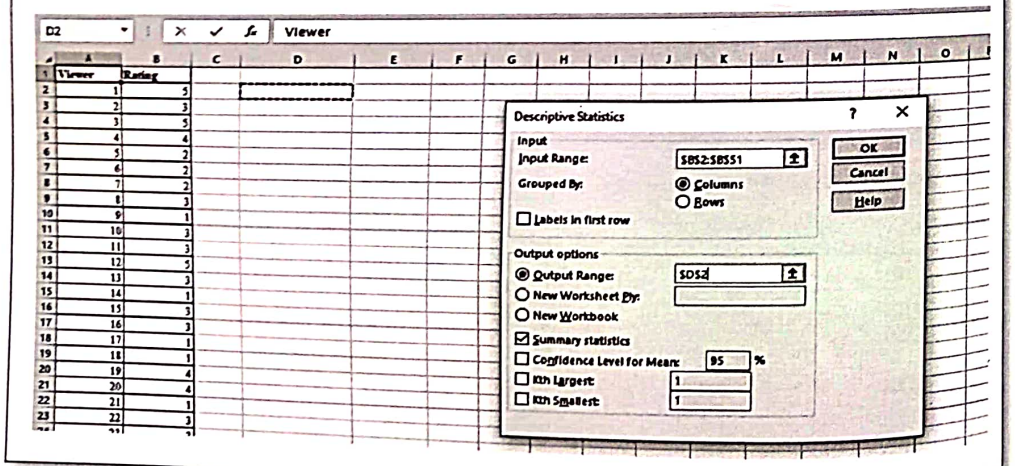
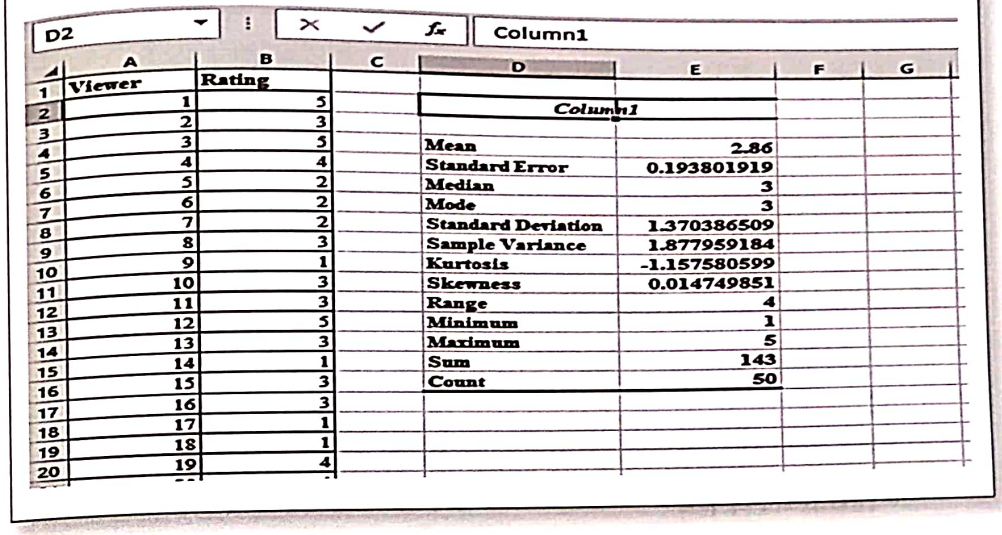


FIGURE 6.16

Step 3 in Data Analysis (Descriptive Statistics)



Viewer	Rating
1	5
2	3
3	5
4	4
5	2
6	2
7	2
8	3
9	1
10	3
11	3
12	5
13	3
14	1
15	3
16	3
17	1
18	1
19	4

Column1	
Mean	2.86
Standard Error	0.193801919
Median	3
Mode	3
Standard Deviation	1.370386509
Sample Variance	1.877959184
Kurtosis	-1.157580599
Skewness	0.014749851
Range	4
Minimum	1
Maximum	5
Sum	143
Count	50

measures asymmetry in the data set determined by the shape of the distribution. Likewise, kurtosis was also explained as a measure of tailedness in a distribution. Computations in general as well as using Excel functions of the measures discussed along with the interpretation of their results remained the focal point. Chapters 5 and 6 combined introduced all the significant measures of descriptive statistics and covered all the basics needed to summarize any data set.[2]

PRACTICE QUESTIONS

1. The following table summarizes the temperature (in Celsius) recorded for a month. Calculate range for the data.

Day	*Temperature*	*Day*	*Temperature*
1	31	16	29
2	28	17	28
3	24	18	30
4	33	19	25

(Continued)

[2] The summary statistics presented in Figure 6.16 has an entry called standard error which will be explained in a later chapter on regression.

(Continued)

Day	Temperature	Day	Temperature
5	28	20	28
6	27	21	28
7	33	22	27
8	29	23	31
9	29	24	30
10	26	25	29
11	32	26	27
12	29	27	33
13	33	28	26
14	32	29	24
15	33	30	27

Ans. 9 degree Celsius.

2. The share prices for the last 10 days of two firms, A and B, are given as follows.

Share Prices (A) (₹)	Share Prices (B) (₹)
3,325	2,657
1,248	529
3,131	2,237
673	2,300
787	1,057
1,345	1,736
1,544	1,863
3,247	1,728
2,583	819

a. Calculate range for Firm A.
b. Calculate range for Firm B.
c. Which firm has comparatively stable prices?
d. Which firm has comparatively volatile prices?

Ans. a. ₹2,652; b. ₹2,915; c. Firm A; d. Firm B.

3. The following table summarizes GDP (in billion US dollars) of 20 countries. Calculate the standard deviation and interpret the result.

Country	GDP (in Billion US Dollars)
1	740
2	9,600

Country	GDP (in Billion US Dollars)
3	95,400
4	15,500
5	8,000
6	3,200
7	8,600
8	9,700
9	12,600
10	7,850
11	9,600
12	3,800
13	1,120
14	9,200
15	12,400
16	7,500
17	4,800
18	13,400
19	8,300
20	2,060

Ans. $19,511.

4. The annual household income of 10 individuals is given as follows.

Individual	Household Income (₹)
A	80,000
B	120,000
C	360,000
D	110,000
E	240,000
F	100,000
G	60,000
H	55,000
I	250,000
J	69,000

Calculate the standard deviation and variance for the data.

Ans. Standard deviation = ₹97,709.98, Variance = 9,547,240,000.

5. Calculate the coefficient of variation for the following data.

Height (cm)	*Weight (kg)*
160	54
172	66
165	59
150	55
182	74
175	69

Ans. We cannot compare the standard deviation for both data sets as they are in different units. Hence, we calculate the coefficient of variation to determine the variability as it is a measure of relative variability.
Coefficient of variation of height = 6.184404545; Coefficient of variation of weight = 13.26249927.
The data of weight has more variation in relation to its mean than that of height.

6. The ratings given by 40 people for a new feature on social media app are given as follows.

Ratings	
4	5
1	2
5	4
1	3
5	4
5	2
4	1
3	2
2	4
3	3
1	5
3	3
1	2
3	3
2	4
5	1
4	3
4	1
1	2
2	5

a. Is the data skewed? Calculate the skewness and comment.
Ans. $S_k = 0.033$, comparatively symmetrical.

7. The number of taxpayers in 30 different cities are given as follows in the table. Calculate the skewness and kurtosis and comment.

City	*No. of Taxpayers*	*City*	*No. of Taxpayers*
1	193,764	16	437,250
2	255,622	17	216,456
3	393,160	18	248,884
4	391,913	19	244,819
5	208,184	20	200,691
6	147,657	21	220,692
7	366,196	22	198,851
8	195,910	23	289,528
9	110,765	24	205,525
10	332,423	25	135,678
11	268,776	26	209,044
12	246,520	27	191,041
13	162,412	28	315,933
14	207,902	29	496,468
15	347,370	30	190,576

Ans. Skewness = 0.9318, asymmetrical distribution, highly skewed to right (positively skewed).
Kurtosis = 0.3719, leptokurtic distribution, heavy tails or larger outliers than a normal distribution.

8. The sales (₹) by a showroom in a month are given as follows. Calculate descriptive statistics using data analysis option.

Sales (₹)		
30,000	68,300	31,800
78,000	28,400	43,900
34,500	89,100	64,400
11,700	12,500	102,000
231,000	45,000	68,900
84,200	22,400	14,900
51,300	33,400	53,500
29,900	68,700	30,200
56,300	32,000	25,900
28,400	81,800	84,000

Ans.

	Column 1
Mean	54,546.66667
Standard error	7,626.892429
Median	44,450
Mode	28,400
Standard deviation	41,774.21027
Sample variance	1,745,084,644
Kurtosis	10.48164561
Skewness	2.707047832
Range	219,300
Minimum	11,700
Maximum	231,000
Sum	1,636,400
Count	30

7 CHAPTER

Compounding: Future Value of Money

LEARNING OBJECTIVES

After reading this chapter, the readers will be able to understand

- The Concept of time value of money
- Simple interest and compound interest
- Calculation and interpretation of the future value of single cash flows using table and Excel
- Calculation and interpretation of the future value of an annuity using table and Excel

One of the most essential and basic concepts in finance is the *time value of money*. Most of the investment decisions in business revolve around the *value of money*, with the underlying assumption that money does not retain its value over time. In other words, the value of money does not remain constant over time as money's purchasing power would fall with the rise in prices.

7.1. TIME VALUE OF MONEY

In any discussion of money, it is said that *a rupee today is worth more than a rupee tomorrow*. The implication is that the value of a rupee *today* will not be the same as the value of the same rupee over

a time period, say *after one year*. Similarly, the value of a rupee after one year will not be the same as it is today.

Let us understand this by taking a simple example. We have won ₹10,000 in a lottery, and we are given two options.

Option A: Take ₹10,000 today
Option B: Take ₹10,000 after 5 years

What would we like to choose? Being a rational individual, we would choose to receive the amount today itself instead of waiting for five years. Why? First, we can invest this money today and receive a higher amount tomorrow because of the interest received on the amount invested. Second, we might be uncertain that we may not get our money back after five years.

The *interest rate* (r) is used for translating money from one time period to another. It represents the *price of time* and the *price of risk*. The price of time is the *opportunity cost* of money, that is, we forego money that could have been used or invested elsewhere. The interest rate, thus, can also be understood as a reward of inflation and is generally positive. The price of risk involves the *risk* or *uncertainty* of not getting back the money in the future. The value of money over time with interest can thus be adjusted using the following two techniques.

- **Compounding:** It computes the worth of today's rupee in the future
- **Discounting:** It refers to today's value of an amount we have in the future

The focus of the present chapter, however, is on *compounding* only. The applications on discounting are discussed subsequently in the next chapter.

7.2. SIMPLE INTEREST

Before learning about *compound interest* (i.e., compounding), which is commonly used in undertaking financial decisions, we need to distinguish it from *simple interest* (SI). While computing SI, interest is paid *only* on the principal amount and calculated using the following formula:

$$SI = Prt$$

where SI = Simple interest,
P = Principal amount,
r = Interest rate (per year) and
t = Time (in years or fraction of a year).

Example: Arun invested ₹1,000 in Bank A for two years. The bank promised him to pay the money at SI rate of 10 per cent per annum. In this example, at the end of the first year, the amount Arun gets from the bank is

$$SI = 1{,}000 \times \frac{10}{100} \times 1 \text{ which is equal to ₹100.}$$

Similarly, at the end of the second year also, he will get the same amount of ₹100 as the simple interest rate is calculated on the basic principal amount ₹1,000 in this case. At the end of two

years, he will get a total of ₹1,000 plus ₹100 (interest amount of the first year) plus ₹100 (interest accrued in the second year) which equals to ₹1,200.

7.3. COMPOUND INTEREST

Interest is *compounded*, if interest is paid on both the *principal amount* and the (*accumulated*) *interest*. For example, if Arun now invests ₹1,000 in Bank A for two years at an interest rate of 10 per cent that is *compounded yearly*, then interest earned by him is as follows:

- At the end of the first year, the interest amount earned will be

$$\left(1{,}000 \times \frac{10}{100}\right) \text{ or ₹100}$$

$$\text{Future value } (FV_1) \text{ after one year} = ₹1{,}000 + \left(₹1{,}000 \times \frac{10}{100}\right) = ₹1{,}100$$

- At the end of the second year, the interest amount earned will be on the principal amount (₹1,000) plus that on the interest earned in the first year (₹100) which equals ₹1,100.

$$\underset{\text{(Principal amount)}}{(1{,}000} + \underset{\text{(Interest earned in first year)}}{100)} \times \underset{\text{(Interest rate)}}{\left(\frac{10}{100}\right)} = \underset{\text{(Interest earned in second year)}}{110}$$

$$\underset{\text{(Principal amount (in second year)}}{FV_2 = 1{,}100} + \underset{\text{(Interest earned in second year)}}{110}$$

$$FV_2 = FV_1 + FV_1\left(\frac{10}{100}\right) = 1{,}210$$

Had he continued to invest in the third year, he would have earned an interest of 10 per cent on ₹1,210, which is ₹121. So the total amount he would receive at the end of the third year is ₹1,210 + ₹121, which equals ₹1,331.

$$FV_3 = FV_2 + FV_2\left(\frac{10}{100}\right) = ₹1{,}331$$

This leads us to a general formula to calculate the compound rate of interest.

$$\text{After 1 year, } FV_1 = P + Pr = P(1 + r)$$

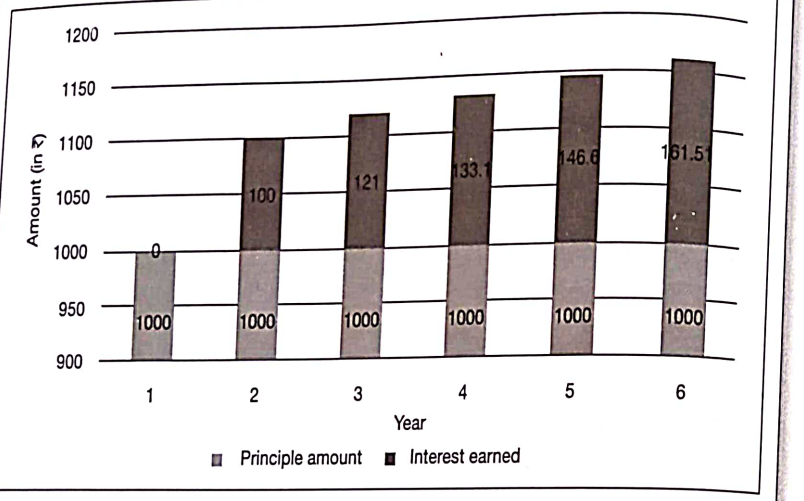
$$\text{After 2 years, } FV_2 = P(1 + r)(1 + r) = P(1 + r)^2$$

$$\text{After 3 years, } FV_3 = P(1 + r)^2(1 + r) = P(1 + r)^3$$

After n years, $FV_n = P(1 + r)^n$, where $(1 + r)$ is equal to the *future value factor*.

FIGURE 7.1

Impact of Compounding in Investment



The compound amount depends on two factors, namely the interest rate (r) and compounding periods (t). Higher the interest rate (r) and the term of the investment (t), the higher will be the income added to the principal amount. Time is the main driving force behind compounding. In finance, even if the interest rate is not changed much but the number of compounding periods is large, the increase in the future returns will be substantial. Therefore, it is generally advised in financial planning to remain invested for a longer period for better returns, compounding being the major reason. Figure 7.1 shows the growth of the amount earned in subsequent years had Arun in our above example continued to invest for more than three years.

The impact of compounding can also be ascertained as follows. If Arun continued with the same investment for 50 years, the compounded amount would have turned out to be nearly ₹117,391. However, reducing the investment by a period of 5 years would have cost him an amount of ₹44,500. This is the power of compounding. Impressed by it, Albert Einstein famously said, *'compound interest is the eighth wonder of the world. He who understands it earns it; he who does not pays it'.*

If the said investment amount is compounded, say semi-annually, then the rate of interest, though would reduce by half, the number of conversion periods would have increased to two, resulting in an even higher amount. Thus, the higher the n, the more is the future value received, making the impact of compounding conspicuous. In this example, the formula becomes

$$FV_n = P\left(1+\frac{r}{2}\right)^{n\times 2}$$

The interest amount earned in the first year, in the ongoing example, would be 102.5 rather than 100 when the compounding was done semi-annually.

Illustration

Ravi deposits an amount of ₹50,000 in his bank for 5 years. The bank offers 6 per cent interest rate compounded annually. How much would he get at the end of the fifth year?

Solution

There are three methods to solve such problems.

First (Direct) Method

In the previous section, we have arrived at the following formula[1]

$$\begin{aligned} FV &= PV(1+r)^n \\ &= 50{,}000(1+0.06)^5 \\ &= 50{,}000(1.06)^5 \\ &= 50{,}000 \times 1.3382 \\ &= ₹66{,}910 \end{aligned}$$

Second Method

Another way of arriving at the future value of *single cash flows* is by using *future value interest factor* (FVIF) tables. The *FVIF* tables can be understood as ratios of the amount received in a year to the amount received in the preceding year.

For instance, depositing ₹50,000 in the bank at the rate of 6 per cent, the future values will be ₹53,000 in the first year, ₹56,180 in the second year, ₹59,550 in the third year, ₹63,125 in the fourth year and ₹66,910 in the fifth year. Dividing the amount received in Year 1, Year 2, Year 3, Year 4 and Year 5 by the principal amount in Year 0, we get *quotients* as 1.06, 1.1236, 1.191, 1.2625 and 1.3382, respectively. These quotients obtained are called *future value interest factor* in the language of compounding. The quotient values for different combinations of n and r are readily available in the FVIF (given in Table 7.1) and therefore need not be calculated manually.[2] We can then easily obtain future values by multiplying these quotients by the principal amount (of Year 0).

For example, the future value at the end of the second year at an interest rate of 6 per cent is read at the intersection point of r at 6 per cent and n of 2 periods. This number gives us the required quotient value. Similarly, the future value at the end of the fifth year at 6 per cent is 1.3382. Thus, multiplying the quotient value of 1.3382 by the principal amount of ₹50,000 straightaway gives the value of ₹66,910. Similarly, the future values of any principal amount can be obtained for any given combination of interest rate (r) and the number of periods (n) by just looking at the *FVIF* table.

[1] P and PV are used interchangeably. PV means present value.

[2] We can reverse this procedure and get PV factor. Dividing the principal amount by the subsequent future values gives the quotient values which are *PVIF*. For example, dividing 50,000 by the future values 53,000, 56,180, 59,550 yield 0.9434, 0,89 and 0.8396, respectively, for the given combination of r (6%) and n = 1, 2 and 3.

TABLE 7.1

Future Value Interest Factor Table

Table A-1 Future Value Interest Factors for One Rupee Compounded at k Percent for n Periods: $FVIF_{k,n} = (1 + k)$

Period	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%	11%	12%	13%	14%	15%	16%	20%
1	1.01	1.02	1.03	1.04	1.05	1.06	1.07	1.08	1.09	1.1	1.11	1.12	1.13	1.14	1.15	1.16	1.2
2	1.0201	1.0404	1.0609	1.0816	1.1025	1.1236	1.1449	1.1664	1.1881	1.21	1.2321	1.2544	1.2769	1.2996	1.3225	1.3456	1.44
3	1.0303	1.0612	1.0927	1.1249	1.1576	1.191	1.225	1.2597	1.295	1.331	1.3676	1.4049	1.4429	1.4815	1.5209	1.5609	1.728
4	1.0406	1.0824	1.1255	1.1699	1.2155	1.2625	1.3108	1.3605	1.4116	1.4641	1.5181	1.5735	1.6305	1.689	1.749	1.8106	2.0736
5	1.051	1.1041	1.1593	1.2167	1.2763	1.3382	1.4026	1.4693	1.5386	1.6105	1.6851	1.7623	1.8424	1.9254	2.0114	2.1003	2.4883
6	1.0615	1.1262	1.1941	1.2653	1.3401	1.4185	1.5007	1.5869	1.6771	1.7716	1.8704	1.9738	2.082	2.195	2.3131	2.4364	2.986
7	1.0721	1.1487	1.2299	1.3159	1.4071	1.5036	1.6058	1.7138	1.828	1.9487	2.0762	2.2107	2.3526	2.5023	2.66	2.8262	3.5832
8	1.0829	1.1717	1.2668	1.3686	1.4775	1.5938	1.7182	1.8509	1.9926	2.1436	2.3045	2.476	2.6584	2.8526	3.059	3.2784	4.2998
9	1.0937	1.1951	1.3048	1.4233	1.5513	1.6895	1.8385	1.999	2.1719	2.3579	2.558	2.7731	3.004	3.2519	3.5179	3.803	5.1598
10	1.1046	1.219	1.3439	1.4802	1.6289	1.7908	1.9672	2.1589	2.3674	2.5937	2.8394	3.1058	3.3946	3.7072	4.0456	4.4114	6.1917
11	1.1157	1.2434	1.3842	1.5395	1.7103	1.8983	2.1048	2.3316	2.5804	2.8531	3.1518	3.4785	3.8359	4.2262	4.6524	5.1173	7.4301
12	1.1268	1.2682	1.4258	1.601	1.7959	2.0122	2.2522	2.5182	2.8127	3.1384	3.4985	3.896	4.3345	4.8179	5.3503	5.936	8.9161
13	1.1381	1.2936	1.4686	1.6651	1.8856	2.1329	2.4098	2.7196	3.0658	3.4523	3.8833	4.3635	4.898	5.4924	6.1528	6.8858	10.699
14	1.1496	1.3196	1.5126	1.7317	1.9799	2.2609	2.5785	2.9372	3.3417	3.7975	4.3104	4.8871	5.5348	6.2613	7.0757	7.9875	12.839
15	1.161	1.3459	1.558	1.8009	2.0789	2.3966	2.759	3.1722	3.6425	4.1772	4.7846	5.4736	6.2543	7.1379	8.1371	9.2655	15.407
16	1.1726	1.3728	1.6047	1.873	2.1829	2.5404	2.9522	3.4259	3.9703	4.595	5.3109	6.1304	7.0673	8.1372	9.3576	10.748	18.488
17	1.1843	1.4002	1.6528	1.9479	2.292	2.6928	3.1588	3.7	4.3276	5.0545	5.8951	6.866	7.9861	9.2765	10.761	12.468	22.186
18	1.1961	1.4282	1.7024	2.0258	2.4066	2.8543	3.3799	3.996	4.7171	5.5599	6.5436	7.69	9.0243	10.575	12.375	14.463	26.623

Third Method (Using Excel)

Excel calculates compounding easily. To arrive at the result, we take the following steps. We proceed by clicking on the **Insert** function f_x option adjacent to the **Formula Bar**, select category **Financial** and choose **FV** as the function. Next, we can write the arguments as follows:

- **Rate**, which we take as 6 per cent. The rate is written either as 0.60 or with a percentage sign, that is, 6%.
- The **Nper** argument is the number of periods. It is 5 years in our example.
- **Pmt** refers to the payment made periodically but since we are taking one-time investment in our example, we leave it blank.
- **Pv** is the principal amount (present value) invested. It must be kept in mind that a (–) sign is written along with the principal amount as it is an outgoing payment.
- Leave **Type** as blank or write 0 implying payment is made at the end of the year/period.
- Alternatively, the same calculation is done by writing = FV (**Rate, Nper, Pmt, Pv**) in the **Formula Bar**. These arguments appear automatically 'as soon as we write '=FV.' We need to fill the values of these arguments accordingly.
- We get the same result of ₹66,911 (Figure 7.2). A marginal difference that is arising is on account of rounding off errors.

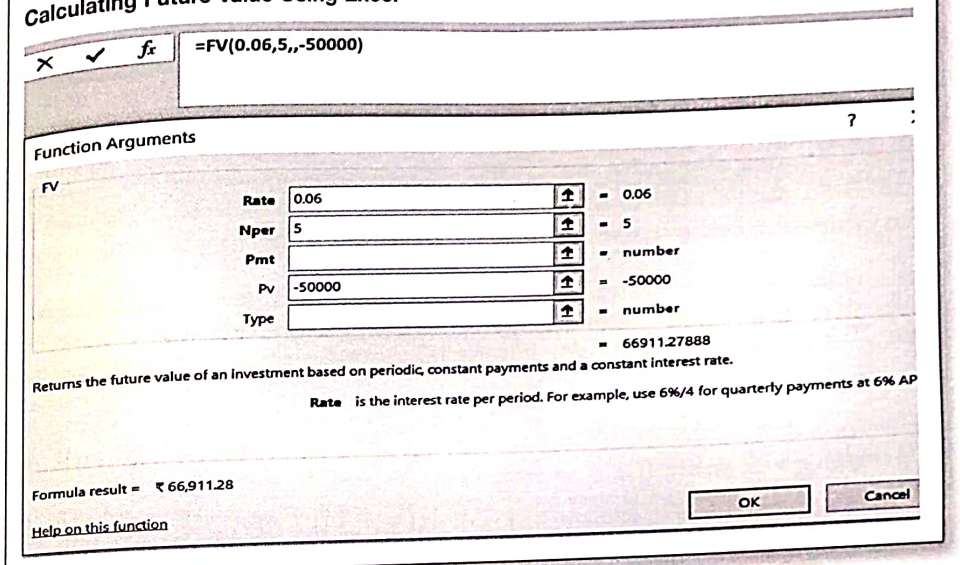
Check Your Understanding

At an interest rate of 10 per cent, it is better to have ₹100 today than ₹120 in 2 years.

Ans. True.

FIGURE 7.2

Calculating Future Value Using Excel



7.4. FUTURE VALUE OF AN ANNUITY

An *annuity* is a series of payments made at equal intervals. Examples are regular deposits to a savings account, monthly home mortgage payments, monthly insurance payments and pension payments. In practice, we seldom make any decision (say investment) based on one-time payment but multiple payments. Hence, it is quite imperative to understand annuity calculations.

The future value of an ordinary annuity is simply a sum of the future values of the individual cash flows. For example, Arun now decides to invest ₹10,000 each in Year 0, Year 1 and Year 2 at the rate of 10 per cent. How much will he get after three years? A timeline is important to understand such calculations that are presented in Table 7.2.

If we look carefully at Table 7.2, then we can notice that the cash flow of Year 0 is 0 as nothing is happening in Year 0, which is today. The Years 1, 2 and 3 represent the end of Year 1, 2 and 3, respectively. *Years to end* refer to how many years are left. For example, in Year 0, years to end are 3 as Year 0 to 1 is counted as 1 year. Similarly, Year 1–2 as the second year and Year 2–3 as the third year. Remember Year 3 is the end of the period in this example.

To determine future values, let us consider one cash flow at a time. At Point 0, the future value must be 0 because conventionally we do not get cash flow in Year 0. The first annuity occurs at the end of Year 1 and the number of years to end is 2, therefore, $C(1 + r)^2$ is the formula as derived before. Similarly, the future values of C in Years 2 and 3 take the formula $C(1 + r)^1$ and C. Thus, having three C's at different points of time is different from having them at the same point of time

TABLE 7.2

Timeline of a Cash Flow

Year	*Cash Flow*	*Years to End*	*Future Value*
0	0	3	0
1	C	2	$C(1+r)^2$
2	C	1	$C(1+r)^1$
3	C	0	C

because of the time value of money. The arguments discussed above can now be used to generalize the formula. The summation of future value for 3-period payments is expressed as $C(1+r)^2 + C(1+r)^1 + C$. Note that FV with power 3 has not been taken into account due to the absence of cash flow in Period 0.

$$FV_3 = C[(1+r)^2 + (1+r)^1 + 1]$$

Multiplying both sides by $(1+r)$, we get

$$FV_3\,(1+r) = C[(1+r)^3 + (1+r)^2 + (1+r)]$$

Subtracting $FV_3\,(1+r)$ from FV_3

$$= C[(1+r)^3 + (1+r)^2 + (1+r)] - C[(1+r)^2 + (1+r)^1 + 1]$$

$$= C[(1+r)^3 - 1)]$$

Solving for $FV_3 = C\left[\dfrac{(1+r)^3 - 1}{r}\right]$

For period n,

$$FV_n = C\left[\frac{(1+r)^n - 1}{r}\right]$$

Illustration

Kumar deposited ₹60,000 per year continuously for 5 years in his bank. The bank pays 8 per cent per annum interest compounded yearly. Find out the total amount available to Kumar at the end of 5 years.

Solution

There can be three methods to solve such problems.

First (Direct) Method

$$FVA = C \times \left[\frac{(1+r)^n - 1}{r}\right]$$

$$= 60{,}000 \times \left[\frac{(1+0.08)^5 - 1}{0.08}\right]$$

$$= ₹351{,}996$$

Second Method (Using Future Value Annuity Table)

Similar to the formation of the FVIF table, future value annuity table is drawn. It is an alternative way of calculating quotients manually. The quotient values are readily available in this table, as shown in Table 7.3. For example, in the ongoing illustration, we can see that the corresponding quotient value of r (at 8%) and n of 5 periods is 5.8666. Multiplying the PV amount of ₹60,000 by 5.8666 delivers an amount of ₹351,996. A simple and straightforward method if no financial calculator is available.

Third Method (Using Excel)

The steps in Excel are almost the same as followed earlier. We proceed by clicking on the **Insert** function f_x option adjacent to the **Formula Bar** and choosing **FV** as the function. Next, we write the arguments as follows.

- **Rate**, which is taken as 8 per cent. The rate is written either as 0.08 or with a percentage sign, that is, 8%.
- The **Nper** argument is the number of periods. It is 5 years in our example.
- **Pmt** refers to the payment made periodically. In our example, we are paying an amount of ₹60,000 per year, hence, we place this figure in **Pmt**. As noted previously, a (–) sign is written along with the principal amount as it is an *outgoing* payment.

TABLE 7.3

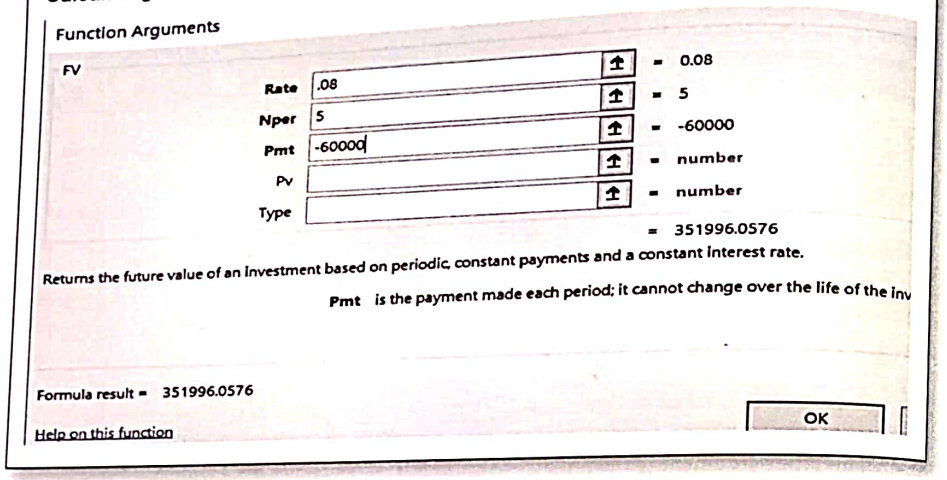
Future Annuity Value Table

Table A-2 Future Value Interest Factors for a One-Rupee Annuity Compounded at *X* Percent for *n* Periods: FVIFA $_{x,n} = [(1+k)^n - 1]/k$

Period	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%	11%	12%	13%	14%	15%	16%	20%	24%	25%	50%
1	1	1.02	1.03	1.04	1.05	1.06	1.07	1.08	1.09	1.1	1.11	1.12	1.13	1.14	1.15	1.16	1.2	1.24	1.25	1.3
2	2.01	2.02	2.03	2.04	2.05	2.06	2.07	2.08	2.09	2.1	2.11	2.12	2.13	2.14	2.15	2.16	2.2	2.24	2.25	2.3
3	3.0301	3.0604	3.0909	3.1216	3.1525	3.1836	3.2149	3.2464	3.2781	3.31	3.3421	3.3744	3.4069	3.4396	3.4725	3.5056	3.64	3.7776	3.8125	3.99
4	4.0604	4.1216	4.1836	4.2465	4.3101	4.3746	4.4399	4.5061	4.5731	4.641	4.7097	4.7793	4.8498	4.9211	4.9934	5.0665	5.368	5.6842	5.7656	6.187
5	5.101	5.204	5.3091	5.4163	5.5256	5.6371	5.7507	5.8666	5.9847	6.1051	6.2278	6.3528	6.4803	6.6101	6.7424	6.8771	7.4416	8.0484	8.207	8.8431
6	6.152	6.3081	6.4684	6.633	6.8019	6.9753	7.1533	7.3359	7.5233	7.7156	7.9129	8.1152	8.3227	8.5355	8.7537	8.9775	9.9299	10.98	11.259	12.756
7	7.2135	7.4343	7.6625	7.8983	8.142	8.3938	8.654	8.9228	9.2004	9.4872	9.7833	10.089	10.405	10.73	11.067	11.414	12.916	14.615	15.073	17.583
8	8.2857	8.583	8.8923	9.2142	9.5491	9.8975	10.26	10.637	11.028	11.436	11.859	12.3	12.757	13.233	13.727	14.24	16.499	19.123	19.842	23.858
9	9.3685	9.7546	10.159	10.583	11.027	11.491	11.978	12.488	13.021	13.579	14.164	14.776	15.416	16.085	16.786	17.519	20.799	24.712	25.802	32.015
10	10.462	10.95	11.464	12.006	12.578	13.181	13.816	14.487	15.193	15.937	16.722	17.549	18.42	19.337	20.304	21.321	25.958	31.643	33.253	42.619
11	11.567	12.169	12.808	13.486	14.207	14.972	15.784	16.645	17.56	18.531	19.561	20.655	21.814	23.045	24.349	25.733	32.15	40.238	42.566	56.405
12	12.683	13.412	14.192	15.026	15.917	16.87	17.888	18.977	20.141	21.384	22.713	24.133	25.65	27.271	29.002	30.85	39.581	50.895	54.208	74.327
13	13.809	14.68	15.618	16.627	17.713	18.882	20.141	21.495	22.953	24.523	26.212	28.029	29.985	32.089	34.352	36.786	48.497	64.11	68.76	97.625
14	14.947	15.974	17.086	18.292	19.599	21.015	22.55	24.215	26.019	27.975	30.095	32.393	34.883	37.581	40.505	43.672	59.196	80.496	86.949	127.913
15	16.097	17.293	18.599	20.024	21.579	23.276	25.129	27.152	29.361	31.772	34.405	37.28	40.417	43.842	47.58	51.66	72.035	100.815	108.887	197.288
16	17.258	18.639	20.157	21.825	23.657	25.673	27.888	30.324	33.003	35.95	38.19	42.753	46.672	50.98	55.717	60.925	87.442	126.011	136.109	218.472
17	18.43	20.012	21.762	23.698	25.84	28.213	30.84	33.75	36.974	40.545	44.501	48.884	53.739	59.118	65.075	71.673	105.931	157.253	173.636	285.014
18	19.615	21.412	23.414	25.645	28.132	30.906	33.999	37.45	41.301	45.599	50.396	56.75	61.725	68.394	75.836	84.141	128.117	195.994	218.045	371.518
19	20.811	22.841	25.117	27.671	30.539	33.76	37.379	41.446	46.018	51.159	56.939	63.44	70.749	78.969	88.212	98.603	154.74	244.033	273.556	463.973
20	22.019	24.297	26.87	29.778	33.066	36.786	40.995	45.762	51.16	57.275	64.203	72.052	80.947	91.025	102.444	115.38	186.688	303.601	342.945	630.165

FIGURE 7.3

Calculating Future Value of Annuity Using Excel



- **Pv** is the present value of an investment. We leave it blank as this is not a single cash flow problem.
- Leave **Type** as blank or write 0 implying payment is made at the end of the year/period.
- Alternatively, the same calculation is done by writing '=FV (**Rate, Nper, Pmt, Pv**)' in the **Formula Bar**. These arguments appear automatically as soon as we write '=FV.' We need to fill the values of these arguments accordingly in the same order as required.
- We can notice the calculated value as ₹351,996 in Figure 7.3.

To sum up, this chapter enables us to calculate the future values of both single and multiple cash flows. This calculation helps in building asset creation as well as in financial planning. Various methodologies to calculate future values have been explained systematically to understand the concept that lies behind these calculations. These concepts are of utmost importance in real life. The foundation of concepts that have been laid through in this chapter forms a base for subsequent topics.

PRACTICE QUESTIONS

1. Rajesh has deposited ₹100 in an account that pays 3 per cent simple interest, yearly. What will be the balance amount in his account at the end of three years principal?
Ans. 109.

2. Ria decided to invest ₹15,000 in her savings account that pays 3 per cent interest, compounded quarterly. How much she will have in her account at the end of five years if she does not make any withdrawals?
Ans. ₹17,417.76.

3. Rohan has ₹1,000 that he will invest in a safe financial instrument expected to return 3 per cent annually. Ananya has ₹500 which she invests in a stock market that is expected to return 7 per cent annually. Who will have more money after 20 years? And how much does he/she have in FV terms?
Ans. Ananya: ₹1,935.

4. Daksh has ₹100,000 to decide on two investment options. Option 1 returns 60 per cent annually for 4 years, but he can invest ₹10,000 at the most. The return on Option 2 is 12 per cent annually for 4 years but it would require the entire investment of ₹100,000. Which option should Daksh choose and what is the benefit over the alternate option? Assume that ₹90,000 not invested in Option 1 would be deposited safely but with no interest.
Ans. Option 2: ₹1,816.

5. Rama is 35 years old and wants to plan for her retirement. She decides to save ₹4,800 starting at the end of this year until she is 65. She can make 7 per cent of returns on her account. How much money will she have at the age of 65 for her retirement?
Ans. ₹453,412.

6. Vanya's parents have been putting money in an account for her since she was five years old. She now has turned 21 and has access to this money. Her mother asks her to guess what her account is worth today after she gives her the hint that they have invested ₹1,000 every year since her fifth birthday and have again made an investment. The interest rate is 3.5 per cent annually. What is the worth of Vanya's account today?
Ans. ₹22,705.

8

CHAPTER

Investment Decision Criteria: NPV and IRR

LEARNING OBJECTIVES

After reading this chapter, the readers will be able to understand

- Meaning and decision rule of net present value (NPV) method
- Computation of NPV using Excel
- Meaning and decision Rule of internal rate of return (IRR) method
- Computation of IRR using Excel
- Application of IRR in endowment plans, moneyback policies and pension plans
- IRR versus extended internal rate of return (XIRR)

In this chapter, the techniques which are used to decide whether the business project and investment are worth pursuing or not are introduced. An investor generally encounters a problem or challenge in deciding the most economically viable project or a venture. He/she is confronted with several alternative options from which he/she will have to choose a particular investment opportunity. The investment may involve a considerable amount of money. Consequently, such an investment involves critical scrutiny, and thus, a careful decision is highly essential.

According to the theory of decision-making, any right decision for a good investment must have the following properties:

- The benefits of the project are more than the cost incurred.
- The criteria should be having units of measurement.

- Simple to calculate.
- Easy to compare different ideas/projects.

Based on the above criteria, two distinct techniques generally used in deciding on business investment are:

- *Net present value (NPV)*
- *Internal rate of return (IRR)*

These two methods act as a decision criterion to determine the idea of undertaking a good project. However, there are several advantages and disadvantages associated with each of these methods. The merits and demerits attributed to each of these methods will be discussed.

8.1. NET PRESENT VALUE (NPV)

NPV refers to the *difference* between future cash-flow streams converted into present values at an appropriate discount rate of the initial investment. While considering the difference, the present values at an appropriate discount rate are regarded as a positive entity, while the initial investment is considered a negative entity. Therefore, in calculating NPV, we consider all cash flows that are both positive and negative entities in nature.

NPV comprises two crucial components:

- The *cash flows* generated in the future. These are mainly the returns that are attributed to the project in the future. Therefore, the values of these returns should be forecasted based on realistic assumptions.
- The *discount rate* captures the opportunity cost of an idea or the investment opportunity that an investor comes across. Alternatively, it can be understood as the rate equivalent to the expected returns prevailing in similar businesses. Thus, one should be quite careful in applying a discount rate in practice.

8.1.1. Decision Rule: Accept/Reject Criteria

The accept/reject criteria of the NPV method works by comparing the present value of the future cash streams with the initial investment.

- If NPV is positive, then the project is worth accepting.
- If the project's NPV is negative, the project is not investment worthy. As a result of a higher discount rate, if NPV turns out to be negative, it is either on account of a higher cost of capital or higher expected returns.
- When NPV is equal to 0, the decision of accepting or rejecting the project becomes indifferent as there is neither a gain nor a loss.

In the case of mutually exclusive projects, when the selection of one project leads to the rejection of other projects, select the project with the highest NPV.

8.1.2. Calculating NPV

NPV is simply a difference between the present value of all future cash-flow streams and initial investments:

NPV = Present value of all future cash-flow streams – Initial investment

$$\text{NPV} = \left(\frac{C_1}{[1+r]} + \frac{C_2}{[1+r]^2} + \frac{C_3}{[1+r]^3} + \ldots\ldots\ldots \frac{C_n}{[1+r]^n}\right) - \text{I}_0 \quad (1)$$

Here,

$C_1, C_2 \ldots C_n$ = Cash flows generated at different points of time in the future
r = Discount rate
n = Estimated life of the project
I_0 = Initial investment

Equation 1 can be summarized as the following formula:

$$\text{NPV} = \sum_{t=1}^{n} \frac{C_n}{(1+r)^n} - \text{I}_0 \quad (2)$$

$$\text{NPV} = \sum_{t=1}^{n} C_n \times \frac{1}{(1+r)^n} - \text{I}_0,$$

$$\text{NPV} = \sum_{t=1}^{n} C_n \times PVF_{r,n} - \text{I}_0 \quad (3)$$

Where $\text{PVF}_{r,n}$ = Present value factor for the given discount rate and the given year.

The illustrations below will depict the detailed steps and calculations involved in computing NPV and the decision-making process for a particular project or investment.

Illustration

Tarun is considering buying a second-hand car to use it as a taxi. The cost of this car is ₹50,000. The car has a life expectancy of five years and no salvage value. The estimated cash inflows from year one to five are as follows: ₹10,000, ₹10,450, ₹11,800, ₹12,250 and ₹16,750. The appropriate discount rate is 10 per cent. State its NPV and if this investment is worth it.

Solution

There are three methods to solve such problems.

First Method

NPV = Present value of cash inflows – Present value of cash outflows

$$\text{NPV} = \sum_{t=1}^{n} \frac{C_n}{(1+r)^n} - \text{I}_0$$

$$\text{NPV} = C_1 \times \frac{1}{(1+0.10)^1} + C_2 \times \frac{1}{(1+0.10)^2} + C_3 \times \frac{1}{(1+0.10)^3} + C_4 \times \frac{1}{(1+0.10)^4} + C_5 \times \frac{1}{(1+0.10)^5} - I_0$$

$$\text{NPV} = 10{,}000 \times 0.909 + 10{,}450 \times 0.826 + 11{,}800 \times 0.751 + 12{,}250 \times 0.683 + 16{,}750 \times 0.621 - 50{,}000$$

NPV = 9,090 + 8,632 + 8,862 + 8,367 + 10,402 – 50,000

NPV = 45,353 – 50,000

= –₹4,647

Since the NPV is negative, the investment is not a good one.

Second Method (Using PVF Table)

Year	*Cash Flows*	*PVF at 10% ($PVF_{10\%,n}$)*	*PV of Cash Flows*
1	10,000	0.909	9,090
2	10,450	0.826	8,632
3	11,800	0.751	8,862
4	12,250	0.683	8,367
5	16,750	0.621	10,402
		Total PV	45,353
		Less: Initial cash outlay	50,000
		NPV	–4,647

Note: PVFs can be calculated using the formula $\frac{1}{(1+r)^t}$ and are readily obtainable from the PVF table.

$$\text{PVF}_{10\%,1} = \frac{1}{(1+0.10)^1} = 0.909 \quad \text{PVF}_{10\%,2} = \frac{1}{(1+0.10)^2} = 0.826 \quad \text{PVF}_{10\%,3} = \frac{1}{(1+0.10)^3} = 0.751$$

$$\text{PVF}_{10\%,4} = \frac{1}{(1+0.10)^4} = 0.683 \quad \text{PVF}_{10\%,5} = \frac{1}{(1+0.10)^5} = 0.621$$

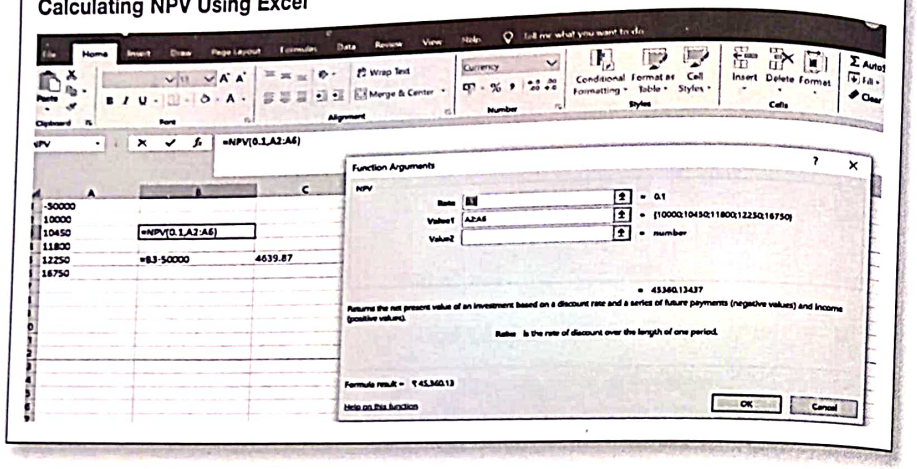
Third Method (Using Excel)

NPV can be calculated using Excel as well.

Step 1. Enter the data of all the cash inflows and outflows.
Step 2. Use the formula NPV and fill the arguments required, which are discount rate and values (of cash inflows).
Step 3. Subtract the result, we get in Step 2 from the initial investment.

FIGURE 8.1

Calculating NPV Using Excel



The calculations are shown in Figure 8.1. The difference between the results we get in methods 2 and 3 is due to rounding off.

When manually computing NPV in the case of annuity streams, the methods, as mentioned above, can be applied in a similar way but with the help of the PVAF table.

Check Your Understanding

Consider an investment that has the following expected cash flows. The discount rate is 10 per cent. What is the NPV of this investment?

Year	*Cash Flows* *A*
Today	–₹10,000
1	₹1,000
2	₹1,000
3	₹9,000

Ans: –1,503

This method is associated with several merits and demerits, as discussed below.

8.2. MERITS AND DEMERITS OF NPV

8.2.1. Merits of NPV

1. It takes into account the time value of money.
2. It takes into consideration all the cash flows arising out of the project in its lifetime.

3. It is an absolute measure of profitability.
4. This method satisfies the value-additivity principle. It gives output in terms of the absolute amount so the NPVs of the project can be added, which is not possible with other methods.
5. It meets the firm's investment goals of wealth maximization.

8.2.2. Demerits of NPV

1. This method ignores uncertainties in cash flows that exist in the real world.
2. The uncertainties are also involved in the calculation of the discount rate.

Another method that is considered in deciding whether a particular investment option is economically viable or not is *IRR*.

8.3. INTERNAL RATE OF RETURN (IRR)

IRR is that *discount rate* which equates the PV of the cash inflows with the PV of the cash outflows or at which NPV is zero. In other words, IRR is a modified compound annual growth rate and is used to calculate return over a holding period when the investments are multiple or in a varied frequency. IRR is also called the marginal efficiency of capital, the marginal productivity of capital, the yield on investment, time adjusted rate of return, the marginal rate of return and so on.

The following condition must be true to arrive at IRR:

PV of cash inflows = Present value of cash outflows

PV of cash inflows – PV of cash outflows = 0

NPV= 0

Symbolically,

$$\sum_{t=1}^{n} \frac{C_n}{(1+r)^n} = \sum_{t=1}^{n} \frac{CO_t}{(1+r)^n}$$

$$\sum_{t=1}^{n} \frac{C_n}{(1+r)^n} - \sum_{t=1}^{n} \frac{CO_t}{(1+r)^n} = 0$$

$$NPV = 0$$

Suppose we put 10 per cent in place of r and both sides become equal (or NPV = 0), then this 10 per cent is the IRR.

Note: If cash outflows are in the zero years only (multiple cash outflows are not there), then there is no need to calculate the PV of the cash outflows.

For calculating IRR directly, we use the trial and error approach. We can either make use of the PVF table in case of mixed streams or PVAF table in case of annuity streams.

- We take a discount rate with given n and compute the PV of the cash inflows of the project. If the PV of cash inflows is equal to the PV of cash outflows (or cash outflows in the zero years), then the percentage taken is the IRR.

- If the NPV is negative, then reduce the discount rate to get a positive NPV and
- If the NPV is positive, then increase the discount rate to get a negative NPV.

Now using the two discount rate—one where NPV is positive and the other where NPV is negative, to arrive at IRR, we apply the following formula:[1]

$$\text{IRR} = \text{LDR} + \frac{\text{NPV}_{\text{LDR}}}{\text{NPV}_{\text{LDR}} - \text{NPV}_{\text{HDR}}} \times |\text{Difference of rates}|$$

or

$$\text{IRR} = \text{LDR} + \frac{\text{PV}_{\text{LDR}} - \text{PV of cash outflows or initial outlay}}{\text{PV}_{\text{LDR}} - \text{PV}_{\text{HDR}}} \times |\text{Difference of rates}|$$

(The above formulae are for the linear interpolation)
Here, LDR = Lower discount rate and HDR = Higher discount rate
The difference in rates is taken after ignoring the signs.

Note

1. We sometimes are not able to decide the discount rate with which to start working. For that, we take any discount rate and compute the PV of cash inflows or NPV. Now, the PV of cash inflows or NPV will give you the direction regarding the next discount rate.
2. The difference between the two discount rates is usually taken as 5.

The IRR decision rule states that:

- If the asset's IRR is less than the standard rate of return, then do not select the asset.
- If the asset's IRR is greater than the standard rate of return, then select the asset.

To comprehend the detailed working of this technique, the following illustrations would be useful both for computing IRR manually and by deploying Excel.

Illustration

The cost of a machine is ₹42,000. Cash inflows are as follows:

Year	*Amount (₹)*
1	12,000
2	12,000
3	14,000
4	15,000
5	16,000

Calculate IRR.

[1] The derivation of this formula is beyond the scope of this book.

Solution
There are two methods to solve such problems.

First Method
In the first method, the trial and error method is used. Let us calculate the NPV at the discount rate of 10 per cent.

Year	*Cash Inflows*	*PVFs at 10% ($PVF_{10\%,n}$)*	*PV at 10%*
1	12,000	0.909	10,908
2	12,000	0.826	9,912
3	14,000	0.751	10,514
4	15,000	0.683	10,245
5	16,000	0.621	9,936
		Total PV	51,515
		Less: Initial cash outlay	–42,000
		NPV	9,515

Since NPV is positive, so we shall increase the discount rate. Now, let us calculate the NPV at the discount rate of 15 per cent.

Year	*Cash Inflows*	*PVFs at 15% ($PVF_{15\%,n}$)*	*PV at 15%*
1	12,000	0.870	10,440
2	12,000	0.756	9,072
3	14,000	0.658	9,212
4	15,000	0.572	8,580
5	16,000	0.497	7,952
		Total PV	45,256
		Less: Initial cash outlay	–42,000
		NPV	3,256

Since NPV is positive, so we shall increase the discount rate. Now, let us calculate the NPV at the discount rate of 20 per cent.

Year	*Cash Inflows*	*PVFs at 20% ($PVF_{20\%,n}$)*	*PV at 20%*
1	12,000	0.833	9,996
2	12,000	0.694	8,328
3	14,000	0.579	8,106
4	15,000	0.482	7,230
5	16,000	0.402	6,432
		Total PV	40,092
		Less: Initial cash outlay	–42,000
		NPV	(1,908)

Now, the NPV is negative at 20 per cent and positive at 15 per cent. It means that the IRR lies between 20 and 15 per cent. So there is a need to apply the formula for the interpolation:

$$\text{IRR} = \text{LDR} + \frac{\text{NPV}_{\text{LDR}}}{\text{NPV}_{\text{LDR}} - \text{NPV}_{\text{HDR}}} \times |\text{Difference of rates}|$$

$$\text{IRR} = 15 + \frac{\text{NPV}_{15\%}}{\text{NPV}_{15\%} - \text{NPV}_{20\%}} \times 5$$

$$\text{IRR} = 15 + \frac{1{,}908}{3{,}256 - (-1{,}908)} \times 5$$

$$\text{IRR} = 15 + 0.6035 \times 5$$

$$\text{IRR} = 15 + 3.0175$$

$$\text{IRR} \cong 18.01$$

or

$$\text{IRR} = \text{LDR} + \frac{\text{PV}_{\text{LDR}} - \text{PV of cash outflows or initial outlay}}{\text{PV}_{\text{LDR}} - \text{PV}_{\text{HDR}}} \times |\text{Difference of rates}|$$

$$\text{IRR} = 15 + \frac{\text{PV}_{15\%} - \text{PV of cash outflows or initial outlay}}{\text{PV}_{15\%} - \text{PV}_{20\%}} \times 5$$

$$\text{IRR} = 15 + \frac{45{,}256 - 42{,}000}{45{,}256 - 40{,}092} \times 5$$

$$\text{IRR} = 15 + 0.6035 \times 5$$

$$\text{IRR} = 15 + 3.0175$$

$$\text{IRR} \cong 18.01$$

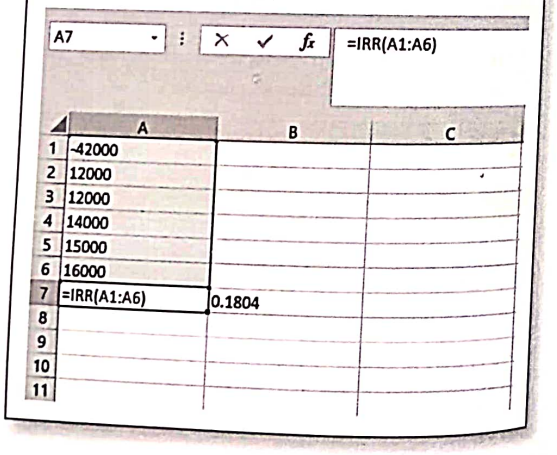
Second Method (Using Excel)

Excel provides an easy way of computing IRR. Instead of going into the complicated manual calculations, we can simply use the IRR function in Excel. As shown in Figure 8.2, the IRR comes out to be 18.04.

- We first need to enter all the values of cash outflows and inflows, respectively.
- Choose the function IRR from the f_x function.

FIGURE 8.2

Calculating IRR Using Excel



- Fill the values of arguments by selecting all the cells containing values.
- Leave the **Guess** argument blank.

Check Your Understanding

Consider an investment decision that has the following expected cash flows. What is the IRR of this investment decision?

Year	*Amount*
0	–₹100,000
1	₹45,000
2	₹45,000
3	₹45,000

Ans: 16.65 per cent

IRR is a widely used technique that is used in our day-to-day life from a simple application such as insurance policy and mutual fund investment. A detailed account of the steps involved in computing IRR-based decisions regarding the viability to invest in any project are explained as follows.

8.4. COMMON APPLICATIONS OF IRR

Mostly, IRR is used in investment decisions such as in insurance policies and mutual fund investments. The methodology to calculate IRR remains the same in either of the investments. However, we take the case of calculating returns in insurance policies only which is a more common form of investment. Insurance plans include endowment plans, moneyback plans, pension plans or any other plan in which we pay premiums. The three examples we consider here are:

Endowment Plans with Moneyback at Maturity

Suppose we have to pay an annual premium of ₹30,000 for 20 years. At the end of the 21st year, we get ₹10 lakh. The IRR (i.e., the return on this investment) can be calculated as shown in Figure 8.3.

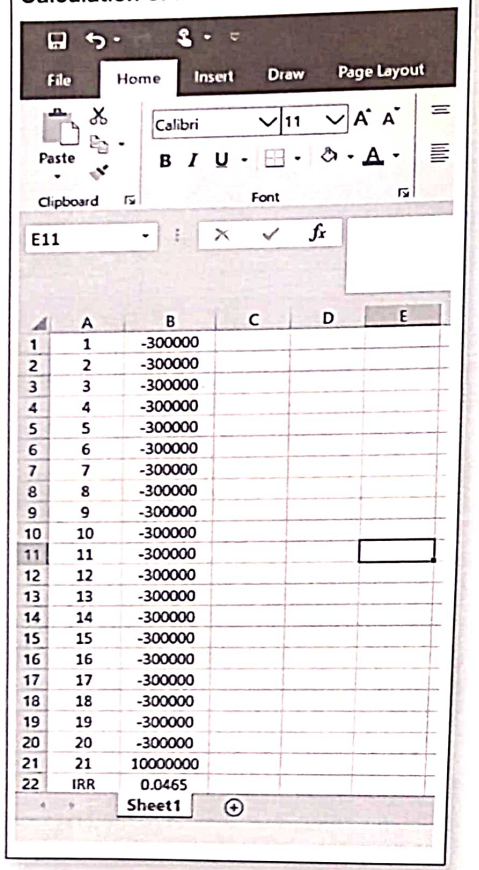
Solving in Excel, using IRR function, we get a return of 4.65 per cent. As indicated earlier, the minus sign represents the outgoing cash flow (i.e., premiums paid for 20 years in our example).

Moneyback Policies with Moneyback in between

Suppose the premium is ₹25,000 per year for 20 years, but we get, say at the end of 4th, 8th, 12th and 16th year an amount of ₹30,000, ₹30,000, ₹30,000 and ₹50,000, respectively. The sum assured at the time of maturity at the end of the 21st year is ₹500,000. In such a case, we need to adjust the amount paid in the 4th, 8th, 12th and 16th year. Accordingly, we take the difference

FIGURE 8.3

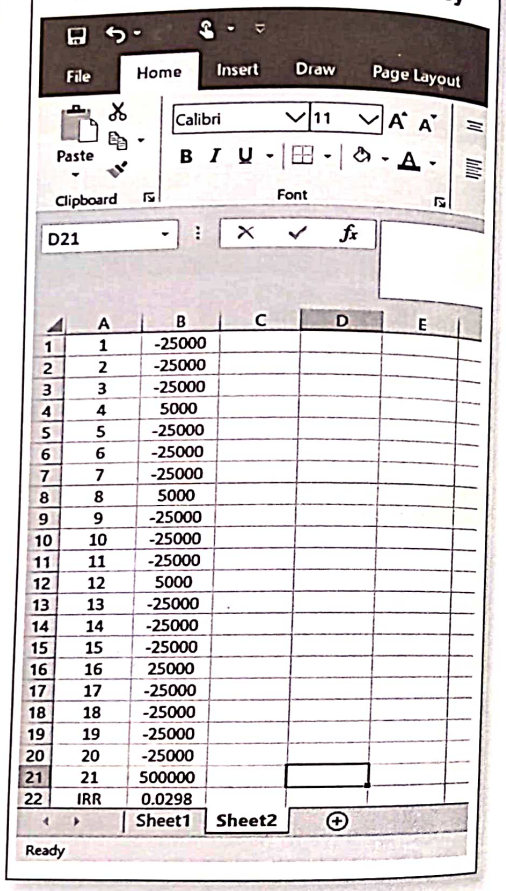
Calculation of IRR of Endowment Plans



	A	B	C	D	E
1	1	-300000			
2	2	-300000			
3	3	-300000			
4	4	-300000			
5	5	-300000			
6	6	-300000			
7	7	-300000			
8	8	-300000			
9	9	-300000			
10	10	-300000			
11	11	-300000			
12	12	-300000			
13	13	-300000			
14	14	-300000			
15	15	-300000			
16	16	-300000			
17	17	-300000			
18	18	-300000			
19	19	-300000			
20	20	-300000			
21	21	10000000			
22	IRR	0.0465			

FIGURE 8.4

Calculating IRR in a Moneyback Policy



	A	B	C	D	E
1	1	-25000			
2	2	-25000			
3	3	-25000			
4	4	5000			
5	5	-25000			
6	6	-25000			
7	7	-25000			
8	8	5000			
9	9	-25000			
10	10	-25000			
11	11	-25000			
12	12	5000			
13	13	-25000			
14	14	-25000			
15	15	-25000			
16	16	25000			
17	17	-25000			
18	18	-25000			
19	19	-25000			
20	20	-25000			
21	21	500000			
22	IRR	0.0298			

between the amount we pay and the amount we get. The adjusted amount will be ₹5,000 (–25,000 + 30,000) in the 4th, 8th, 12th and ₹25,000 (–25,000 + 50000) in the 16th year. Calculating IRR in Excel will yield 2.98 per cent (Figure 8.4).

A Simple Pension Plan

Pension plans are quite popular now. Let us take an example to understand its return before considering an investment in it. We have been asked to pay a premium of ₹25,000 for 10 years. For the next 10 years, we do not get anything. However, for the next 40 years, the company promises us to pay a pension of ₹25,000 each year. If we analyse it in absolute terms, we spend an

amount of ₹250,000 and get in return ₹1000,000. It sounds attractive, and a lot of investors may fall for it. To analyse it correctly, we have to take into account the application of IRR. IRR is a cruel tool as it makes us come to terms with reality. Calculating IRR in excel, we enter –25,000 in the first 10 cells, 0 in the next 10 cells and 25,000 in the next 40 cells. This will yield a return of 4.37 per cent. Now, it is up to us whether we want to invest at this rate of return.

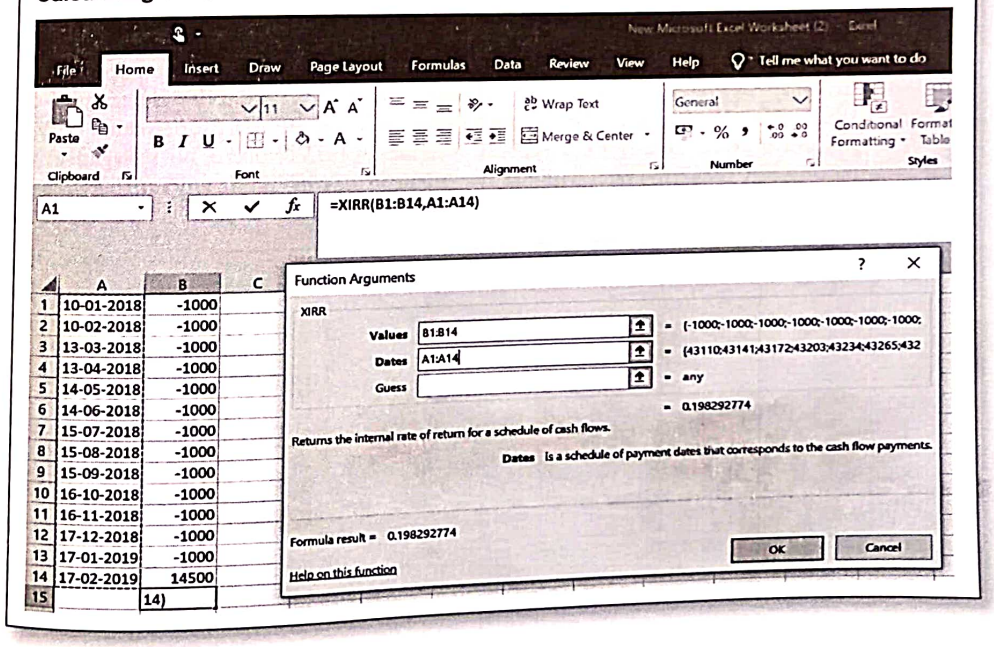
8.5. IRR VERSUS EXTENDED INTERNAL RATE OF RETURN (XIRR)

The only difference between IRR and XIRR is that the latter allows us to assign a date to each cash flow. Let us take an example. We make a payment of ₹1,000 each month for 13 months on *specific dates*, as shown in Figure 8.4. After the end of this period, its redemption value is ₹14,500. The XIRR formula in Excel additionally brings in **Dates** argument. We generally leave the **Guess** argument blank. It can be seen in formula result that this investment is yielding a return of 19.8 per cent. Note that we have to be very careful about the format of the dates. Excel only recognizes the date format of which is given in Figure 8.5.

There are several advantages and disadvantages associated with the method of IRR, as discussed in Section 8.6.

FIGURE 8.5

Calculating XIRR



	A	B	C
1	10-01-2018	-1000	
2	10-02-2018	-1000	
3	13-03-2018	-1000	
4	13-04-2018	-1000	
5	14-05-2018	-1000	
6	14-06-2018	-1000	
7	15-07-2018	-1000	
8	15-08-2018	-1000	
9	15-09-2018	-1000	
10	16-10-2018	-1000	
11	16-11-2018	-1000	
12	17-12-2018	-1000	
13	17-01-2019	-1000	
14	17-02-2019	14500	
15		14)	

8.6. MERITS AND DEMERITS OF IRR

8.6.1. Merits of IRR

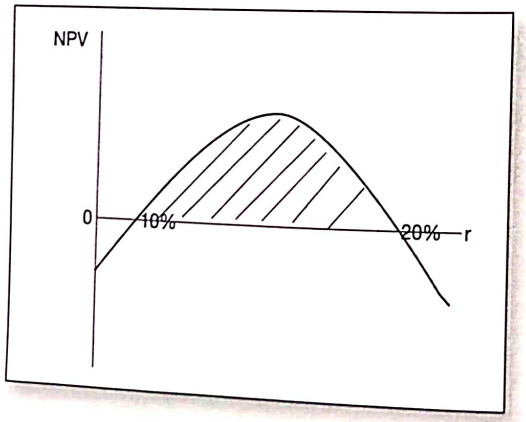
1. This method recognizes the time value of money, and all the cash flows over the life of the project/investment.
2. This method itself determines a rate.
3. It helps in achieving the objective of wealth maximization.
4. This method gives a rate that can be compared with the standard rate of return to evaluate the project.

8.6.2. Demerits of IRR

1. The calculation is difficult and tedious.
2. Sometimes this method gives more than one IRR, especially when the sign of cash flows changes. This leads to confusion. To understand it better, let us take the following example.

Cash Flows	*Years from Now*
–100	0
230	1
–132	2

Calculating IRR in Excel yields 10 per cent. Let us start by guessing a random IRR, say 20. We get NPV as 0 even at 20 per cent, and our guess turns out to be correct. This clearly shows that more than one IRR is present in this case. So which IRR should be chosen? The solution lies in the graphical representation shown below. When r = 0, NPV is negative (–2). As *r* tends to be 10 per cent, NPV approaches 0. Between 10 and 20 per cent, NPV is positive, and at 20 per cent, NPV is again 0. Increasing *r* further makes NPV negative. We will undoubtedly prefer the rates with positive NPV. So we can choose any rate between 10 and 20 per cent, which is comparable to similar businesses.



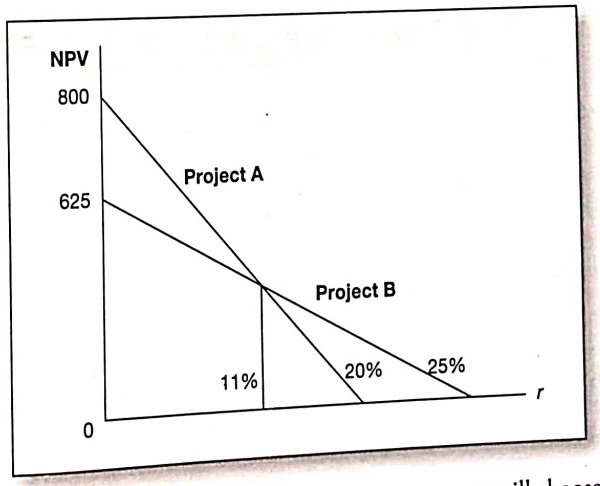
3. When we compare two mutually exclusive projects, an intuitive decision rule would be to accept the project with the highest IRR. This rule is unfortunately incorrect as we show in the following example.

Consider two projects, A and B with the following cash flows. Which one would we choose?

Year	*Project A*	*Project B*
0	–2,000	–2,000
1	400	2,000
2	2,400	625
IRR	20%	25%
NPV (@ 5%)	558	472
NPV (@ 20%)	0	101
NPV (@ 11%)	308	309

Based on IRR, we will be inclined towards Project B as its IRR is higher. But IRR is plagued by a short-term bias. Even though the two projects are of the same physical length, Project A yields higher returns in the later years, while Project B yields a higher return in the earlier years. IRR favours Project B because of earlier high returns. But comparing 20 per cent with 25 per cent directly is not the right thing to do. We are comparing this internally, *but we should compare it with a benchmark, for example, the cost of capital (r)*. We are more interested in knowing *how well we are doing relative to the competitors* and not in our internal rate of return.

We here calculate NPV using three rates—5, 20 and 11 per cent as shown above. If *r* = 5 per cent, we will choose Project A (as its NPV is more), if *r* = 20 per cent, we will then select Project B. However, at *r* = 11 per cent, we are indifferent between the two projects as the NPV is almost the same.



It is clear from the graph that if *r* is less than 11 per cent, we will choose Project A, and if *r* is greater than 11 per cent, we will choose Project B (because of higher NPV). In any

case, we are not going to select any project if r is more than 25 per cent because of negative NPV.

4. IRR also suffers from a small investment bias. Consider two projects, A and B with the following cash flows:

Year	*Project A*	*Project B*
0	–5,000	–50,000
1	7,500	64,000
IRR	50%	28%

In Project A, since the investment is smaller (–₹5,000), the return will turn out to be higher because of the smaller denominator as compared to Project B (Growth = [FV – PV] / PV). IRR is a percentage and favours little things over large. This shows there is a built-in bias of scale.

5. IRR implicitly assumes that the obtained positive cash-flows are *reinvested at the rate at which they are generated.* This assumption, however, may not always hold true. To get rid of this limitation, an alternative method, namely *modified internal rate of return* (MIRR) is used. *MIRR* assumes the reinvestment taking place at the rate of the cost of capital.

So to summarize, this chapter is featured to explain two important criteria in decision-making in the selection of the project and/or investment. These two criteria, NPV and IRR, are closely linked to each other and have their own sets of problems. However, in practice, IRR is a prevalent method and widely used than NPV. This is because it is improper to calculate NPV unless the appropriate discount rate is known. Moreover, in investment decisions, individuals and analysts seem to prefer rates of return rather than absolute values. Similarly, IRR is a straightforward mode of disclosing information about a project.

PRACTICE QUESTIONS

1. Hariti needs ₹12,000 to purchase the mobile she wants to purchase. Her father offers to loan her the money but wants to teach her the time value of money. He offers to have her repay the loan in the future using pocket money. They agree that she will repay ₹4,500 in 3, 6 and 15 months from now. Assuming a 6 per cent cost of capital and that compounding is done monthly, what is the NPV of the loan?
Ans. ₹976.13.

2. Dolly runs a boutique. She decides to expand her business. The additional rental space will cost her ₹7,200 per year at the end of each year and an immediate one-time investment of ₹10,000 for other machines. The boutique will generate an extra income of ₹10,000 per year at the end of each year. The business is expected to last 20 years. The discount rate is 10 per cent. What is the NPV of the boutique project? (round off decimals).
Ans. ₹13,838.

3. Consider two investments, A and B with the following cash flows.

Year	*Cash Flows*	
	A (₹)	*B (₹)*
Today	–5,000	–5,000
1	1,000	0
2	2,000	0
3	2,000	4,000
4	1,500	1,000
5	1,000	2,500

Which investment is more profitable if the cost of capital of both investments is 7 per cent?
Ans. Investment A is more preferred because it is more profitable at ₹1,171 than ₹810.55 in Investment B.

4. Which investment would you choose if you are offered the following two investments?

Year	*Cash Flows*	
	A	*B*
Today	–₹1,000	–₹1,000
1	₹700	₹0
2	₹1,200	₹2,100

The cost of capital for both investments is 10 per cent.
Ans. Investment B.

5. Varun owned a factory that got burnt down in an accident. The worth of vacant land is ₹70,000. He also received ₹200,000 from his insurance company. Varun plans to build another factory on the same land which will cost ₹275,000. He expects that the value of the finished factory in the market will be ₹385,000 next year. The discount rate is 10 per cent. What are the NPV and IRR of this decision?
Ans. ₹5,000 and 11.59 per cent.

6. Sunshine technology is considering the following three projects, and it has to choose only one. Project A has an expense of ₹100,000 today, and the cash flows over the next three years are ₹20,000, ₹40,000 and ₹70,000. Project B has an outflow of ₹110,000 and cash flows of ₹40,000, ₹80,000 and ₹20,000. The cost of Project C is ₹120,000 and cash inflows of ₹0, ₹20,000 and ₹141,500. Which project this company should choose if the cost of capital for similar projects is 6 per cent.
Ans. Project B.

9

CHAPTER

Loan Amortization

LEARNING OBJECTIVES

After reading this chapter, the readers will be able to understand

- The difference between flat and reducing balance type of loans
- The calculation of equated monthly instalment (EMI)
- Loan amortization schedule
- Calculating various components in loan schedule using Excel
- Effective annual rate

Loan amortization is one of the most basic applications used in the time value of money. A loan mainly involves a lending exercise. The lender lends money to the borrower to get back his money along with an interest on it. The calculation of loan requires the following basic concepts:

1. **Equated monthly instalment (EMI):** It is a periodic payment made on the loan. It primarily depends upon the rate of interest (r) and the number of periods (n) over which the amount is to be paid. The *higher* the number of periods and/or *lower* the r, the lesser will be the EMI and *vice versa*. The term is equivalent to **PMT** (short for payment) in Excel.
2. **Principal payment:** It refers to the amount of money returned out of the principal amount borrowed.
3. **Interest payment:** It is the amount of interest paid on loan to the lender.

9.1. TYPE OF LOAN: FIXED INTEREST VERSUS REDUCING BALANCE METHOD

Two common methods adopted to calculate the loan are fixed interest and reducing balance. The difference between the two is that in the fixed interest method, interest is calculated on the *principal amount borrowed through the loan period.* While, in reducing balance method, the interest rate is calculated on the remaining principal amount at any time. In this section, calculations based on the fixed rate method is elucidated.

One of the advantages of a fixed rate type of loan is its simplified computation of EMI and its components, namely interest amount and principal payment. Besides, even if there is an increase in the interest rate on loans, we continue to pay the same fixed amount as the interest rate on loan taken is already fixed. Yet in practice, loans based on fixed rate are usually not favoured. We will understand this statement subsequently. The formula to calculate EMI is the following:

$$\frac{P+(P\times R\times N)}{N\times 12}$$

where P = Principal amount, R = Rate of interest, N = Number of years (multiplied by 12 in case calculations are to be made monthly).

That is, the total of principal and total interest payable for the time for which the loan has been taken is divided by the number of instalments. Let us take an example to understand the calculations. We consider a loan of ₹100,000 that is to be paid off with three *annual payments* with a fixed rate of 10 per cent rate. Substituting the values in the stated formula, the EMI of this loan is:

$$=\frac{100{,}000+(100{,}000\times 0.10\times 3)}{3}=₹43{,}333$$

Alternatively, using Excel, the syntax of the EMI amount is shown in Figure 9.1.

FIGURE 9.1

Calculating EMI Using Fixed Rate Method

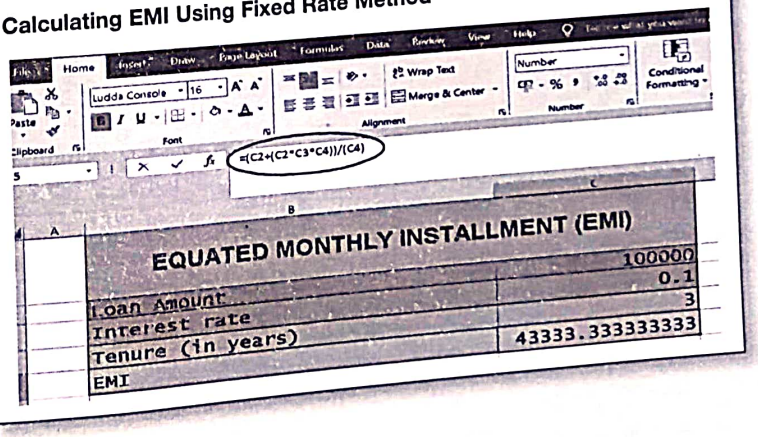


TABLE 9.1

Loan Amortization Table in Fixed Rate Method

Year	*EMI Amount (in ₹)*	*Interest Amount Paid (in ₹)*	*Principal Amount Paid (in ₹)*
1	₹43,333	₹10,000	₹33,333
2	₹43,333	₹10,000	₹33,333
3	₹43,333	₹10,000	₹33,333

Given the basis of calculation methodology, the repayment table showing interest and payment amount can be drawn as shown in Table 9.1. During the tenure of the loan, the interest amount paid is ₹30,000, and the total repayment amount is ₹130,000.

9.1.1. Arithmetic Derivation of Payment (PMT)/EMI in Reducing Balance Method

We can derive the payment factor mathematically in the following way:

Let us consider three years of timeline with r as the rate of interest per annum.

The PV of a cash flow or payment (C) occurring after a year would be:

$$PV = \frac{C}{(1+r)}$$

The PV of a cash flow occurring after two years would be:

$$PV = \frac{C}{(1+r)^2}$$

Similarly, the PV of cash flow after three years would be:

$$PV = \frac{C}{(1+r)^3}$$

The PV of annuity of these three years together will be:

$$PV = \frac{C}{(1+r)} + \frac{C}{(1+r)^2} + \frac{C}{(1+r)^3}$$

Or it can be rewritten as follows:

$$PV = C\left(\frac{1}{[1+r]} + \frac{1}{[1+r]^2} + \frac{1}{[1+r]^3}\right)$$

$$PV(1+r) = C\left(1 + \frac{1}{[1+r]} + \frac{1}{[1+r]^2}\right)$$

$$PV(1+r) - PV = C\left(1 + \frac{1}{[1+r]} + \frac{1}{[1+r]^2}\right) - C\left(\frac{1}{[1+r]} + \frac{1}{[1+r]^2} + \frac{1}{[1+r]^3}\right)$$

$$PV(r) = C\left(1 - \frac{1}{[1+r]^3}\right)$$

$$PV = \frac{C\left(1 - \frac{1}{[1+r]^3}\right)}{r}$$

$$\text{or } C \text{ or } PMT = \frac{PV}{\frac{\left(1 - \frac{1}{[1+r]^3}\right)}{r}}$$

If number of years = n, then the generalized formula is:

$$PMT = \frac{PV}{\frac{\left(1 - \frac{1}{[1+r]^n}\right)}{r}}$$

If the loan amount (i.e., PV), interest rate and the number of years for which loan is known, we can use the above equation to calculate the PMT (EMI). Assuming r as 10 per cent for n for 3 years, the following expression will turn out as 2.4869:

$$\frac{\left(1 - \frac{1}{[1+r]^n}\right)}{r}$$

9.1.2. Calculating Payment or EMI Using Present Value Interest Factor of Annuity (PVIFA)

We can avoid the tedious denominator calculation and simply refer to the PVIFA table. For example, a loan amount of ₹100,000 is taken for three years at an interest rate of 10 per cent. We can quickly solve the denominator part by looking at the intersection of the 10 per cent interest rate and the number of periods equal to 3 in the PVIFA table.

- The interactive figure is 2.4869 (compare it with the expression calculated)
- Dividing this value by PV (i.e., 100,000) gives us the EMI value of ₹40,211 (rounded off)

Note that this EMI value is calculated yearly, which means each *year* payment of ₹40,211 has to be made. For calculating *monthly* EMI payments, we know that the rate of interest needs to be adjusted by the number of months and will be in decimals. *Since the PVIFA table does not contain the interest rate in decimals, we, therefore, cannot use the table when calculating the EMI amount monthly.*

Check Your Understanding

Consider a loan of ₹200,000 that is to be paid with 10 annual payments with a 10 per cent rate of interest. Calculate the amount of EMI per year.

Ans. ₹25,901 (rounded off).

9.2. LOAN AMORTIZATION WITH TABLE

Let us understand the loan amortization concept with the help of a table. We start by taking the same loan example in which we borrow ₹100,000 for three years at the rate of 10 per cent. The balance at the beginning of the *first year* or *time 0* is ₹100,000 as it is the amount we are starting with. The first payment will occur at the end of the first year, which is ₹40,211 as calculated earlier. Since the interest rate is fixed, the yearly payments for all the years will be ₹40,211. At the end of the first year, we break the amount paid to the lender, that is, ₹40,211 into two components—*interest* and the *principal amount repaid.*

The interest amount paid at the rate of 10 per cent of the amount borrowed is ₹10,000 (Table 9.2). Subsequently, we need to find out how much of the loan amount is paid. The loan repayment is the *difference* between the yearly annual payment (PMT) and the interest paid, that is, ₹40,211 – ₹10,000 = ₹30,211 in the first year. This means that a repayment of ₹30,211 has been done out of the total loan amount. Payment of ₹30,211 implies that at the beginning of the second year, the amount of loan left for repayment is calculated by subtracting the principal amount repaid from the beginning balance, that is, ₹100,000 – ₹30,211 which is equal to ₹69,789.

Similarly, in the second year, the interest payment will be 10 per cent of the beginning amount of ₹69,789, which is ₹6,979. The principal amount repaid will be the difference between the

TABLE 9.2

Loan Amortization Table

Beginning Year	*Balance at the Beginning of the Year*	*Yearly Payment*	*Interest Paid*	*Principal Amount Paid*	*Balance at the End of the Year*
Today	–	–	–	–	100,000
1	100,000	40,211	10,000	30,211	69,789
2	69,789	40,211	6,979	33,232	36,557
3	36,557	40,211	3,656	36,555	2[a]

Note: [a]Display of ₹2 instead of 0 is on account of rounding off errors.

annual payment of ₹40,211 and ₹6,979 (interest amount), which equals to ₹33,232. Notice that for the third year, the EMI payment (₹40,211) is coming out to be more than the beginning balance of the third year. The reason behind this is that the outstanding balance of ₹36,557 at the beginning of the third year also has an interest payment due on it, that is of ₹3656. In this manner, the entire loan gets paid off. A point of observation here is that the proportion of interest amount decreases, whereas the principal amount increases in the EMI component over time.

Check Your Understanding

Suppose you borrow ₹600,000 and pay it back over 20 years in annual payments. If the interest rate is 5 per cent, complete the information corresponding to the second- and third-year payments.

Likewise, this procedure can also be followed for calculating *monthly* EMI, as shown in the table. However, we cannot use the PVIFA table in this case because of the reason stated earlier. Hence, PMT needs to be calculated manually at this point in case of the unavailability of any financial calculator. In this case, the monthly EMI would turn out to be an amount of ₹3,227. This leads us to an important conclusion that as the number of periods increases, the EMI amount decreases.

9.3. AMOUNT OF LOAN AFTER n PERIODS

Suppose someone is interested in finding out the balance loan amount at the beginning of the 16th month, it would be a very cumbersome task to find it out manually by following the procedure described in Table 9.3. *We must understand that the answer to this question is equivalent to finding out the PV of the loan amount at the beginning of the 16th month.*

To find out the PV, we need the following:

- **r** which is 0.1/12 (compounding monthly).
- **PMT** which is –₹3,227 (the amount is negative as it is outgoing payment)
- Leave future value (**Fv**) blank.
- Write 0 as the **type** argument (Type 0 implies payment at the end of the period).
- The most important thing is to decide what to take as the total number of periods (**Nper**). *We count Nper in such cases as the number of periods left of the loan amount.* For example, in this case, the number of periods left is 21 (by subtracting the number of periods in which we have repaid the loan amount, that is 15, from the total number of periods for which loan is taken, that is 36).

The balance of loan amount at the beginning of the 16th month is ₹61,927. The figures can be realized from the calculations done in Table 9.3 and Figure 9.2.

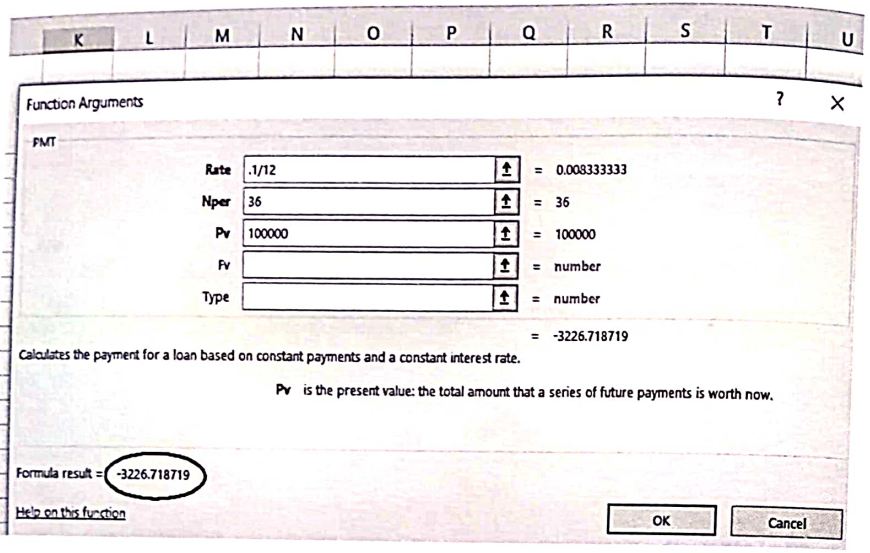
Check Your Understanding

Amit borrows ₹100,000 at 8 per cent compounded annually. Equal annual payments are to be made for six years. However, at the time of the fourth payment, he chose to pay off the loan. How much should be paid?

Ans. ₹60,207.

FIGURE 9.3

Calculating EMI Amount Using Excel



the dialogue box named **Function Arguments**, we fill in the values of different parameters asked. The illustration of the same example discussed before is presented in Figure 9.3. In this example,

- **Rate** is taken as 0.1/12 (compounded monthly)
- **Nper** as 36 (total number of months are 36)
- **PV** (loan amount) as ₹100,000

The corresponding value of EMI can be seen in Figure 9.3 as ₹3,227 (rounded off). The amount is the same as in Table 9.3. Note that, this function can also alternatively be written as '= PMT (0.1/12, 36, 100,000)' in the **Formula Bar**.

9.5. CUMULATIVE PRINCIPAL AND INTEREST AMOUNT USING EXCEL

If we wish to find out how much of the total principal amount we have repaid at the end of, say the 15th period, we use the function **CUMPRINC**. Instead of finding the PV value as we did

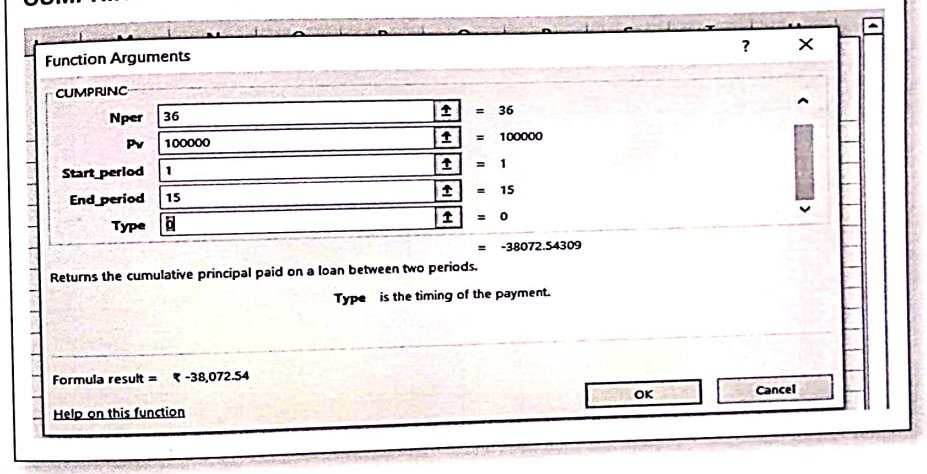
earlier, we can simply make use of this function. Select the function **CUMPRINC** after clicking on f_x. In the dialogue box, we start filling in the argument values.

- **Rate** is taken, as usual, that is, 0.1/12 as in our example.
- **Nper** is 36, that is, the total number of periods in the calculation.
- **PV** (loan amount) is ₹100,000.
- **Start_period** is taken as 1 since we want to calculate from the starting period. It is not always necessarily 1 as we may wish to calculate by taking any starting period.
- **End_period** is taken as 15 as in our example as the number of periods completed is 15.
- **Type** is taken as 0 as type refers to the timing of the payment.

Thus, at the end of the 15th period, the total principal amount that has been repaid is ₹38,073 (Figure 9.4). The remaining balance is ₹100,000 – ₹38,073 = ₹61,927. This amount can be cross verified from Figure 9.2 as well. Note that the (–) sign represents the cash outflow or the payment made.

FIGURE 9.4

CUMPRINC Function in Excel



Check Your Understanding

Manisha takes a loan of ₹5,000,000 from a bank to buy a house for 20 years. The bank agrees to provide a loan at an interest rate of 8 per cent per annum compounded monthly. At the end of five years, Manisha is interested in knowing how much of the total principal amount she has paid in the last two years?

Ans. ₹40,358.

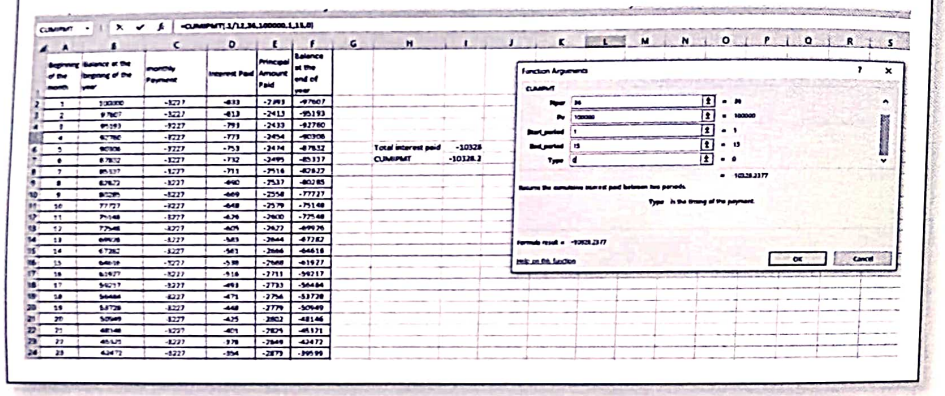
Similarly, the total amount paid between any two periods can be found out using the Excel function **CUMIPMT**. The dialogue box after selecting this function requires the following parameters:

- **Rate** is 0.1/12 as in our example.
- **Nper** is 36, that is, the total number of periods in the calculation.
- **PV** (loan amount) is ₹100,000.
- **Start_period** is taken as one since we want to calculate from the starting period. It is not always necessarily one as we may wish to calculate by taking any starting period.
- **End_period** is taken as 15; as in our example, we want to find out the interest paid at the end of the 15th period.
- **Type** is taken as 0 because the type refers to the timing of the payment.

Thus, the total interest amount paid on the loan amount until the 15th period is ₹10,328. We can cross-check the result by adding the interest amounts paid up to the 15th period, which is coming out to be the same (Figure 9.5).

FIGURE 9.5

CUMIPMT Function in Excel



Check Your Understanding

Now, Manisha wants to know how much total interest amount she will pay in the first 15 years? Ans. ₹4,590,556.

9.6. PPMT AND IPMT

There are two more related functions besides the functions already discussed, which are often used in loan repayments:

- **PPMT function:** It helps us to find how much of the principal amount we have to repay in nth month.
- **IPMT function:** It enables us to find how much of the interest amount we will have to repay in the nth month.

9.6.1. PPMT Function

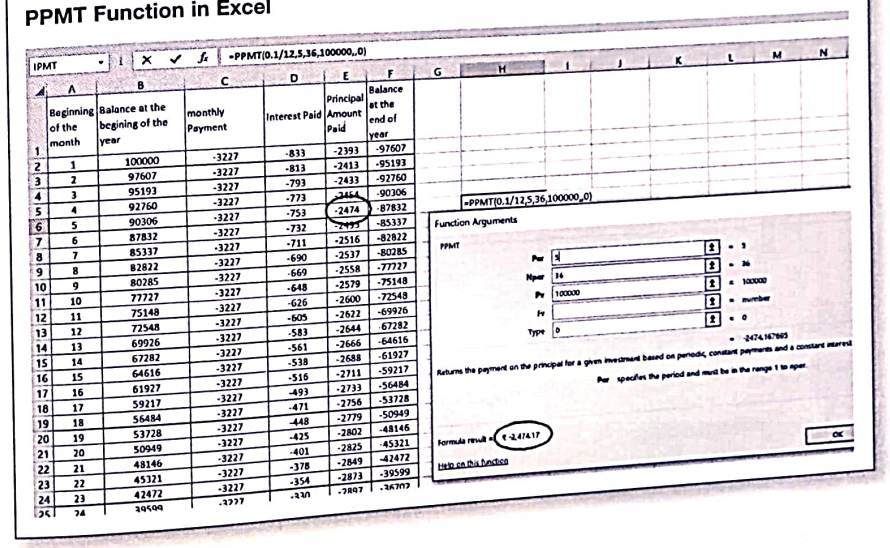
In the ongoing example, if we randomly want to find how much of the principal amount we will have to repay, say in the 5th month, we use this function. After selecting the **PPMT** function, the following parameters are required:

- **Rate** as 0.1/12.
- **Per** is 5 as it is the period in which we want to find the principal amount of repayment.
- **Nper** is the total number of periods, that is, 36.
- **PV** is the loan amount taken which is ₹100,000.
- **Fv** is the future value and is left blank.
- **Type** is taken as 0.

Performing this task would have been time-consuming, had we used the table procedure to find this. The **PPMT** function has simplified the work to a great extent, as we just have to insert the corresponding values. Both these methods are leading to the same results as shown in Figure 9.6.

FIGURE 9.6

PPMT Function in Excel



Check Your Understanding

Mr X has taken a car loan of ₹50,000 at 6 per cent per annum compounded monthly for three years. Calculate the principal amount paid in his 25th instalment?

Ans. ₹1,433.

9.6.2. IPMT Function

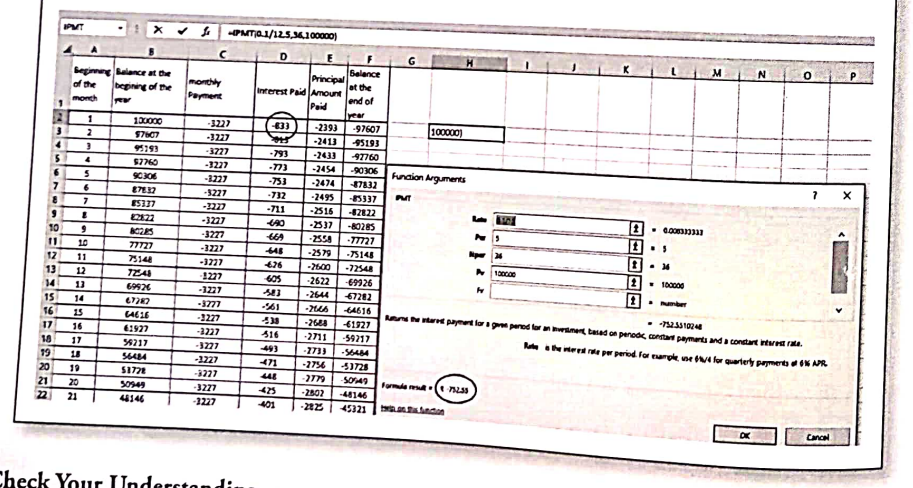
We may want to find out the interest payment on the loan amount, say in the 5th month. The IPMT function will help us in carrying out this task by requiring the values of the same arguments as in the **IPMT** function.

- **Rate** as 0.1/12.
- **Per** is 5 as it is the period in which we want to find the principal amount of repayment.
- **Nper** is the total number of periods, that is, 36.
- **PV** is the loan amount taken which is ₹100,000.
- **Fv** is the future value and is left blank.
- **Type** is taken as 0.

It can be seen in Figure 9.7 that the interest amount that is to be paid at the end of the 5th month is ₹753. This function has made the work much simpler and easier.

FIGURE 9.7

IPMT Function in Excel



Check Your Understanding

Mr X now also wants to find out the interest amount that he will have to pay in his 25th instalment?

Ans. ₹88.36.

9.7. EFFECTIVE ANNUAL INTEREST RATE

Arun is facing a situation whether to take a loan of ₹1,000 for one year at an interest rate of 10 per cent per annum compounded annually or at an interest rate of 10 per cent compounded quarterly. Which option should he opt for? To find out a solution for Arun, let us first consider Option 1. By the usual formula of compound interest rate, we can find out how much he will have to repay at the end of 1 year:

$$\text{Fv of this loan} = P + \text{Pr}$$

$$\text{Equivalently, } P\,(1 + r)^n \qquad (1)$$

$$\text{that is, } 1{,}000\,(1 + 0.10)^1$$

$$= ₹1{,}100$$

Let us look at Option 2 now. Notice carefully that the rate of interest is 10 per cent per annum but compounding is done quarterly, which means, it is compounded 4 times in a year. Adjusting the R and N, the formula becomes as follows:

$$\text{Fv} = 1{,}000\left(1+\frac{10}{4}\right)^4 \qquad (2)$$

$$= ₹1{,}103.813$$

In Option 2, Arun has to pay ₹1,103.813 as compared to ₹1,100 in Option 1. The amount in Option 2 is more as the compounding is done is four times in a year as compared to Option 1 where the compounding is done only once in the year.

But instead of working through this formula which does not take compounding effect into account, we can directly find out the effective interest rate by converting the stated interest rate. The effective interest rate formula can be derived as follows:

$$\text{From Equation (2), } ₹1{,}103.813 = 1{,}000\left(1+\frac{10}{4}\right)^4$$

$$\frac{1{,}103.813}{1{,}000} = \left(1+\frac{10}{4}\right)^4$$

$$\frac{1{,}000+103.813}{1{,}000} = \left(1+\frac{10}{4}\right)^4$$

(1,000 + 103.813 => principal amount plus interest amount)

It is equivalent to:

$$1 + 0.1038 = \left(1+\frac{10}{4}\right)^4$$

$$0.1038 = \left(1+\frac{10}{4}\right)^4 - 1$$

So 0.1038 or 0.1038 × 100, which is equal to 10.38 is the effective interest rate. The above formula for the effective annual rate (EAR) can be generalized as follows:

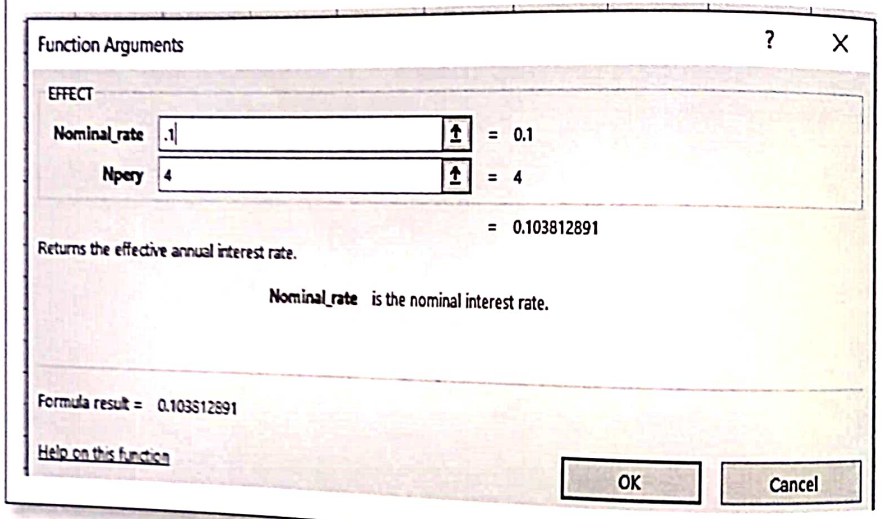
$$EAR = \left(1+\frac{R}{N}\right)^{n} - 1$$

Thus, by comparing the EAR, we can conclude that Option 1 is better for Arun as the interest rate of 10 per cent is less than 10.38 per cent in Option 2. It is worth emphasizing that in the case of any investment or loan, we must not only look at the interest rate but also at the frequency which is being compounded.

This exercise can also be simply performed using the **EFFECT** function in Excel. To use this function, we click on f_x function and choose **EFFECT** from the dialogue box. The two arguments required to solve it are **Nominal_rate** which is the nominal interest rate, that is, 10 per cent in our case. The second one is **Npery** which is the number of times the compounding is taking place, that is, 4. The answer turns out to be the same at 10.38 per cent as shown in Figure 9.8.

FIGURE 9.8

Effective Annual Rate Function



To sum up, this chapter has provided the nitty-gritty of loans and has also introduced us with the functions which can help in reducing the burden of extensive manual calculations. The application of the concepts of loans discussed is widely used in our day to day life. Therefore, it is imperative to study such concepts diligently and thoroughly.

PRACTICE QUESTIONS

1. Ram borrows ₹3,000 to be paid back at the compound rate of 5 per cent per annum by the end of three years in three equal yearly instalments. How much will each instalment be? (Round to the nearest whole rupee).
   Ans. ₹1,102.

2. Mr X borrows ₹50,000 at 8 per cent compound interest. How much he has to pay equally at the end of each year, to settle his loan in five years? (Round off decimals).
   Ans. ₹12,523.

3. Ten years from now, Mr A will start receiving a pension of ₹3,000 a year. The payment will continue for 16 years. How much is the pension worth now, if the interest rate is 10 per cent?
   Ans. ₹9,954.

4. Ramesh is purchasing his first car. The car costs ₹1,000,000. They need to pay 20 per cent down payment and refinance the rest of the amount. He is considering a fixed-rate 30-year mortgage at a 5.25 per cent annual percentage rate (APR) with monthly payments. How much will be his first monthly payment be?
   Ans. ₹4,418.

5. You have been living in the house you bought 10 years ago for ₹300,000. At that time, you took a loan for 80 per cent of the house at a fixed rate of a 15-year loan at an annual stated rate of 9 per cent. You have just paid the 120th monthly payment. The interest rate has meanwhile dropped steadily at 6 per cent per year, and you think it is time to refinance the remaining balance. However, the fee to refinance the loan is ₹4,000. Should you refinance the remaining balance? How much would you save/lose if you decided to refinance?
   Ans. A gain of ₹4,647.

6. Two years ago, Ram purchased the mobile at ₹10,000, he paid ₹2,000 down and borrowed the rest. He took a fixed-rate 60-month instalment at a rate of 8 per cent per year. The interest rate has fallen during the last two years. He has the option of refinancing the balance amount at a new fixed rate of 6 per cent per year for three years. Should the amount be refinanced? How much the amount will be saved per month for the next three years in case refinance?
   Ans. Yes, ₹5.

7. Which of the following represents the lowest cost of credit on an effective annual rate of interest rate basis?
   a. 10 per cent APR, interest compounded semi-annually
   b. 9.5 per cent APR, interest compounded quarterly
   c. 9.25 per cent APR, interest compounded monthly.
   Ans. C, 9.65 per cent.

8. If a mortgage requiring monthly payments is advertised of having a 6 per cent APR, what is the effective annual rate on this mortgage?
   Ans. 6.16 per cent.

9. Suppose your credit card states that interest on unpaid balances is 22 per cent APR, with interest compounded monthly. What is the effective annual rate of interest on unpaid balances?
   Ans. 24.36 per cent.
10. Complete the following table for a loan of ₹80,000 that is repaid with five annual payments of ₹21,104.

Period	*Beginning Balance*	*Payment*	*Interest*	*Principal Repayment*	*Ending Balance*
Today	–	–	–	–	₹80,000
1	₹80,000	₹21,104			
2		₹21,104			
3		₹21,104			
4		₹21,104			
5		₹21,104			

10

CHAPTER

Correlation Coefficient

LEARNING OBJECTIVES

After reading this chapter, the readers will be able to understand

- The two measures of association—covariance and correlation
- The concept of covariance
- The concept of correlation and how it is different from covariance
- Methods to estimate correlation coefficients—Pearson's and Spearman's method
- The use of Excel in calculating correlation coefficients

In this chapter, we will introduce the concept, application and computation of another set of descriptive measures that relate to an association between two or more variables. To illustrate, we might be interested in finding out the sales of ice creams as temperature rises. The measures that attempt to quantify such a relationship between the variables and further help to answer the strength of relationship are called *measures of association*. We will be studying two such measures in this chapter, which are closely related to each other. These are *covariance* and *correlation*.

10.1. COVARIANCE

Covariance is a statistical measure that analyses the *linear relationship* between two random variables. It evaluates how the two variables vary together or covary. For instance, if one variable

goes up or down or remains constant, then what happens to the other variable. Accordingly, we can have the following types of linear relationships or covariances:

- A *positive covariance* indicates a direct relationship between the variables, that is, the two variables tend to move together in the same direction either upward or downward. So when one variable increases the other variable increases as well, or when one variable decreases the other variable decreases.
- A *negative covariance* indicates an inverse relationship between variables, that is, the two variables tend to move away or in the opposite direction from each other. So when one variable increases, the other one decreases or vice versa.
- A *zero covariance* indicates if the two variables are independent or not related to each other, the covariance will be zero (0). However, a zero covariance does not necessarily imply that the variables are independent. A zero covariance means that there is *no linear relationship* between the two variables and do not exhibit any sort of pattern.

It may be reiterated that the covariance measures the directional relationship only, and we are, therefore, interested merely in the sign of the value of covariance and thus ignore the size of the number. We cannot interpret anything about the strength of the relationship from the computations arrived at.

For two random variables X and Y, the covariance (denoted by Cov (X, Y) or σ_{xy}) is calculated using the following formula. The unit of measurement of covariance, nevertheless, is the product of the units of individual variables. For instance, the unit of measurement of covariance between height (in cm) and weight (in kg) will be cm × kg.

$$\text{Cov}(X, Y) = \frac{\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})}{N}$$

Here,

Σ = Sigma/operator for summation
$\bar{X}$ = Arithmetic mean of random variable X
$\bar{Y}$ = Arithmetic mean of random variable Y
N = Number of observations

Important note: When we calculate the covariance for a sample to infer about the population as a whole from which it is drawn, we use n – 1 instead of n in the denominator.

Therefore, the population covariance (σ_{xy}) is $= \dfrac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$

And the sample covariance (s_{xy}) is $= \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$

Here, μ_x = population mean of a random variable X, μ_y = population mean of a random variable Y, $\bar{x}$ = sample mean of a random variable X, $\bar{y}$ = sample mean of a random variable, N = number of observations in the population and n = number of observations in a sample.

10.1.2. Illustration

Sensex and Nifty 50 are the two benchmark stock market indices used in India. Both the indices measure the overall performance of the stock market and are indicators of the strength of the Indian economy. The table below summarizes the monthly rates of return on Sensex (X) and Nifty 50 (Y) for the year 2019. Calculate covariance and interpret the results.

Time Period (2019)	*Rates of Return (in Percentages)*	
	Sensex	*Nifty 50*
January	0.52	0.35
February	–1.07	0.32
March	7.82	4.58
April	0.93	3.13
May	1.75	–0.58
June	–0.80	2.34
July	–4.86	–2.47
August	–0.40	–4.46
September	3.57	1.41
October	3.78	3.3
November	1.66	4.23
December	1.13	1.11

Source: https://www.bseindia.com/; https://www1.nseindia.com/

First Method (Direct Method)

Time	*Sensex (X)*	*Nifty 50 (Y)*	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X}) \times (Y-\bar{Y})$
January	0.52	0.35	–0.65	–0.76	0.00494
February	–1.07	0.32	–2.24	–0.79	0.017696
March	7.82	4.58	6.65	3.47	0.230755
April	0.93	3.13	–0.24	2.02	–0.00485
May	1.75	–0.58	0.58	–1.69	–0.0098
June	–0.8	2.34	–1.97	1.23	–0.02423
July	–4.86	–2.47	–6.03	–3.58	0.215874
August	–0.4	–4.46	–1.57	–5.57	0.087449
September	3.57	1.41	2.4	0.3	0.0072
October	3.78	3.37	2.61	2.26	0.058986
November	1.66	4.23	0.49	3.12	0.015288
December	1.13	1.11	–0.04	0	0
Total	14.03	13.33			0.5993

$$\text{Cov}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{N}$$

Here, $\bar{X} = 14.03/12 = 1.169\%$

$\bar{Y} = 13.33/12 = 1.111\%$

$$\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y}) = 0.5993\%$$

$N = 12$

$$\text{Cov}(X, Y) = \frac{0.5993\%}{12}$$

$= 0.04994\%$

Second Method (Using Excel)

Use the **COVARIANCE.P** function to calculate the covariance of a population. If using the entire population is not possible, we can calculate the covariance of a sample drawn from the population using the Excel function **COVARIANCE.S**. To calculate the covariance using Excel, follow the stated steps:

Step 1. Enter the returns on Sensex (X) and Nifty 50 (Y) in the Excel sheet.

Step 2. Use the **COVARIANCE.P** function to calculate the covariance of the population. Type the '=' (equal to) sign in the cell where the result is desired, say 'C15', and enter the function name **COVARIANCE.P** and then select the cell range of numbers (data points of variable X), that is, **B2:B13** for **Array 1** and the cell range of numbers (data points of variable Y), that is, **C2:C13** for **Array 2.** The formula will be **=COVARIANCE.P (B2:B13, C2:C13)** for the covariance of population and **=COVARIANCE.S (B2:B13, C2:C13)** for the covariance of a sample as remarked in Figures 10.1 and 10.2.

Step 3. Press **Enter** for the result.

FIGURE 10.1

Step 1 in Calculation of Covariance

AVERAGE | =COVARIANCE.P(B2:B13,C2:C13)

	A	B	C	D	E
1	Time	Sensex (X)	Nifty 50 (Y)		
2	January	0.52%	0.35%		
3	February	-1.07%	0.32%		
4	March	7.82%	4.58%		
5	April	0.93%	3.13%		
6	May	1.75%	-0.58%		
7	June	-0.80%	2.34%		
8	July	-4.86%	-2.47%		
9	August	-0.40%	-4.46%		
10	September	3.57%	1.41%		
11	October	3.78%	3.37%		
12	November	1.66%	4.23%		
13	December	1.13%	1.11%		
14					
15	COVARIANCE	for population	=COVARIANCE.P(B2:B13,C2:C13)		
16		for sample	=COVARIANCE.S(B2:B13,C2:C13)		

FIGURE 10.2

Step 2 in Calculation of Covariance

Clipboard | Font

C15 | =COVARIANCE.P(B2:

	A	B	C	D
1	Time	Sensex (X)	Nifty 50 (Y)	
2	January	0.52%	0.35%	
3	February	-1.07%	0.32%	
4	March	7.82%	4.58%	
5	April	0.93%	3.13%	
6	May	1.75%	-0.58%	
7	June	-0.80%	2.34%	
8	July	-4.86%	-2.47%	
9	August	-0.40%	-4.46%	
10	September	3.57%	1.41%	
11	October	3.78%	3.37%	
12	November	1.66%	4.23%	
13	December	1.13%	1.11%	
14				
15	COVARIANCE	for population	0.04994%	
16		for sample	0.05448%	

Interpretation of Results

- A positive covariance of 0.04994 per cent indicates that there is a direct linear relationship between the variables, that is, returns on Sensex (X) and returns on Nifty 50 (Y).
- As the returns on Sensex rises, the returns on Nifty rises as well and vice versa.
- As the returns on Sensex decrease, so does the returns on Nifty and vice versa.
- The returns tend to move together in the same direction either upward or downward.

The limitation of using covariance, as mentioned earlier, is that it does not tell us anything about the strength of the relationship between the variables. Thus, covariance is not an appropriate measure when it comes to assessing the strength of the relationship. To overcome this limitation, we study correlation which deals with the strength of the relationship.

10.2. CORRELATION

Correlation is a statistical technique that measures both the direction and the strength of the linear relationship between the variables. The study of correlation is important as it quantifies the degree of association between the variables and shows how strongly the variables are related. If two

variables are correlated it means that if one variable increases or decreases, then the other variable also changes either in the same or opposite direction besides assessing the strength of the relationship. An example of correlation may be the increase in the price of a stock of a company and an increase in the company's profitability and revenue.

However, *correlation does not necessarily mean causation.* That is, correlation in no way confirms that variable X causes variable Y to vary or vice versa. For example, we may observe that during summers, when ice cream sales increase, the groundwater level reduces. Does that mean the consumption of ice cream causes depletion in the groundwater level or vice versa? In reality, this does not make any sense. It is the third factor, that is, the summer season, which influences the movement in both the variables. As it may be possible that due to the increase in temperature during summer, people tend to eat more ice cream, and the groundwater begins to evaporate due to hot temperature leading to depletion of water level. Therefore, correlation does not imply any causal relationship or cause and effect relationship between the variables.

There can be both linear and non-linear relationships that can exist between two variables, but in this chapter, we restrict ourselves mainly to the linear one. Correlation can be categorized as follows:

FIGURE 10.3

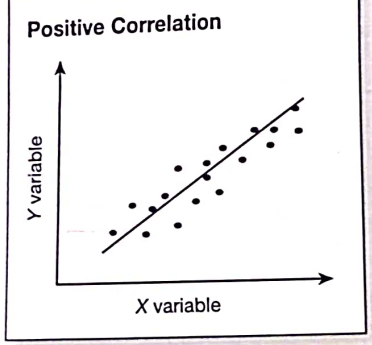
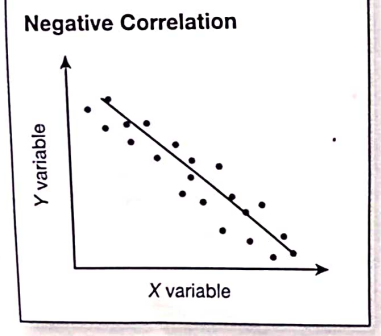


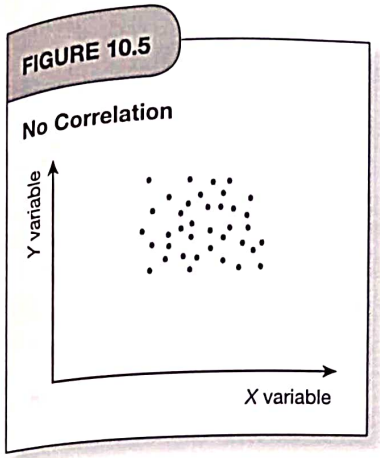
FIGURE 10.4



- A *positive correlation* indicates a direct relationship between the variables, that is, the two variables tend to move together in the same direction either upward or downward. So when one variable increases, the other variable increases as well, or when one variable decreases, the other variable also decreases. It is shown as an upward sloping line indicating movement of the variables in the same direction (Figure 10.3).
- A *negative correlation* indicates an inverse relationship between variables, that is, the two variables tend to move away or in the opposite direction from each other. So when one variable increases, the other one decreases. It is shown as a downward sloping line indicating movement of the variables in the opposite direction (Figure 10.4).
- A *zero correlation* indicates no relationship between the data sets at all. However, a zero correlation does not necessarily imply that the variables are independent. A zero correlation will indicate that there is *no linear relationship* between the two variables, and they do not exhibit any sort of pattern. It is shown by the widely scattered points, and there is no upward rising or downward sloping line (Figure 10.5).

The degree of association between data sets is determined by the *correlation coefficient.* A correlation

FIGURE 10.5



coefficient expresses or defines the strength of the association between variables, that is, whether it is strong or weak or non-existent. The value of the coefficient of correlation lies between +1 and –1. The positive and negative signs will indicate the direction of the relationship, that is, direct or inverse relation. The closer the value of the coefficient is to 1 (+ve or –ve), the stronger is the direct or inverse relationship.

If the value of the correlation coefficient attains the value of +1, it indicates a *perfect positive correlation.* It means when one variable increases by some percentage then the other variable also increases at the same percentage, or if one variable decreases by some percentage, then the other variable also decreases at the same percentage. The value of correlation coefficient between 0 and 1 also indicates a *positive correlation* although the change is not identical between the variables. Contrarily, the closer the value of the coefficient is to 0, the weaker is the relationship. In case the coefficient of correlation is equal to 0, it indicates that the variables are uncorrelated or have no linear relationship. Generally, the coefficient of correlation of |0.5| and above is considered to show a strong relationship between variables.

Correspondingly, the correlation coefficient of –1 indicates a *perfect negative correlation.* It means when one variable increases by some percentage, then the other variable decreases identically. The values between 0 and –1 indicate a *negative correlation.* It means when one variable increases by some percentage, then the other variable decreases but with lesser percentage, or if one variable decreases by some percentage, then the other variable increases but with a lesser percentage.

As mentioned before, the correlation coefficient of 0 indicates *no correlation between the variables.* It means that the change in one variable is not linked to another variable. The correlation coefficient can be estimated with the help of the following methods.

10.3. KARL PEARSON'S COEFFICIENT OF CORRELATION

Karl Pearson's correlation coefficient is the most widely used method to measure the correlation between two variables. It is also referred to as the *Pearson product–moment correlation coefficient* denoted by r. Pearson correlation coefficient has no unit of measurement. Thus, relationships across the variables can be measured regardless of the unit of measurements they are in. For example, the correlation coefficient between height (in cm) and weight (in kg) is neither expressed in cm nor kg. It is a number independent of the unit of measurement.

It may be noted that the Pearson coefficient is also independent of the change in the scale and the origin of the variables. Thus, the magnitude of the coefficient is not affected by the change in the unit of measurements of the variables. For example, the correlation coefficient between height (in cm) and weight (in kg) will be equal to the correlation coefficient between height (in m) and weight (in g).

Karl Pearson's coefficient of correlation between two variables, X and Y, is calculated using the formula given below:

$$r_{xy} = \frac{\text{Covariance between } X \text{ and } Y}{\text{Product of standard deviation of } X \text{ and standard deviation of } Y} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$\text{or} \quad r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{X})^2 \times \sum_{i=1}^{n}(y_i - \bar{Y})^2}}$$

Here,
x_i = ith observation of X variable
y_i = ith observation of Y variable
Σ = Sigma/operator for summation
$\bar{X}$ = Arithmetic mean of random variable X
$\bar{Y}$ = Arithmetic mean of random variable Y
n = Number of observations

or

$$r_{xy} = \frac{N\left(\sum x_i y_i\right) - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{\left(N\sum x_i^2 - \left[\sum x_i\right]^2\right)\left(N\sum y_i^2 - \left[\sum y_i\right]^2\right)}}$$

Here,
Σx_i = Sum of all observations of X variable
Σy_i = Sum of all observations of Y variable
$\Sigma x_i y_i$ = Sum of the product of X and Y
Σx_i^2 = Sum of squares of X variable
Σy_i^2 = Sum of the square of Y variable
N = Number of observations

10.3.1 Illustration

In the previous illustration of the monthly rates of return on Sensex (X) and Nifty 50 (Y) for the year 2019, we found that there is a direct linear relationship between the variables, that is, returns on Sensex (X) and returns on Nifty (Y). But we could not interpret how strong is the relation between the variables. Now, let us compute the correlation coefficient to determine the strength of this relationship.

First Method (Direct Method)

Time	Sensex (X; %)	Nifty 50 (Y; %)	X^2 (%)	Y^2 (%)	XY (%)
January	0.52	0.35	0.0027	0.0012	0.0018
February	–1.07	0.32	0.0114	0.0010	–0.0034
March	7.82	4.58	0.6115	0.2098	0.3582
April	0.93	3.13	0.0086	0.0980	0.0291

Time	Sensex (X; %)	Nifty 50 (Y; %)	X^2 (%)	Y^2 (%)	XY (%)
May	1.75	–0.58	0.0306	0.0034	–0.0102
June	–0.80	2.3	0.0064	0.0548	–0.0187
July	–4.86	–2.47	0.2362	0.0610	0.1200
August	–0.40	–4.46	0.0016	0.1989	0.0178
September	3.57	1.41	0.1274	0.0199	0.0503
October	3.78	3.37	0.1429	0.1136	0.1274
November	1.66	4.23	0.0276	0.1789	0.0702
December	1.13	1.11	0.0128	0.0123	0.0125
Total	14.03	13.33	1.219	0.952	0.755

Now, $\Sigma x = 0.1403$, $\Sigma y = 0.1333$, $\Sigma xy = 0.00755$, $\Sigma x^2 = 0.01219$, $\Sigma y^2 = 0.00952$
Substituting these values in the formula below:

$$r = \frac{N\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left(N\sum x^2 - \left[\sum x\right]^2\right)\left(N\sum y2 - \left[\sum x\right]^2\right)}}$$

$$\text{we get} = \frac{12 \times 0.00755 - 0.1403 \times 0.1333}{\sqrt{\left(12 \times 0.01219 - [0.1403]^2\right)\left(12 \times 0.00952 - [0.1333]^2\right)}}$$

$$= \frac{0.0906 - 0.01870}{\sqrt{(0.14628 - 0.01968)(0.11424 - 0.017769)}}$$

$$= \frac{0.0719}{\sqrt{(0.1266)(0.09647)}}$$

$$= \frac{0.0719}{\sqrt{0.012213}}$$

$$= \frac{0.0719}{0.1105} = 0.650$$

Second Method (Using Excel)
Use the **CORREL** function to calculate the correlation between the two data sets. To calculate the correlation using Excel, follow these steps:

Step 1. Enter the returns on Sensex (X) and Nifty 50 (Y) in the Excel sheet.
Step 2. Use the **CORREL** function to calculate the correlation. Type the '=' (equal to) sign in the cell where the result is desired, say 'C15', and enter the function name **CORREL** and then select the cell range of numbers (data points of variable X), that is, **B2:B13** for **Array 1** and the cell range of numbers (data points of variable Y), that is, **C2:C13** for **Array 2**. The formula will be **=CORREL (B2:B13, C2:C13)** as shown in Figure 10.6.
Step 3. Press **Enter**. The result will be displayed in the cell as exhibited in Figure 10.7.

FIGURE 10.6

Step 1 in Calculation of Correlation

AVERAGE | =CORREL(B2:B13,C2:C13)

	A	B	C	D	E	F	G
1	Time	Sensex (X)	Nifty 50 (Y)				
2	January	0.52%	0.35%				
3	February	-1.07%	0.32%				
4	March	7.82%	4.58%				
5	April	0.93%	3.13%				
6	May	1.75%	-0.58%				
7	June	-0.80%	2.34%				
8	July	-4.86%	-2.47%				
9	August	-0.40%	-4.46%				
10	September	3.57%	1.41%				
11	October	3.78%	3.37%				
12	November	1.66%	4.23%				
13	December	1.13%	1.11%				
14							
15	CORRELATION COEFFICIENT	=CORREL(B2:B13,C2:C13)					
16							

FIGURE 10.7

Step 2 in Calculation of Correlation

C15 | =CORREL(B2:B13,C2:C13)

	A	B	C	D	E	F	G
1	Time	Sensex (X)	Nifty 50 (Y)				
2	January	0.52%	0.35%				
3	February	-1.07%	0.32%				
4	March	7.82%	4.58%				
5	April	0.93%	3.13%				
6	May	1.75%	-0.58%				
7	June	-0.80%	2.34%				
8	July	-4.86%	-2.47%				
9	August	-0.40%	-4.46%				
10	September	3.57%	1.41%				
11	October	3.78%	3.37%				
12	November	1.66%	4.23%				
13	December	1.13%	1.11%				
14							
15	CORRELATION COEFFICIENT		0.65021992				
16							

Third Method (Using Different Function in Excel)

Use the **PEARSON** function to calculate the *Pearson product–moment correlation coefficient* between the two data sets. The steps are as follows:

Step 1. Enter the returns on Sensex (X) and Nifty 50 (Y) in the Excel sheet.
Step 2. Use the **PEARSON** function to calculate the correlation.
Type the '=' (equal to) sign in the cell where the result is desired, say 'C15', and enter the function name **PEARSON** and then select the cell range of numbers (data points of variable X), that is, **B2:B13** for **Array 1** and the cell range of numbers (data points of variable Y), that is, **C2:C13** for **Array 2**.
The formula will be **=PEARSON (B2:B13, C2:C13)** as shown in Figure 10.8.
Step 3. Press **Enter** for the results in the cell (Figure 10.9).

Fourth Method (Using Data Analysis in Excel)

There is another method to compute the correlation coefficient using Excel by clicking the option of **Data Analysis**. To calculate the correlation using **Data Analysis**, follow the below-mentioned steps:

Step 1. Enter the returns on Sensex (X) and Nifty 50 (Y) in the Excel sheet.
Step 2. Go to **Data** tab, click on the option of **Data Analysis** in the **Analysis** group on the right side.

FIGURE 10.8

Step 1 in Calculation of Pearson Correlation Coefficient

AVERAGE | =PEARSON(B2:B13,C2:C13)

	A	B	C	D	E	F	G
1	Time	Sensex (X)	Nifty 50 (Y)				
2	January	0.52%	0.35%				
3	February	-1.07%	0.32%				
4	March	7.82%	4.58%				
5	April	0.93%	3.13%				
6	May	1.75%	-0.58%				
7	June	-0.80%	2.34%				
8	July	-4.86%	-2.47%				
9	August	-0.40%	-4.46%				
10	September	3.57%	1.41%				
11	October	3.78%	3.37%				
12	November	1.66%	4.23%				
13	December	1.13%	1.11%				
14							
15	PEARSON CORRELATION COEFFICIENT	=PEARSON(B2:B13,C2:C13)					
16							

FIGURE 10.9

Step 2 in Calculation of Pearson Correlation Coefficient

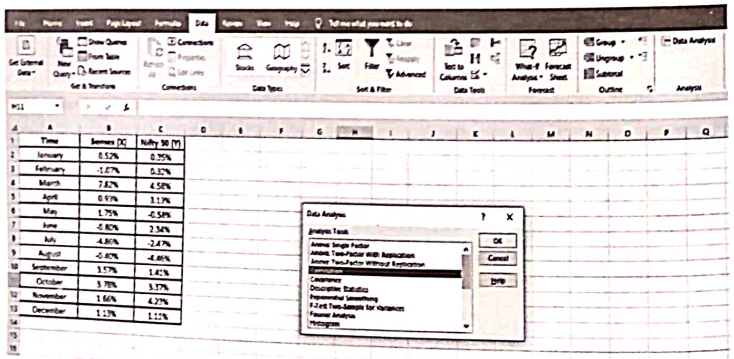
C15 =PEARSON(B2:B13,C2:C13)

	A	B	C	D	E	F	G
1	Time	Sensex (X)	Nifty 50 (Y)				
2	January	0.52%	0.35%				
3	February	-1.07%	0.32%				
4	March	7.82%	4.58%				
5	April	0.93%	3.13%				
6	May	1.75%	-0.58%				
7	June	-0.80%	2.34%				
8	July	-4.86%	-2.47%				
9	August	-0.40%	-4.46%				
10	September	3.57%	1.41%				
11	October	3.78%	3.37%				
12	November	1.66%	4.23%				
13	December	1.13%	1.11%				
14							
15	PEARSON CORRELATION COEFFICIENT		0.65021992				
16							

Step 3. A dialogue box will appear. Select the option of **Correlation** from the list and click **OK** (Figure 10.10).

FIGURE 10.10

Step 1 in Calculation of Correlation using Data Analysis



Step 4. A **Correlation** dialogue box will appear. Fill the following details:

- **Input Range:** The cell range of numbers for which the correlation is to be calculated, that is, B1:C13.
- **Labels in the first row:** Check on **Labels in the first row**. Remember if you choose the label/name of the variables, that is, Sensex and Nifty 50 in the input range, then check the box **Labels in the first row.**
- **Output Range:** The cell reference where the result is desired, say E1 as shown in Figure 10.11.

Step 5. Click **OK.** The result will be displayed in the cell (Figure 10.12).

FIGURE 10.11

Step 2 in Calculation of Correlation using Data Analysis

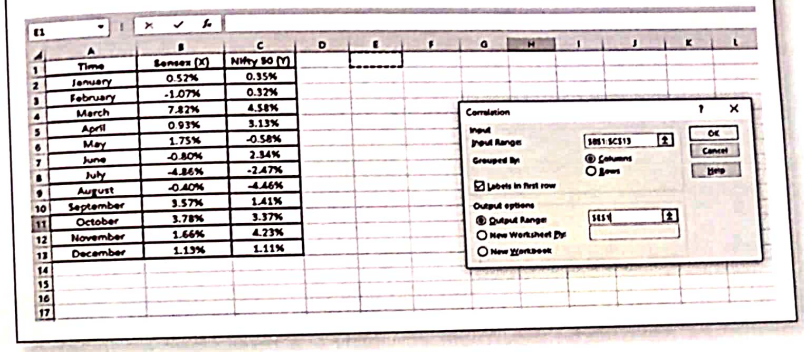


FIGURE 10.12

Step 3 in Calculation of Correlation using Data Analysis

E1

	A	B	C	D	E	F	G	H
1	Time	Sensex (X)	Nifty 50 (Y)			Sensex (X)	Nifty 50 (Y)	
2	January	0.52%	0.35%		Sensex (X)	1		
3	February	-1.07%	0.32%		Nifty 50 (Y)	0.650219918	1	
4	March	7.82%	4.58%					
5	April	0.93%	3.13%					
6	May	1.75%	-0.58%					
7	June	-0.80%	2.34%					
8	July	-4.86%	-2.47%					
9	August	-0.40%	-4.46%					
10	September	3.57%	1.41%					
11	October	3.78%	3.37%					
12	November	1.66%	4.23%					
13	December	1.13%	1.11%					
14								
15								
16								

Interpretation of Results

We can see that the calculated correlation coefficient is the same irrespective of the method chosen. The calculated estimate of **+0.650** suggests:

- The + sign describes a *direct or positive linear relationship* between the variables, that is, returns on Sensex (*X*) and returns on Nifty (*Y*). It signifies that the returns from both the indices tend to *move together in the same direction* either upward or downward.
- Correlation coefficient of *0.650 (r > 0.5) indicates strong positive correlation between the data sets.* It means that when the returns on one index *increases* by 100 per cent, then the returns on another index *also increases* but by lesser percentage, that is, by 65 per cent, or when the return on one index *decreases* by 100 per cent, then the returns on other also decreases but by lesser percentage, that is, 65 per cent.

We can also get an idea of the relationship between two variables with the help of a scatter plot in Excel, the preparation of which has already been explained earlier (Chapter 4). The *direction of the data points* helps to determine in which direction the variables change together or covary.

- *A positive or direct linear relationship* is indicated by an *upward sloping straight line.* The two variables tend to move in the same direction.
- *A negative or inverse linear relationship* is indicated by a *downward sloping straight line.* The two variables tend to move away or in the opposite direction from each other.

The *degree of the closeness of the data points* helps us to examine the strength of the relationship between variables, which is given below:

- If the data points lie on the straight line, there is a perfect correlation between the variables.
- If the data points are close or near to the line, then there is a strong relationship between the variables.
- If the data points are widely scattered or dispersed around the line, there is a low correlation between the variables.
- If the data points are widely scattered, and there is no upward rising or downward sloping line, the data sets are said to have no linear relationship.

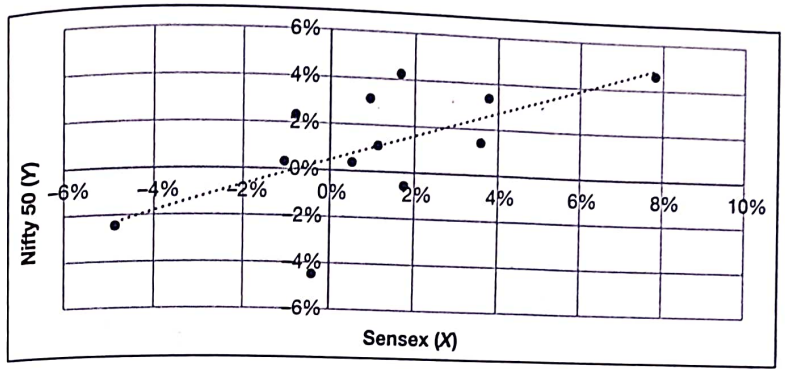
The following scatter plot, as in Figure 10.13, illustrates the direction as well as the strength of the relationship between the monthly rates of return on Sensex (*X*) and Nifty 50 (*Y*).

A perusal of the fitted trendline suggests the following:

- The shape or the pattern of the data points follows a linear pattern.
- The *trendline* which exhibits the nature of the trend that exists between the two variables is upward sloping to the right indicating a direct relationship or positive correlation
- It shows that when returns on one index move in a certain direction, the returns on the other index tend to move in the same direction.
- The data points scattered around the upward sloping line that shows the variables have a strong positive relationship.

FIGURE 10.13

Scatter Plot Showing Linear Relationship between the Monthly Rate of Returns on Sensex (X) and Nifty 50 (Y)



Check Your Understanding

Does the strong positive correlation mean that the increase in the returns in Nifty 50 causes the increase in the returns of Sensex or the increase in returns of Sensex causes the increase in the returns of Nifty 50?

Ans. No. Correlation does not imply any cause–effect relationship.

10.4. SPEARMAN'S COEFFICIENT OF CORRELATION

The Spearman correlation coefficient, unlike Pearson's correlation coefficient method, is calculated for the data measured on ordinal or rank scale. Denoted by 'r_s' or Greek letter rho 'ρ', this method measures both the direction and strength of the monotonic relationship between the rankings in the two data sets. It is a non-parametric measure of correlation, giving a numerical value between +1 and –1. As discussed before,

- $r_s = +1$ shows a *perfect positive relationship*
- $r_s = -1$ shows a *perfect negative relationship*
- $r_s = 0$ *no correlation*
- The closer the value of the coefficient is to 1 (+ve or –ve), the higher is the similarity in rankings of the two data sets.
- The closer the value of the coefficient is to 0, the lower is the similarity in the rankings of the two data sets.

The steps to calculate the Spearman coefficient of correlation are as follows.

Step 1. The ranks are assigned to all the values in both the data sets in the specified manner as shown. The first rank (1st) is given either to the highest value in both the data sets or to the lowest value in both the data sets. The subsequent ranks are given

accordingly in the same manner. In an example of marks of seven students in an exam, the assigned ranks are shown as follows.

Student	*Scores*	*Rank*
A	80	2
B	70	3
C	40	5
D	60	4
E	10	7
F	20	6
G	90	1

However, if two or more observations are identical, and there is a tie for the same rank, then the rank assigned is the *mean of the ranks that they occupy*. In the example given below, ranks are assigned to the scores of seven students in an exam out of which two students obtained the same score of 80.

Student	*Scores*	*Rank*
A	80	2.5
B	70	4
C	80	2.5
D	60	6
E	50	7
F	65	5
G	90	1

In this case, there is a tie between Student A and Student C for the second rank, with a score of 80 in both cases. Here, each of them with identical scores will be assigned the mean of the ranks $\frac{(2nd+3rd)}{2} = 2.5$th rank. Now, the next highest score will be assigned the 4th rank, as Students A and C together occupy the mean of 2nd and 3rd rank.

Step 2. The difference between the ranks of the two values in each data set is then calculated, denoted by d.

Step 3. The square of the differences (d^2) is computed, and then, the sum of all the squared values is obtained.

Step 4. The Spearman coefficient of correlation is calculated using the following formula:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Here,

d_i = Difference between ranks of the ith pair of the two variables,

$\sum d_i^2$ = Sum of the squares of the difference between ranks,

n = Number of observations in each data set

10.4.1. Illustration

Final year students of a college got placed in different companies. The table below summarizes the marks of students (in percentages) at the graduate level and their first salary (per annum). Calculate the Spearman coefficient of correlation to find out the association between the two variables.

Result in % (Graduation)	*First Salary (per Annum; ₹)*
76	5,65,000
44	5,70,000
80	6,50,000
87	9,20,000
90	8,60,000
92	7,80,000
66	6,30,000
57	4,00,000
97	8,20,000
88	9,50,000

First Method (Direct Method)

Result (Graduation; X)	*First Salary (Y; ₹)*	*Rank (X)*	*Rank (Y)*	*d*	d^2
76	565,000	7	9	−2	4
44	570,000	10	8	2	4
80	650,000	6	6	0	0
87	920,000	5	2	3	9
90	860,000	3	3	0	0
92	780,000	2	5	−3	9
66	630,000	8	7	1	1
57	400,000	9	10	−1	1
97	820,000	1	4	−3	9
88	950,000	4	1	3	9
					46

$$\sum d^2 = 46,\ n = 10$$

Substituting these values in the formula below:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$$r_s = 1 - \frac{6 \times 46}{10(10^2-1)}$$

$$= 1 - \frac{276}{990}$$

$$= 1 - 0.27878$$

$$= 0.7212$$

Second Method (Using Excel)

There is no direct or inbuilt function to calculate the Spearman coefficient of correlation using Excel. To compute the coefficient, we will first rank the values in each data set and then calculate the simple correlation between the ranks of the data. To do this, we will use the **RANK** and **CORREL** function to calculate the Spearman correlation coefficient between the two data sets.

To calculate the spearman coefficient, follow these steps:

Step 1. Enter the result score and amount of the first salary in columns in the Excel sheet.

Step 2. Use the **RANK** function to rank the data.

1. To rank the variable *X* (scores):
   a. Type the '=' (equal to) sign in the cell where the result is desired, say 'C2', and enter the function name **RANK**.
   b. Select the cell number of the first value in variable *X*, that is, A2, and then select the cell range of numbers (all data points of variable *X*), that is, absolute reference, **$A$2:$A$11** as a reference of the list of numbers against which the rank is to be assigned.
   c. Type **'0'** to assign the first rank to the highest value, that is, to rank in descending order.
   The formula will be **=RANK (A2, $A$2:$A$11, 0)**
   d. Press **Enter** (see Figures 10.14 and 10.15).
   The result will appear in cell C2. Then click on the fill handle (the small black square in the bottom right corner when you select the cell) and drag it vertically to automatically rank the other values.
2. Similarly, to rank the variable *Y* (first salary):
   a. Type the '=' (equal to) sign in the cell where the result is desired, say 'D2', and enter the function name **RANK**.
   b. Select the cell number of the first value in variable *X*, that is, B2, and then select the cell range of numbers (all data points of variable *X*), that is, absolute reference **$B$2:$B$11** as a reference of the list of numbers against which the rank is to be assigned.
   c. Type **0** to assign the first rank to the highest value, that is, to rank in descending order.
   The formula will be **=RANK (B2, $B$2:$B$11, 0)**
   d. Press **Enter** (see Figures 10.14 and 10.15).
   The result will appear in cell D2. Then click on the fill handle (the small black square in the bottom right corner when you select the cell) and drag it vertically to automatically rank for the other values.

Step 3. Finally, to compute the Spearman correlation, follow the mentioned steps:

a. Type the '=' (equal to) sign in the cell where the result is desired, say 'C15', and enter the function name **CORREL**.
b. Then select the cell range of numbers (data points of variable X), that is, **C2:C13** for **Array 1** and the cell range of numbers (data points of variable Y), that is, **D2:D13** for **Array 2**. The formula will be **=CORREL (C2:C13, D2:D13)**.

Step 4. Press **Enter**. The result will be displayed in the cell (see Figures 10.16 and 10.17).

Important Points

1. We need to use absolute cell reference **$A$2:$A$11** in our formula so that it will not change when the formula is copied down to cells below as the range **A2:A11** (reference of the list of numbers against which the ranks is assigned) *remains constant* for all the values.
2. The 'order' argument (to rank in either descending or ascending order) in the formula is optional.
   - Choose 0 or leave the argument empty to rank the numbers in descending order (largest number gets a rank of 1)
   - Choose 1 to rank the numbers in ascending order (lowest number gets a rank of 1)
3. If two or more observations are identical, then there is a tie for the same rank.
   We use the **RANK.AVG** function to assign the average rank to more than one value having the same rank.

FIGURE 10.14

Step 1 in Calculation of Spearman Correlation Coefficient

C2 | =RANK(A2,$A$2:$A$11,0)

	A	B	C	D
1	Result (Graduation)	First salary	Rank (X)	Rank (Y)
2	76	565000	=RANK(A2,$A$2:$A$11,0)	=RANK(B2,$B$2:$B$11,0)
3	44	570000		
4	80	650000		
5	87	920000		
6	90	860000		
7	92	780000		
8	66	630000		
9	57	400000		
10	97	820000		
11	88	950000		
12				

Interpretation of Results

The correlation coefficient of +0.7212 suggests

- There is a *direct or positive linear relationship* between the ranks of the variables. The result score and amount of the first salary tend to *move together in the same direction* either upward or downward.
- Correlation coefficient of *0.7212 (r_s > 0.5) suggests strong positive correlation between the data sets.*
- Therefore, we can conclude that the student who scores a high percentage is likely to get a higher amount of the first salary.

FIGURE 10.15

Step 2 in Calculation of Spearman Correlation Coefficient

C2 | =RANK(A2,$A$2:$A$11,0)

	A	B	C	D	E	F	G
1	Result (Graduation)	First salary	Rank (X)	Rank (Y)			
2	76	₹ 5,65,000	7	9			
3	44	₹ 5,70,000	10	8			
4	80	₹ 6,50,000	6	6			
5	87	₹ 9,20,000	5	2			
6	90	₹ 8,60,000	3	3			
7	92	₹ 7,80,000	2	5			
8	66	₹ 6,30,000	8	7			
9	57	₹ 4,00,000	9	10			
10	97	₹ 8,20,000	1	4			
11	88	₹ 9,50,000	4	1			
12							

FIGURE 10.16

Step 3 in Calculation of Spearman Correlation Coefficient

AVERAGE | =CORREL(C2:C11,D2:D11)

	A	B	C	D	E	F	G
1	Result (Graduation)	First salary	Rank (X)	Rank (Y)			
2	76	₹ 5,65,000	7	9			
3	44	₹ 5,70,000	10	8			
4	80	₹ 6,50,000	6	6			
5	87	₹ 9,20,000	5	2			
6	90	₹ 8,60,000	3	3			
7	92	₹ 7,80,000	2	5			
8	66	₹ 6,30,000	8	7			
9	57	₹ 4,00,000	9	10			
10	97	₹ 8,20,000	1	4			
11	88	₹ 9,50,000	4	1			
12							
13	CORRELATION (between the ranks)		=CORREL(C2:C11,D2:D11)				
14							

FIGURE 10.17

Step 4 in Calculation of Spearman Correlation Coefficient

D13 | =CORREL(C2:C11,D2:D11)

	A	B	C	D	E	F	G
1	Result (Graduation)	First salary	Rank (X)	Rank (Y)			
2	76	₹ 5,65,000	7	9			
3	44	₹ 5,70,000	10	8			
4	80	₹ 6,50,000	6	6			
5	87	₹ 9,20,000	5	2			
6	90	₹ 8,60,000	3	3			
7	92	₹ 7,80,000	2	5			
8	66	₹ 6,30,000	8	7			
9	57	₹ 4,00,000	9	10			
10	97	₹ 8,20,000	1	4			
11	88	₹ 9,50,000	4	1			
12							
13	CORRELATION (between the ranks)			0.72121212			
14							

So far, the discussion has been centred around *linear relationships*. Correlation can, on the other hand, be negligible in two cases. First, when the variables are *independent* of each other and second, when there is *no linear relationship* between the two variables. Suppose the two variables are independent of each other, and when one does not change with change in another variable, the correlation will be zero (0). The correlation coefficient having zero (0) value does not imply that the variables are independent. We can only conclude that there is no linear relationship between them. Nonetheless, the two variables may have a strong *non-linear relationship*.

A nonlinear relationship refers to the relationship between the variables where if one variable changes, the other variable also changes but at varying rates. When the data points of two variables having a non-linear relationship are plotted on a graph and connected, it does not form a straight line rather a curved line. Many relationships in economics are nonlinear relationships. Few examples of nonlinear relationships are quadratic polynomial, cubic polynomial or logarithm relationships.

Consider the following hypothetical illustration, exhibiting the data related to the hours of work, output and the average cost per unit of production in a firm, assuming a fixed cost of ₹500 and variable cost of ₹1 per unit. Note that both fixed and variable costs rise after 18 hours. The data indicates that as hours of work increase, output increases uniformly, but the average cost does not come down uniformly as fixed costs get distributed over progressively.

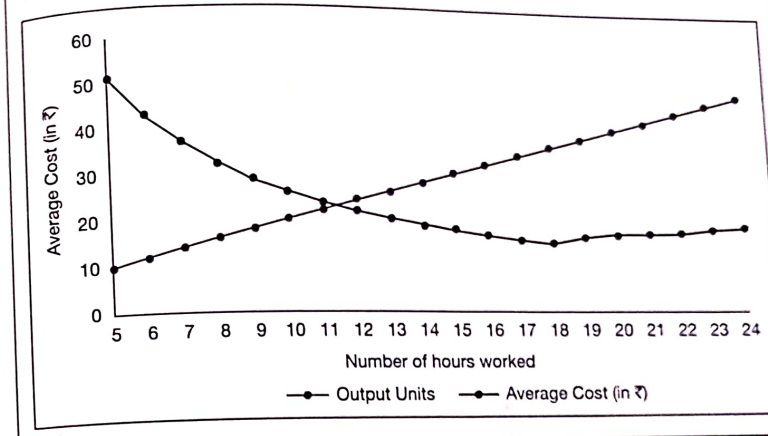
Hours of Work	*Output*	*Cost*
5	10	51
6	12	42.67
7	14	36.71
8	16	32.25
9	18	28.78
10	20	26.00
11	22	23.73
12	24	21.83
13	26	20.23
14	28	18.86
15	30	17.67
16	32	16.63
17	34	15.71
18	36	14.89
19	38	16.47
20	40	17.00
21	42	17.20
22	44	17.50
23	46	18.00
24	48	18.50

A systematic approach before establishing any correlation between two variables is to plot the data points and have a graphical look of the data. For instance, Figure 10.18 endeavours to establish a correlation between the number of hours worked and average cost. It may be seen in the graph plotted below a non-linear relationship between the variables. However, calculating the correlation coefficient between the two variables yields r = –0.8225. *It is important to note here that the correlation coefficient value on the non-linear relationship is not meaningful. Correlation itself does not reveal anything on either the linear or non-linear relationship. Although if a relationship is linear, then correlation measures the strength of the relationship.*

To recap, this chapter principally discussed some of the techniques that attempt to quantify the linear relationship between the variables. Although covariance analysis advise us on the directional relationship between two variables, insight into the strength of relationships is important in statistical analysis. The study of the correlation coefficient helps to determine the strength of the relationship between variables. In particular, we focused on Karl Pearson's method and Spearman's method to estimate the coefficient of correlation. It was clear that these measures in no way suggest any cause and effect relationship. To establish any relationship among variables and along with finding out the change in a (dependent) variable caused by the change in other (independent) variables, we will study regression analysis in the next chapter.

FIGURE 10.18

Relationship between Hours of Work and Average Cost



PRACTICE QUESTIONS

1. Zero correlation implies that *X* and *Y* variables are independent. Comment.
2. The table below summarizes the export values and import values of India (in the US $ million for 10 years starting from the year 2009–2010 to 2018–2019. Calculate the covariance and Pearson's correlation coefficient and comment on the relationship.

Time Period (2018)	*Export*	*Import*
2009–2010	178,751	288,373
2010–2011	249,816	369,769
2011–2012	305,764	489,319
2012–2013	300,401	490,737
2013–2014	314,405	450,200
2014–2015	310,338	448,033
2015–2016	262,291	381,008
2016–2017	275,852	384,357
2017–2018	303,526	465,581
2018–2019	331,020	507,436

Source: Department of Commerce, Ministry of Commerce and Industry, Government of India. Annual Report 2018-19. Retrieved from https://commerce.gov.in/writereaddata/uploadedfile/MOC_637036322182074251_Annual%20Report%202018-19%20English.pdf

Ans. Covariance for population = **2,609,286,231.88** (positive or direct relationship)
Covariance for sample = **2,899,206,924.31**
Correlation is 0.9416, very strong positive relation.

3. Calculate the Karl Pearson coefficient of correlation between the net profits (in crores) and non-performing assets (in crores) of State bank of India for 10 years starting from the year 2009–2010 to 2018–2019.

Financial Year	*Net Profits*	*Non-performing Assets*
2009–2010	12,020.54	10,870.17
2010–2011	11,179.94	12,346.89
2011–2012	15,829.45	15,828.85
2012–2013	18,322.99	21,956.48
2013–2014	14,489.47	31,096.07
2014–2015	17,517.37	27,590.58
2015–2016	12,743.29	55,807.02
2016–2017	–390.67	58,277.38
2017–2018	–4,187.41	110,854.7
2018–2019	3,069.07	65,894.74

Source: Data retrieved from https://www.moneycontrol.com/financials/statebankindia/balance-sheetVI/sbi

Ans. Pearson's correlation coefficient = –0.8294 (strong negative correlation).

4. The exam scores of 15 students in economics and statistics in a class are given in the table below. Calculate the Spearman correlation coefficient to estimate the relationship between the scores.

Student ID	*Economics*	*Statistics*
1	94	88
2	95	60
3	35	84
4	75	100
5	100	80
6	46	90
7	93	70
8	81	75
9	60	84
10	45	54
11	92	77
12	76	86
13	44	78
14	87	65
15	84	52

Ans. Spearman correlation coefficient = –0.2478 (low negative correlation), that is, the student scoring high in the economics exam is less likely to score high in statistics exam.

5. The following are the marks of 8 students in the physics and chemistry test. Find Karl Pearson and Spearman coefficient of correlation.
Ans. Pearson correlation coefficient = +0.5954
Spearman correlation coefficient = +0.63474

Based on correlation coefficients, it can be concluded that an increase in 10 marks in one subject will lead to an increase in 5.9 and 6.3 marks respectively in other subject, according to Pearson and Spearman estimations.

Marks in Physics	*Marks in Chemistry*
40	30
65	80
70	10
35	30
90	90
45	50
55	70
80	85

11

CHAPTER

Regression Analysis

LEARNING OBJECTIVES

After reading this chapter, the readers will be able to understand

- The concept of regression analysis
- Developing a linear regression model
- Estimation and interpretation of the regression coefficients
- Predicting the values of the dependent variable
- Derivation and interpretation of the Coefficient of Determination (R^2)
- Least squares method to obtain the best fitting line
- Performing the regression analysis using Data Analysis in Excel

In the previous chapter, we studied the measures of association that help estimate the degree of relationship between two data sets. It was pointed out that correlation analysis does not suggest any causal relationship between the variables. In regression analysis, the theory postulates a relationship between the variables, and analysis is performed to estimate the influence of one or more variables on the other.

Regression analysis can be linear or non-linear based on the nature of the relationship between variables. An important assumption of linear regression is that the relationship between the variables must be linear, that is, it follows a straight line relationship. On the other hand, a non-linear regression model studies variables having no linear relationship, that is, the curve of regression is not a straight line. The present chapter focuses on the linear regression model throughout the analysis. In addition to understanding the concept of regression analysis, its relevance and applicability, we will also be learning how to develop, estimate and interpret a linear regression model. Based on this, we will make predictions and determine the fit of the model.

11.1. REGRESSION ANALYSIS

Regression analysis refers to a statistical technique that quantifies the relationship between the variables and predicts the value of one variable from another set of variables. The variable which predicts the value of the other variable is known as the *independent variable* or the *explanatory variable* or the *predictor* or simply the *X variable*. Likewise, the variable whose value is predicted is known as the *dependent variable* or *variable of interest* or the *Y variable*. Linear regression analysis essentially models the relationship between a dependent and an independent variable by fitting a linear equation. Let us look at the steps involved in developing a complete linear regression model.

11.2. DEVELOPING A LINEAR REGRESSION MODEL

To create a regression model, we first identify and establish the relationship between the dependent and independent variables. The *simple* linear regression equation can be written in the form of:

$$Y = \beta_0 + \beta_1 X$$

Here,

Y = Dependent variable
X = Independent variable
β_0 = Y-intercept
β_1 = Slope coefficient of Y with respect to X

Correspondingly, the *multiple* linear regression equation for 'n' number of independent variables can be written in the form of:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots + \beta_n X_n$$

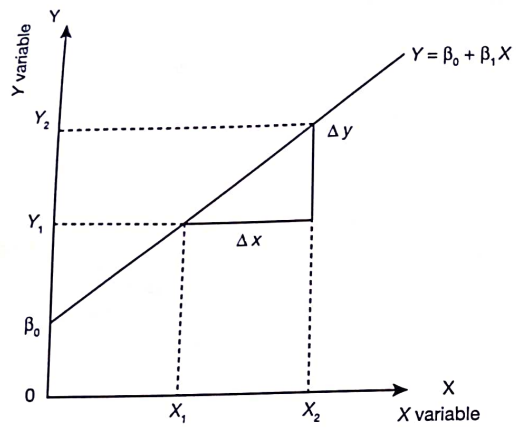
Here,

Y = Dependent variable
X_i = *ith* independent variable, $i \in (1, 2\ldots n)$
β_0 = Y intercept
β_i = Slope coefficient of Y on X_i

The regression coefficient, denoted as β_i, is known as the regression coefficient of Y on X (when the value of variable Y depends on X). The β_0, known as the constant or the Y-intercept, estimates the value of Y when the value of X is zero. The magnitude of β_i signifies the amount of change in Y due to a change in X_i by one unit, provided all the other explanatory variables are held constant. Together β_0 and β_1 are the parameters of the regression model that describe the relationship between dependent and independent variables mathematically. The regression model can be shown graphically as in Figure 11.1, where the Y-axis (vertical axis) depicts the dependent variable and the X-axis (horizontal axis) shows the independent variable.

FIGURE 11.1

Graphical Representation of Regression Line



In this figure, the upward-sloping straight line, from the lower left to the upper right, is the regression line represented by the equation of $Y = \beta_0 + \beta_1 X$. β_0 is the intercept of the line, crossing from the Y-axis which indicates the value of the dependent variable when the explanatory variable is zero. β_1 is the *slope* of the regression line which measures *the rate of change in the Y corresponding to the unit change in the value of X,* that is:

$$\frac{\Delta y}{\Delta x} \text{ or } \frac{Y_2 - Y_1}{X_2 - X_1}$$

The slope coefficients may either be *positive, negative or zero,* indicating the direction of change. The positive slope means for any given change in the value of *X*, the change in the value of *Y* is in the same direction, that is, a positive change in *X* is accompanied by a positive change in *Y* and a negative change in *X* is followed by a negative change in *Y* (Figure 11.2). The negative slope implies that for any given change in the value of X, the change in the value of *Y* is in the opposite direction, an increase in *X* accompanied by a decrease in *Y* and a decrease in *X* by an increase in *Y* (Figure 11.3). However, if the slope is 0, it means that for any value of *X*, the value of *Y* remains constant. It is represented by a straight line horizontal to the X-axis as shown in Figure 11.4.

FIGURE 11.2

Regression Line with Positive Slope ($\beta_1 > 0$)

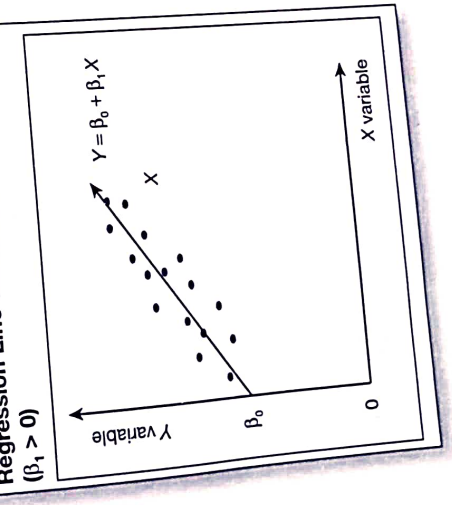


FIGURE 11.3

Regression Line with Negative Slope ($\beta_1 < 0$)

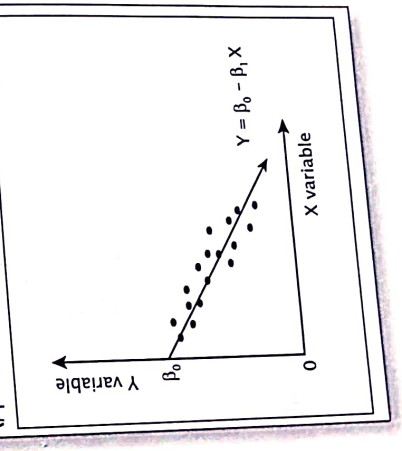
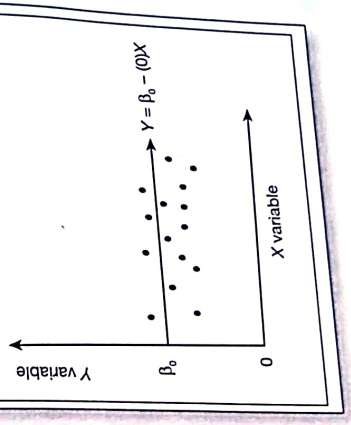


FIGURE 11.4

Regression Line with Constant Slope ($\beta_1 = 0$)



11.3. ESTIMATION OF THE COEFFICIENTS IN THE REGRESSION MODEL

The second step is about estimating the values of the coefficients in the regression model. For a dependent (*Y*) and independent variable (*X*), we have already developed a regression model and established the relationship between them in the form of an equation, that is, $Y = \beta_0 + \beta_1 X$. The formula[1] for the slope coefficient is given by the ratio of the *covariance* of *X* and *Y* to the *variance* of *X*, that is:

$$\beta_1 = \frac{\mathrm{Cov}(X,Y)}{\sigma_x^2}$$

Or

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{X})^2}$$

$$\beta_1 = \frac{\sum xiyi}{\sum xi^2}$$

[1] Summing up the given linear regression equation, we get:

$$\sum Y = na + b\sum X \quad (1)$$

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

Solving Equations (1) and (2) simultaneously for a and b, we get:

$$b = \frac{(\sum xy) - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} \quad (3)$$

$$= \frac{n(\sum xy) - \sum x \times \sum y}{n\sum x^2 - (\sum x)^2} \quad (4)$$

It may be recalled that the covariance between X and Y is given by:

$$\mathrm{Cov}(X,Y) = \frac{\sum XY}{n} - \left(\frac{\sum X}{n}\right)\left(\frac{\sum Y}{n}\right) \quad (5)$$

Further, the variance of X is given by

$$\sigma_x^2 = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2 \quad (6)$$

From 3, 5 and 6, we get:

$$b = \frac{\mathrm{Cov}(X,Y)}{\sigma_x^2}$$

Here,

$\sum$ = Sigma/operator for summation
$\bar{X}$ = Arithmetic mean of independent variable
$\bar{Y}$ = Arithmetic mean of dependent variable
x_i = Value of independent variable
y_i = Value of the dependent variable

$$\text{Alternatively, } \beta_1 = \frac{n(\sum xy) - \sum x \times \sum y}{n\sum x^2 - (\sum x)^2}$$

or

$$\beta_1 = r \times \frac{\sigma_y}{\sigma_x}$$

Here,

r = Correlation coefficient
σ_x = Standard deviation of independent variable
σ_y = Standard deviation of the dependent variable

The value of parameter β_0 has no economic interpretation but can be calculated as follows:

$$\beta_0 = \bar{Y} - \beta_1\bar{X}$$

Illustration

A company incurs promotion expenses to boost their revenue. The table below summarizes the monthly promotional expenses and total revenue of the company for the last two years. The sales manager wants to examine how the total revenue of the company depends upon the promotion expenses. Develop a regression model and estimate the regression coefficients.

Months	*Promotion Expenses (X; in Crores)*	*Revenue (Y; in Crores)*
1	400	5,500
2	410	5,540
3	430	5,550
4	450	5,600
5	480	5,600
6	510	5,600
7	530	5,700
8	540	5,700
9	550	5,800
10	560	5,800
11	560	5,900
12	570	5,900
13	570	5,950
14	580	6,000

Months	Promotion Expenses (X; in Crores)	Revenue (Y; in Crores)
15	620	6,040
16	630	6,050
17	640	6,100
18	640	6,200
19	660	6,300
20	590	6,600
21	630	6,850
22	640	6,800
23	680	7,000
24	690	7,500

First Method (Direct Method; in Crores)

Months	Promotion Expenses (X)	Revenue (Y)	$x-\bar{X}$	$(x-\bar{X})^2$	$y-\bar{Y}$	$(x-\bar{X})\times(y-\bar{Y})$
1	400	5,500	–165	27,225	–565.83	93,362.50
2	410	5,540	–155	24,025	–525.83	81,504.17
3	430	5,550	–135	18,225	–515.83	69,637.50
4	450	5,600	–115	13,225	–465.83	53,570.83
5	480	5,600	–85	7,225	–465.83	39,595.83
6	510	5,600	–55	3,025	–465.83	25,620.83
7	530	5,700	–35	1,225	–365.83	12,804.17
8	540	5,700	–25	625	–365.83	9,145.83
9	550	5,800	–15	225	–265.83	3,987.50
10	560	5,800	–5	25	–265.83	1,329.17
11	560	5,900	–5	25	–165.83	829.17
12	570	5,900	5	25	–165.83	–829.17
13	570	5,950	5	25	–115.83	–579.17
14	580	6,000	15	225	–65.83	–987.50
15	620	6,040	55	3,025	–25.83	–1,420.83
16	630	6,050	65	4,225	–15.83	–1,029.17
17	640	6,100	75	5,625	34.17	2,562.50
18	640	6,200	75	5,625	134.17	10,062.50
19	660	6,300	95	9,025	234.17	22,245.83
20	590	6,600	25	625	534.17	13,354.17
21	630	6,850	65	4,225	784.17	50,970.83
22	640	6,800	75	5,625	734.17	55,062.50
23	680	7,000	115	13,225	934.17	107,429.17
24	690	7,500	125	15,625	1,434.17	179,270.83
Total	13,560	145,580		162,200		827,500

Mean of independent variable $(\bar{X}) = 13{,}560/24 = 565$
Mean of dependent variable $(\bar{Y}) = 145{,}580/24 = 6{,}065.8333$

We already know that $\beta_1 = \dfrac{\sum_{i=1}^{n}(x_i-\bar{X})(y_i-\bar{Y})}{\sum_{i=1}^{n}(x_i-\bar{X})^2}$

Substituting the values in the formula above, we get:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{X})(y_i-\bar{Y})}{\sum_{i=1}^{n}(x_i-\bar{X})^2}$$

$$= \frac{827{,}500}{162{,}200} = 5.101726264$$

Similarly, $\beta_0 = \bar{Y} - \beta_1\bar{X}$
$= 6{,}065.8333 - 5.101726264 \times 565$
$= 6{,}065.8333 - 2{,}882.47533916$
$= 3{,}183.357994$

As discussed previously, in a regression model, we first need to identify the dependent variable (*Y*) and independent variables. In the present example, revenue depends on promotional expenses, hence, revenue is a dependent variable and promotional expense is an independent or explanatory variable. In a mathematical or statistics parlance, revenue is a function of the promotional expenses. The linear equation is written as follows:

$$Y = \beta_0 + \beta_1 X$$

or

$$\text{Revenue} = \beta_0 + \beta_1 \times \text{Promotion expenses}$$

We rewrite the regression equation by substituting the value of beta coefficients, that is, β_1 and β_0 calculated above:

$$\text{Revenue} = 3{,}183.3579 + 5.101726 \times \text{Promotion expenses}$$

Second Method of Calculating Parameters (Using Excel)[2]

To estimate the regression model using Excel, we use the option of **Data Analysis** and follow the stated steps:

Step 1. Enter the promotion expenses (*X*) and revenue (*Y*) in the Excel sheet (Figure 11.5).
Step 2. Go to **Data** tab, click on the option of **Data Analysis** in the **Analysis** group on the right side (Figure 11.5).
Step 3. A dialogue box will appear. Select the option of **Regression** from the list and click **OK** (Figure 11.6).
Step 4. A **Regression** dialogue box will appear. Fill the following details (Figure 11.7).

- **Input Y Range:** The cell range of the values of the dependent variable (revenue) for which the regression model is developed, that is, **C1:C25**.

[2] Excel calculates parameters using ordinary least squares (OLS) method.

FIGURE 11.5

Step 1 in Calculating Regression Parameters

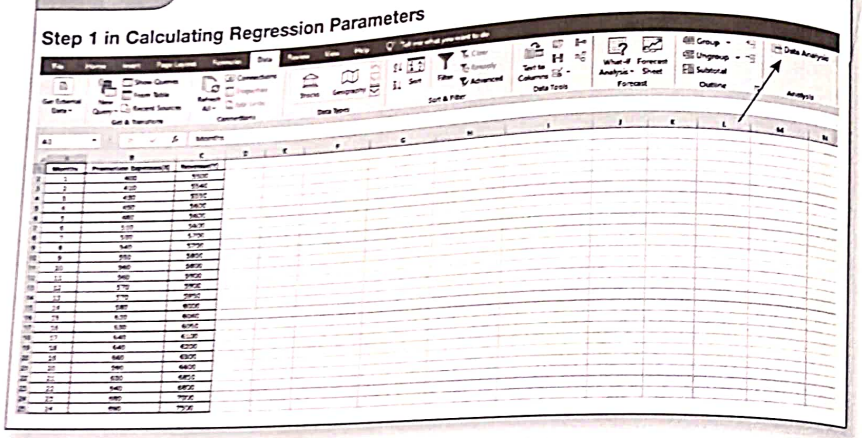


FIGURE 11.6

Step 2 in Calculating Regression Parameters

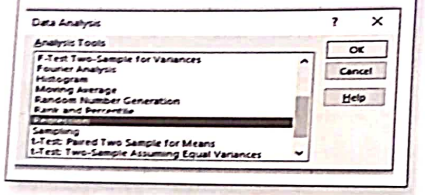
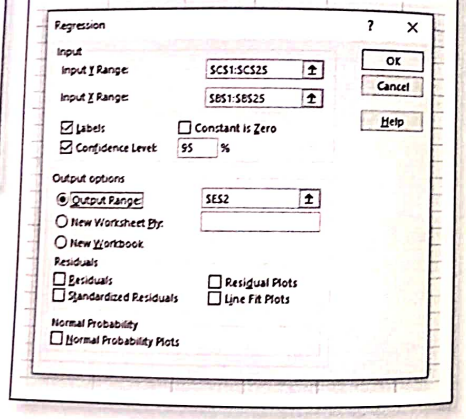


FIGURE 11.7

Step 3 in Calculating Regression Parameters



- **Input X Range:** The cell range of the values of the independent variable (promotion expenses), that is, **B1:B25**.
- **Check on Labels:** Remember if you choose the label/name of the variables, that is, promotion expenses and revenue in the **Input Range**, then check the box **Labels**.
- Check on the **Confidence Level** and set it at **95 per cent**, which is also the default value.
- **Output Range:** Under the output options, choose the cell reference where the result is desired in the existing worksheet, say **E2**.

Step 5. Click **OK**.

The regression result is displayed as a summary output table (Figure 11.8). The calculated values of parameters are identical in both the methods.

FIGURE 11.8

Step 4 in Calculating Regression Parameters

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.809705674
R Square	0.655623278
Adjusted R Square	0.63996979
Standard Error	317.483346
Observations	24

ANOVA	df	SS	MS	F	Significance F
Regression	1	4221678.483	4221678.483	41.88352807	1.6435E-06
Residual	22	2217504.85	100795.675		
Total	23	6439183.333			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	3183.357994	450.0836347	7.07281436	4.282E-07	2249.941666	4116.774322	2249.941666	4116.774322
Promotion Expenses(X)	5.101726264	0.788307256	6.471748456	1.6435E-06	3.466877077	6.73657545	3.466877077	6.73657545

11.4. INTERPRETATION OF THE ESTIMATED COEFFICIENTS IN THE REGRESSION MODEL

In a regression model, interpretation of the analysis plays a very important role. It is indeed one of the primary objectives. As mentioned before, β_0 and β_1 are the parameters with their respective meanings. Let's interpret the values of the coefficients estimated.

- $\beta_0 = 3{,}183.3579$ shows that the total revenue earned by the company is ₹3,183.3579 crore when the amount spent on promotion expenses is ₹0, that is, if the company does not spend any money on promotional activities for the product. Although from a technical point of view, the amount of revenue earned with no promotion expenses is correct. But in managerial terms, this value is unrealistic or irrational as the company is expected to spend some amount on its promotional activities or sales strategies. Therefore, we tend to ignore the value of β_0 in economic analysis.
- The positive sign of β_1 indicates that as the promotion expenses increase, the revenue of the company also tends to increase. The value of β_1 as 5.10 demonstrates that the promotional expenditure of one crore rupees (a unit of the independent variable) will boost the revenues by ₹5.10 crore (dependent variable). The statistical significance of these values, as well as of the model, will be examined in the ensuing sections.

11.5. FORECAST THE VALUES OF THE DEPENDENT VARIABLE

Regression analysis is mostly used to make predictions. To predict the values of the dependent variable, we bring in the calculated values of parameters along with the values of the independent variable in the regression model. The predicted value of Y is customarily denoted by $\hat{y}$ known as *y 'hat'*.

Illustration

Let us clarify it with the ongoing illustration. The sales manager of the company has allocated the budget for the promotion expenditure of the company for the next five months. The budget amount for such expenses is tabled below. The sales manager wishes to find out the predicted sale revenue for the next five months based on these promotion expenditures.

Months	*Budget for Promotion Expenses*
1	650
2	700
3	760
4	720
5	800

First Method

The regression equation under discussion is:

$$\text{Revenue} = \beta_0 + \beta_1 \times \text{Promotion expenses}$$

Substituting the values of beta coefficients, that is, β_1 and β_0 calculated above, we get:

$$\text{Revenue} = 3{,}183.3579 + 5.101726 \times \text{Promotion expenses}$$

The predicted revenue (Y) when the promotion expenses are ₹650 crore, ₹700 crore, ₹760 crore, ₹720 crore and ₹800 crore can be estimated as follows.

Months	*Budget for Promotions (X; in Crores)*	*Estimated Revenue (ŷ; in Crores)*
1	650	3,183.3579 + 5.101726 × 650 = 6,499.48
2	700	3,183.3579 + 5.101726 × 700 = 6,754.57
3	760	3,183.3579 + 5.101726 × 760 = 7,060.67
4	720	3,183.3579 + 5.101726 × 720 = 6,856.60
5	800	3,183.3579 + 5.101726 × 800 = 7,264.74

The estimated revenue is ₹6,499.48 crore, ₹6,754.57 crore, ₹7,060.67 crore, ₹7,264.74 crore for the 1st, 2nd, 3rd, 4th and 5th month respectively.

Second Method (Using Excel)

Step 1. Enter the budgeted promotion expenses (X) in the Excel sheet as shown in Figure 11.9.

Step 2. Type the '=' (equal to) sign in the desired cell where the estimated revenue (Y) is to be calculated, say G22, and then select the cell number containing the value of the *constant* in the summary output, that is, F18, then insert a plus sign (+) to add the product of the value of coefficient (in summary output), that is, F19 and value of promotion expense (X) in the first month, that is, H22.

The formula will be $F$18 + ($F$19 × F22).

Step 3. Press **Enter**. The result will appear in cell G22. Then click on the fill handle (the small black square in the bottom right corner when you select the cell) and drag it vertically to automatically calculate the predicted revenue for the next months (Figure 11.10).

FIGURE 11.9

Step 1 in Regression Analysis (Predicting the Dependent Variable)

AVERAGE | =$F$18+($F$19*F22)

	E	F	G	H	I	J	K	L	M
1									
2	SUMMARY OUTPUT								
3									
4		*Regression Statistics*							
5	Multiple R	0.809705674							
6	R Square	0.655623278							
7	Adjusted R Square	0.63996979							
8	Standard Error	317.483346							
9	Observations	24							
10									
11	ANOVA								
12		*df*	*SS*	*MS*	*F*	*Significance F*			
13	Regression	1	4221678.483	4221678.483	41.88352807	1.6435E-06			
14	Residual	22	2217504.85	100795.675					
15	Total	23	6439183.333						
16									
17		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
18	Intercept	3183.357994	450.0836347	7.07281436	4.282E-07	2249.941666	4116.774322	2249.941666	4116.774322
19	Promotion Expenses(X)	5.101726264	0.788307256	6.471748456	1.6435E-06	3.466877077	6.73657545	3.466877077	6.73657545
20									
21	Months	Promotion Expenses(X)	Revenue(Y)						
22	1	650	=$F$18+($F$19*F22)						
23	2	700							
24	3	760							
25	4	720							
26	5	800							

FIGURE 11.10

Step 2 in Regression Analysis (Predicting the Dependent Variable)

G22 | =$F$18+($F$19*F22)

	E	F	G	H	I	J	K	L	M
1									
2	SUMMARY OUTPUT								
3									
4		*Regression Statistics*							
5	Multiple R	0.809705674							
6	R Square	0.655623278							
7	Adjusted R Square	0.63996979							
8	Standard Error	317.483346							
9	Observations	24							
10									
11	ANOVA								
12		*df*	*SS*	*MS*	*F*	*Significance F*			
13	Regression	1	4221678.483	4221678.483	41.88352807	1.6435E-06			
14	Residual	22	2217504.85	100795.675					
15	Total	23	6439183.333						
16									
17		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
18	Intercept	3183.357994	450.0836347	7.07281436	4.282E-07	2249.941666	4116.774322	2249.941666	4116.774322
19	Promotion Expenses(X)	5.101726264	0.788307256	6.471748456	1.6435E-06	3.466877077	6.73657545	3.466877077	6.73657545
20									
21	Months	Promotion Expenses(X)	Revenue(Y)						
22	1	650	6499.48						
23	2	700	6754.57						
24	3	760	7060.67						
25	4	720	6856.60						
26	5	800	7264.74						

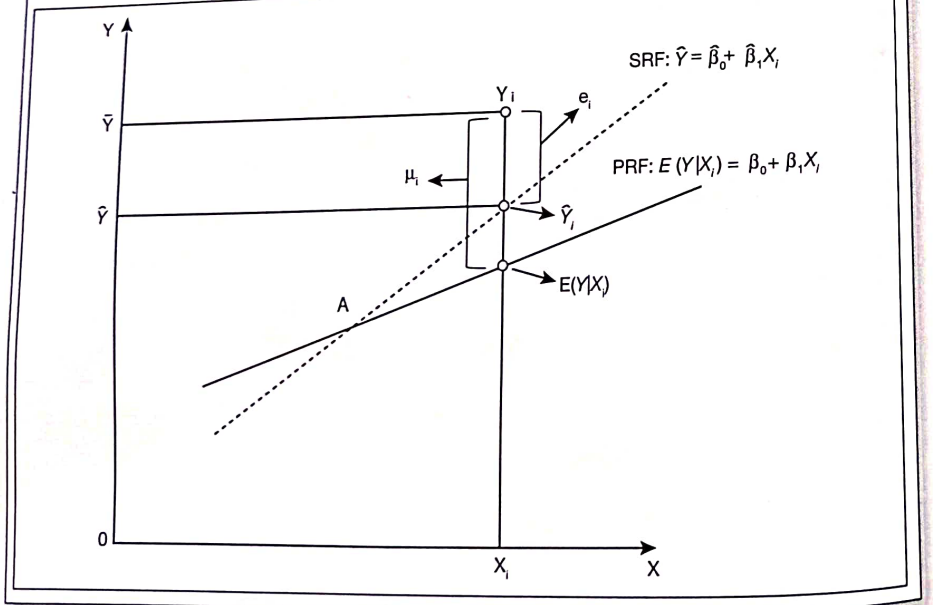
We need to use absolute cell reference in our formula so that it will not change when the formula is copied down to cells below as the *value of constant and slope coefficient* remains the same for all the values of Y variable.

The regression coefficients which are estimated from the regression model are based on the sample data and are just the estimates for the population parameters and not the true regression coefficients. Figure 11.11 shows the population and the sample regression functions, that is, PRF and SRF respectively. $\hat{Y}$ is the estimated value of the actual value of the Y variable, that is, E $(Y|X_i)$. We observe that the $\hat{Y}$ overestimates the true value of Y, that is, E $(Y|X_i)$ for a given value of X_i to the right of A as shown. Similarly, for any value of X_i to the left of Point A, the SRF underestimates the true PRF. Such a difference in the values is due to sampling fluctuations.

As we know, the SRF is an approximation of the PRF. We need to find out a mechanism to build the sample regression function so that the beta coefficients $(\hat{\beta}_0, \hat{\beta}_1)$ estimated from the regression model are as close as possible or near to the true value of the coefficients (β_0, β_1). The concern is addressed in the upcoming section when we talk about the OLS method or the least square criterion. The method reveals the procedures constructing the sample regression function which is almost a counterpart of or a good estimator of the population regression function.

FIGURE 11.11

Population (PRF) and Sample Regression Functions (SRF)



11.6. RESIDUALS OR ERRORS AND OLS METHOD

The foregoing section brings about regression equation $Y = \beta_0 + \beta_1 X$ that has attempted to predict the values of the dependent variable, given the values of the independent variable. However, these predicted values may be different from the *actual* or the *observed* values of dependent variable Y. Based on the estimated regression model, we can then calculate the difference between the actual values of Y and the values predicted from the regression model. This difference is known as *errors* or *residuals* in the regression model. The residual is denoted by Greek small letter epsilon, that is, 'ε'. Therefore, the regression equation showing the linear relationship between the dependent and independent variable can now be written as:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

or

$$\varepsilon = Y - (\beta_0 + \beta_1 X)$$

That is, errors/residuals = Actual value (Y) – Predicted value ($\hat{Y}$)

If the actual value is more than the predicted value, thus residual being positive, we can infer that the regression model is *underpredicting* the value of the dependent variable. Equivalently, if the actual value is less than the predicted values, residual being negative, we construe that the regression model is *overpredicting* the value of the dependent variable. However, the model predicts accurately if the error or residual values are absent.

Drawing a scatter plot helps identify the residuals. The first step involved is plotting the *actual values* of the variable Y (dependent variable) for the given value of the X variable (independent variable) on a scatter plot. The line joining different data points does not necessarily fall on a straight line but fundamentally follows a linear pattern.

In the second step, based on the regression model followed, a *regression line* or the *line of best fit* is drawn based on the calculated parameters for the given data set. The regression line obtained will seemingly be straight. It may conspicuously be observed that some of the actual values of the Y variable *fall above* the regression line and some *fall below* the regression line. The actual values plotted above the regression line (i.e., actual Y value > predicted Y value) are *positive residuals*. The residuals are negative if the actual values get plotted below the regression line (i.e., actual Y value < predicted Y value). Deviation of the dots (actual values) from the straight line (predicted or estimated values) are the residuals shown by a vertical bar as displayed in Figure 11.12. The larger is the deviation, the worse is the fit of the model to the data. Conversely, the closer the expected residual value is to 0, the better is the fit.

The error terms cancel out each other due to the positive and the negative signs. The objective of finding a best-fitted regression line is not to compensate the positive terms with the negative terms but to minimize the error terms. To avoid the sign problem of error terms, we use the method of the *least square criterion* popularly known as the method of *Ordinary Least Squares* (OLS), which determines the equation of the straight line of best fit. This method creates a linear regression model that *minimizes the sum of squared errors* (SS_E) to obtain the regression line that best fits through the given set of data points and predicts the movement in the variable of interest. To find out the extent of fit of the regression model, another measurement called the coefficient of determination is exercised.

11.7. COEFFICIENT OF DETERMINATION

The coefficient of determination is a statistical tool of the goodness of fit of a regression model to the data. The coefficient of determination figures out the proportion of the variation in the Y variable explained by the X variable. To put differently, it tells us how many predicted data points would fall on the regression line. It is denoted by R^2 or R square. It is similar to the coefficient of correlation and simply be calculated by squaring the correlation coefficient between the two variables, that is, $R^2 = (r)^2$. The value of the R^2 ranges between 0 to +1. For instance, $R^2 = 0.9$ indicates that 90 per cent of the variation in the dependent variable can be explained by the independent variable, ceteris paribus. On the other hand, a value closer to zero implies a poor fit model, that is, a smaller proportion of total variation in Y is explained by the explanatory variable.

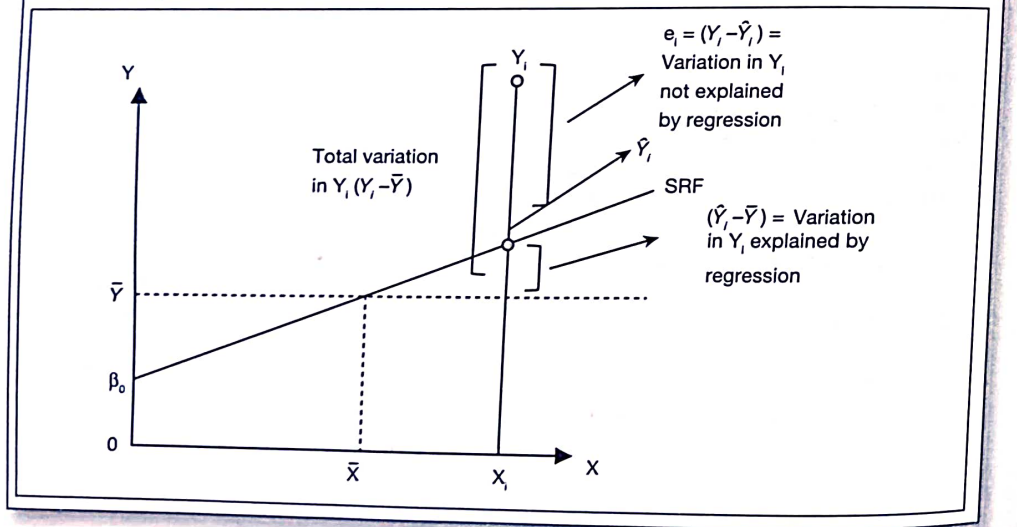
Here, the *explained variation* (SS_R) or regression sum of squares is the variation in the Y variable due to the X variable. Whereas the variation in the Y variable due to the errors or the variation in the Y variable not accounted for by the X variable is known as the unexplained variation or residual sum of squares (SS_E). Total variation (SS_y) in the dependent variable is the sum of the explained variation and unexplained variation, that is, $SS_y = SS_R + SS_E$.

Mathematically, these three types of variations are written as below which are derived from Figure 11.12:

Explained variation $(SS_R) = \sum_{i=1}^{n}(\hat{y}-\bar{Y})^2$. It is the sum of the squares of the deviations of the predicted values from the mean value of the dependent variable.

FIGURE 11.12

Graphical Representation of Three Types of Variations



Residual sum of squares $(SS_E) = \sum_{i=1}^{n}(Y_i - \hat{y})^2$. This sum of squares is the divergence between the actual data and the estimated model.

Total variation $(SS_y) = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$. The sum of squares represents the difference between the actual Y value and its mean.

Here, $\hat{y}$ = Value estimated by the regression line
$\bar{Y}$ = Mean of the dependent variable
Y_i = ith observation of dependent variable

R^2 can now be expressed as the ratio between the explained variation to the total variation, that is:

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

$$= \frac{SS_x}{SS_y}$$

or

$$R^2 = 1 - \frac{\text{Unexplained variation}}{\text{Total variation}}$$

$$= 1 - \frac{SS_E}{SS_y}$$

$$= 1 - \frac{\sum \hat{u}_i^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

In the case of multiple regression as well, the coefficient of multiple determination (R^2) represents the proportion of the variation in the Y variable explained by all the explanatory variables together. However, one problem with R^2 is that as we increase the number of X variables, the value of R^2 also increases even if there is no relationship between the additional X variable and Y variable, thus leading to an overfitting model. Thus, we need to adjust the R^2 and instead use *adjusted* R^2 ($\bar{R}^2$). It can be computed as follows:

$$R^2_{adj} = 1 - \left(\frac{[1-R^2][N-1]}{N-k-1}\right)$$

Here,
R^2 = Coefficient of multiple determination
N = Number of observations in the variable or the sample size
k = Number of independent variables

It may be noted that the value of adjusted R^2 is always less than R^2. The reason being that as the number of explanatory variables (k) increases in the model, the denominator becomes smaller and the value in parenthesis will become larger. Resultantly, $\bar{R}^2$ the value will come out to be smaller than the R^2 value.

The adjusted R^2 increases or decreases depending on whether the additional variables improve the explanatory power of the model or not. Upon the addition of X variables, the adjusted R^2 will increase only if the additional independent variables contribute more to the model than what would have been expected by chance. On the other hand, it will decrease if the additional independent variables improve the model by less than expected by chance. Hence, adjusted R^2 should be used to compare the models having different number of independent variables. Therefore, adjusted R^2 determines the proportion of the variation in Y variable that is explained by all the explanatory variables adjusted for the number of independent variables in the model.

Illustration

Calculate the value of the residuals or the errors for the previous illustration. Also, compute the value of the coefficient of determination and interpret the results.

First Method

The continuing regression equation can be written as:

$$\text{Revenue} = \beta_0 + \beta_1 \times \text{Promotion expenses}$$

Substituting the value beta coefficients, that is, $\beta_1 = 5.101726$ and $\beta_0 = 3{,}183.3579$, we get:

- Estimated/predicted of revenue = 3,183.3579 + 5.101726 × Promotion expenses (Column 3)
- Residual/errors = Actual value of Y variable – Predicted value of Y variable (Column 4)

(in crores)

Months	Promotion Expenses (X)	Revenue (Y)	Predicted Revenue ($\hat{y}$)	Residuals ($Y_i - \hat{y}$)	$(Y_i - \bar{Y})^2$	$(Y_i - \hat{y})^2$
1	400	5,500	5,224.05	275.95	320,167.36	76,149.23
2	410	5,540	5,275.07	264.93	276,500.69	70,190.15
3	430	5,550	5,377.10	172.90	266,084.03	29,894.31
4	450	5,600	5,479.13	120.87	217,000.69	14,608.39
5	480	5,600	5,632.19	–32.19	217,000.69	1,035.98
6	510	5,600	5,785.24	–185.24	217,000.69	34,313.26
7	530	5,700	5,887.27	–187.27	133,834.03	35,071.14
8	540	5,700	5,938.29	–238.29	133,834.03	56,782.21
9	550	5,800	5,989.31	–189.31	70,667.36	35,837.31
10	560	5,800	6,040.32	–240.32	70,667.36	57,755.96
11	560	5,900	6,040.32	–140.32	27,500.69	19,691.02
12	570	5,900	6,091.34	–191.34	27,500.69	36,611.75
13	570	5,950	6,091.34	–141.34	13,417.36	19,977.55
14	580	6,000	6,142.36	–142.36	4,334.03	20,266.15
15	620	6,040	6,346.43	–306.43	667.36	93,898.29
16	630	6,050	6,397.45	–347.45	250.69	120,718.40
17	640	6,100	6,448.46	–348.46	1,167.36	121,426.33
18	640	6,200	6,448.46	–248.46	18,000.69	61,733.76
19	660	6,300	6,550.50	–250.50	54,834.03	62,748.91

Months	Promotion Expenses (X)	Revenue (Y)	Predicted Revenue ($\hat{y}$)	Residuals ($Y_i - \hat{y}$)	$(Y_i - \bar{Y})^2$	$(Y_i - \hat{y})^2$
20	590	6,600	6,193.38	406.62	285,334.03	165,342.68
21	630	6,850	6,397.45	452.55	614,917.36	204,805.54
22	640	6,800	6,448.46	351.54	539,000.69	123,578.40
23	680	7,000	6,652.53	347.47	872,667.36	120,734.11
24	690	7,500	6,703.55	796.45	2,056,834.03	634,334.01
Total	13,560	145,580			6,439,183.33	2,217,504.85

Mean of dependent variable (Y) = 145,580/24 = 6,065.8333

$$\text{Total variation} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = 6{,}439{,}183.33$$

$$\text{Unexplained variation} = \sum_{i=1}^{n}(Y_i - \hat{y})^2 = 2{,}217{,}504.85$$

Explained variation = Total variation – Unexplained variation = 4,221,678.48

$$\text{Coefficient of determination } (R^2) = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{4{,}221{,}678.48}{6{,}439{,}183.33} = 0.65562$$

Second Method (Using Excel)

To estimate the residual and coefficient of determination in the regression model using Excel, we use the option of data analysis to obtain the results in the form of a summary output and residual output. We create a new worksheet and follow similar steps as performed for computing the coefficients and constant in the previous illustration.

Step 1. Enter the promotion expenses (X) and revenue (Y) in the Excel sheet.

Step 2. Go to **Data** tab, click on the option of **Data Analysis** in the **Analysis** group on the right side.

Step 3. A dialogue box will appear. Select the option of **Regression** from the list and click **OK**.

Step 4. A **Regression** dialogue box will appear. Fill the following details (Figure 11.13):

FIGURE 11.13

Regression Analysis Dialogue Box

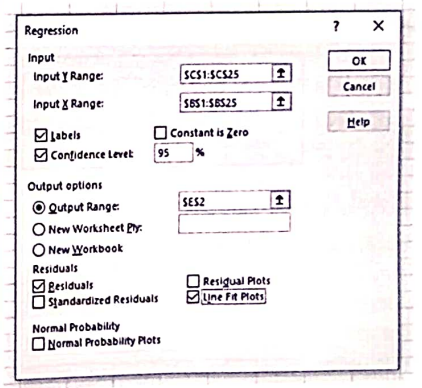


FIGURE 11.14

Regression Analysis (Summary Output)

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.809705674
R Square	0.655623278
Adjusted R Square	0.63996979
Standard Error	317.483346
Observations	24

FIGURE 11.15

Regression Analysis (Residual Output)

RESIDUAL OUTPUT

Observation	*Predicted Revenue(Y)*	*Residuals*
1	5224.0485	275.9515002
2	5275.065762	264.9342376
3	5377.100288	172.8997123
4	5479.134813	120.865187
5	5632.186601	-32.1866009
6	5785.238389	-185.2383888
7	5887.272914	-187.2729141
8	5938.290177	-238.2901767
9	5989.307439	-189.3074394
10	6040.324702	-240.324702
11	6040.324702	-140.324702
12	6091.341965	-191.3419647
13	6091.341965	-141.3419647
14	6142.359227	-142.3592273
15	6346.428278	-306.4282778
16	6397.44554	-347.4455405
17	6448.462803	-348.4628031
18	6448.462803	-248.4628031
19	6550.497328	-250.4973284
20	6193.37649	406.6235101
21	6397.44554	452.5544595
22	6448.462803	351.5371969
23	6652.531854	347.4681463
24	6703.549116	796.4508837

- **Input Y Range:** The cell range of the values of the dependent variable (revenue) for which the regression model is developed, that is, **C1:C25**.
- **Input X Range:** The cell range of the values of the independent variable (promotion expenses), that is, **B1:B25**.
- **Check on 'Labels':** Remember if you choose the label/name of the variables, that is, promotion expenses and revenue in the input range, then check the box **Labels**.
- Check on the **Confidence Level** and set it at **95 per cent**, which is also the default value.
- Under the residuals, options **Select** the **Residuals** checkbox to get the residual output as the results. Check on the **Line Fit Plots** (optional) to create a scatter plot of the line of best fit.
- **Output Range:** Under the output options, choose the cell reference where the result is desired in the existing worksheet, say **E2**.

Step 5. Click **OK**. The **Regression** result will be displayed in the form of a summary output (Figure 11.14), residual output table (Figure 11.15), and line fit plot in Figure 11.16.

Interpretation of Results

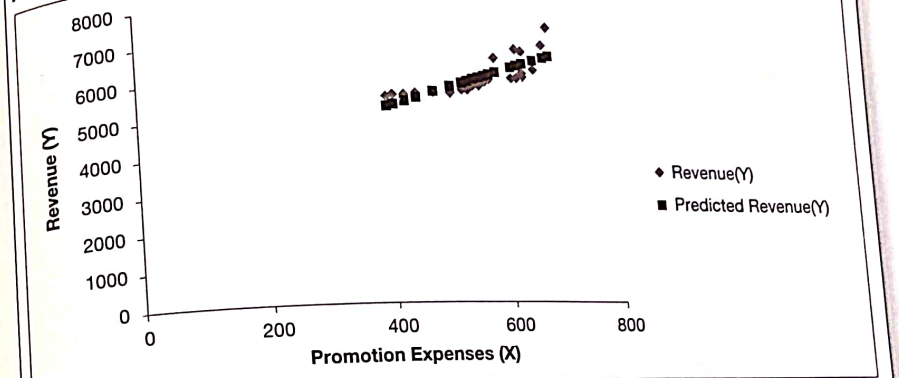
Figure 11.14 shows the coefficient of determination or the R^2 is produced as a part of the regression output. It comes out to be 0.65562. It means that 65.52 per cent of the variation in the revenue can be explained by the promotion expenses. The revenue earned by the company is strongly related to the amount spent on promotional activities as these expenses account for 65.52 per cent of the variation in the sales revenue of the company. The remaining goes unexplained. The adjusted R^2 is equal to 0.6399.

Interpretation of Results

Figure 11.15 shows the predicted values of revenue with the help of the regression equation and values of the residuals or errors the difference between the actual values of Y and the values predicted from the regression model. For instance, the residual of 275.95 in the 1st month implies that the regression model is *underpredicting* the revenue by 275.95 crore or the residual of (–32.1866) in the 5th month implies that the regression model is *overpredicting* the revenue by 32.1866 crore.

FIGURE 11.16

Regression Analysis: Promotional Expense Line Fit Plots



Interpretation of Results

Figure 11.16 displays the scatter plot of the line of best fit determined using the least squares method obtained with regression output in Excel. The actual values of revenue plotted above the regression line are positive residuals whereas the actual values of revenue plotted below the regression line are negative residuals. Distance of the *actual values* (represented by grey dots) from the *predicted or estimated values* (represented by the straight line in black colour) are the residuals.

Summing up, it can be concluded that regression analysis is one of the most important types of data analysis. In this chapter, we learned how to predict the value of a dependent variable from an independent one, for which we developed and estimated a regression model. The strength of the relationship, established by the value of R^2, describes the fit of the model. We used the ordinary least squares method to obtain the best fitting line which attempts to minimize the sum of squared errors. Now the question arises as to how we can find out whether the estimated coefficients are statistically significant. To figure it out, we will explain the process of conducting the hypothesis testing in the next chapter.

PRACTICE QUESTIONS

1. The working hours and the electricity consumption (in megawatts) in a factory are recorded for 20 days. Using an OLS method, estimate the values of the coefficients and interpret the results.

Working Hours (X)	*Electricity Consumption (Y)*
14	920
8	730

Working Hours (X)	*Electricity Consumption (Y)*
10	840
9	850
5	700
5	720
8	760
8	790
9	820
7	710
6	710
7	750
7	790
13	900
8	870
12	930
11	910
5	650
13	870
14	970

Ans. β_0 (constant) = 558.9582717 and β_1 (slope) = 27.9934892
The regression equation can be written as:
Electricity consumption = 558.9582717 + 27.9934892 × Working hours
This can be interpreted as the working hours increase, the electricity consumption in the factory also tends to increase. The value of β_1 as 27.9934 demonstrates that increasing the working hours by an hour (a unit of the independent variable) will increase the electricity consumption by 27.9934 megawatts (dependent variable).

2. The interest rate on the loan charged by a bank and the demand for the loan (in lakh) is given below. Estimate the values of the coefficients and comment on the goodness of the fit of the model using the OLS method.

The Interest Rate on Loan (%; X)	*Demand for Loans (in Lakh; Y)*
22	200
19	350
36	100
28	170
32	140
19	390
29	250

The Interest Rate on Loan (%; X)	*Demand for Loans (in Lakh; Y)*
38	80
35	100
38	70
11	400
25	180

Ans. β_0 (constant) = 557.4474 and β_1 (slope) = –12.8294260
The regression equation can be written as:
Demand for loans = 557.4474 – 12.8294 × Interest rate on the loan
R^2 (coefficient of determination) is equal to 0.863318, which indicates a good fit model. It tells that the interest rate on the loan (X) explains 86.3318 per cent of the variation in the demand for loans (Y).

3. The rent (in ₹) and the area of the flats (in sq. ft) for 20 houses in Delhi are given in the table below. Estimate the regression coefficients and also find the value of rent when the area of the flat is 7,500 sq. ft.

Area of the House (in Sq. Ft; X)	*Rent (Y)*
2,000	16,000
2,600	18,000
2,800	20,000
3,300	21,000
4,000	25,000
4,000	27,000
4,000	31,000
4,300	32,500
4,800	33,000
6,000	37,000
6,500	39,000
6,600	42,000
7,200	43,000
7,200	44,000
7,400	51,000
8,400	51,000
8,400	52,000
8,500	55,000
8,700	56,000
8,800	59,000

Ans. β_0 (constant) = 3,465.17985 and β_1 (slope) = 5.9151203

The regression equation is:
Rent = 3465.17985 + 5.9151203 × Area of the flat
The value of rent (predicted) when the area of the flat is 7,500 sq. ft is:
3,465.17985 + 5.9151203 × 7,500 = ₹47,828.58

4. The annual profits (in crores) for a company and its average annual share prices (₹) for the last 15 years is given below. Calculate the residuals and interprets the results.

Profits (in Crores: X)	*Share Prices (Y)*
600	7,800
560	6,000
540	6,800
460	6,800
400	5,900
360	5,700
330	5,500
330	4,500
280	4,300
270	2,800
220	2,500
220	2,600
200	2,100
170	1,000
120	1,800

Ans.

Observation	*Predicted Share Prices*	*Residuals*
1	7,931.314789	–131.315
2	7,394.566344	–1394.57
3	7,126.192122	–326.192
4	6,052.695232	747.3048
5	5,247.572564	652.4274
6	4,710.824119	989.1759
7	4,308.262785	1,191.737
8	4,308.262785	191.7372
9	3,637.327229	662.6728
10	3,503.140117	–703.14
11	2,832.204561	–332.205
12	2,832.204561	–232.205
13	2,563.830339	–463.83
14	2,161.269005	–1,161.27
15	1,490.333449	309.6666

Interpretation: In (1st year, 2nd year, 3rd year, 10th year, 11th year, 12th year, 13th year and 14th year), the negative residual implies that the regression model is overpredicting the share prices. Similarly, in (4th year, 5th year, 6th year, 7th year, 8th year, 9th year, 15th year), the positive residual implies that the regression model is underpredicting the share prices.

5. The price of artificial jewellery (₹) and the orders placed (units) for the last 15 months are given below. You are required to interpret the regression output and the residual output obtained using the data analysis option using Excel.

Price (X)	*Orders Placed (Y)*
4,500	25,541
3,950	27,963
4,100	30,733
3,840	31,816
3,560	31,974
3,430	32,078
3,600	36,356
3,280	39,326
3,150	42,256
2,980	42,950
2,950	47,886
2,700	51,106
2,680	52,775
2,500	54,612
2,670	55,623

Ans. Summary output

	Regression Statistics
Multiple R	0.952024899
R^2	0.906351408
Adjusted R^2	0.89914767
Standard error	3,251.083536
Observations	15

Multiple R is the correlation coefficient that measures the strength of a linear relationship between two variables. The correlation coefficient of 0.95 indicates a strong correlation between the dependent and independent variables.
R^2 of 0.90635 indicates that 90.635 per cent of the variation in the orders placed (dependent variable) can be explained by the price of jewellery (independent variable).
Adjusted R^2 of 0.899 is the R^2 adjusted for the number of independent variables in the model. It is useful to compare the models in case of an additional independent variable is added in the model.

Analysis of variance (ANOVA)

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	132,982,6457	1.33E + 09	125.8168221	4.67505E – 08
Residual	13	137,404,074	10,569,544		
Total	14	1,467,230,531			

SS (regression): Explained variation = 1,329,826,457
SS (residual): Sum of Squares of errors or unexplained variation = 137,404,074
SS (total): Total variation = 1,467,230,531

	Coefficients
Intercept	94,509.14895
Price	–16.32876797

Intercept (β_0) is equal to 94,509.14895 indicating the units of orders placed when the price of the jewellery is ₹0. As stated, we ignore the interpretation of β_0, as it will be irrational in practice to find the price of jewellery as ₹0.

The negative sign of slope coefficient (β_1) means that as the price increase, the number of orders placed tends to decrease. The value of β_1 = –16.329 (rounded off) demonstrates that by increasing the price by one rupee (a unit of the independent variable) the number of orders will decrease by 16 units (dependent variable).

12

CHAPTER

Hypothesis Testing in Regression Analysis

LEARNING OBJECTIVES

After reading this chapter, the readers will be able to understand

- The concept of hypothesis testing
- Performing hypothesis testing to test the significance of the regression coefficients using:
  - The *t*-statistics approach
  - The *p*-value approach
  - Confidence interval approach
- Interpreting the statistical significance of the regression model and the coefficients with the help of regression results obtained in Excel

In the last chapter, we have observed that the population refers to the aggregate of all the items under study, but at times, it may not be possible to collect the data for the entire population. In that case, we select a sample from the population. The sample is the set of items taken from the population and is representative of it. The findings from this sample are generalized to the population as a whole, and on its basis, we can infer about the population.

The population parameters describe the characteristics of the entire population, while sample statistics describe the characteristics of the sample. The regression coefficients estimated from the regression model are based on the sample data and are the estimates for the population parameters (true values but unknown). Therefore, we use hypothesis testing to test the significance of the

regression coefficients or to test whether the sample belongs to the population referred to. A regression output in Excel exhibits the following type of statistic: (a) *t*-statistics, (b) *p*-values and (c) confidence intervals. The following sections attempt to understand the meaning of these tests so that they are interpreted aptly.

12.1. HYPOTHESIS TESTING USING THE *t*-STATISTICS

In hypothesis testing, we make assumptions or statements about the population parameters. However, in reality, the assumptions made may not be true. Based on these assumptions, the hypothesis is formed, which can be tested using statistical tests wherein the validity of a belief, claim or assumption made about the population based on the data is tested. The first approach is associated with the calculation of *t*-statistic, which tests the individual significance of the parameters estimated from the regression model. Hypothesis testing involves the following steps:

Step 1. Formulation of the Hypothesis

The first step involved is the formulation of the *null hypothesis* (H_0) and the *alternative hypothesis* (denoted by H_1 or H_A). The null hypothesis is assumed to be true, and the assumption is tested for rejection. With the rejection of the null hypothesis, we accept the alternative hypothesis. The alternative hypothesis, on the other hand, is the opposite of the null hypothesis, which states that the population parameter is smaller or greater than the *null value* (*one-tailed*) or it is *non-directional* (*two-tailed*). Note that it is the null hypothesis only that is tested and not the alternative hypothesis.

$H_0: \mu = 0$
$H_1: \mu >$ or < 0 (one-tailed test)
$H_1: \mu \neq 0$ (two-tailed test)

The null hypothesis always has strict equality. The following examples will help get a clear picture of the two tests. In the one-tailed test, for example, an Indian website claims that the average salary of finance managers is greater than 10 lakh per annum. The hypotheses can be stated as:

$H_0: \mu \leq 10$
(no change, i.e., the average salary of finance managers is equal to or less than 10 lakh)

$H_1: \mu > 10$
(the average salary of finance managers is greater than 10 lakh)

On the other hand, in a two-tailed test, for example, if the objective is to examine the perception of the average typing speed of a student is 50 words per minute, the hypotheses for the study can be stated as:

$H_0: \mu = 50$
(no difference, i.e., the average typing speed of a student is equal to 50 words per minute)

$H_1: \mu \neq 50$
(average typing speed of a student is not equal to 50 words per minute)

TABLE 12.1

Type 1 and Type 2 Errors

True State of Null Hypothesis → / *Statistical Decision* ↓	*Null hypothesis is true*	*Null hypothesis is false*
Reject the null hypothesis	Type 1 error (α; incorrect decision)	Power ($1 - \beta$; correct decision)
Do not reject the null hypothesis	Confidence level ($1 - \alpha$; correct decision)	Type 2 error (β; incorrect decision)

Step 2. Setting the Significance Level

Secondly, we set a significance level which is the probability at which we reject the null hypothesis when it is true. This is termed as the *Type 1 error*, denoted by alpha (α). For instance, $\alpha = 0.05$ indicates that we are willing to take a 5 per cent chance that we may be wrong when we reject the null hypothesis when it is true. *Type 2 error*, denoted by beta (β), is the probability of not rejecting the null hypothesis when it is false. The Type 1 error and Type 2 error are inversely related to each other, which means that as one of the errors increases, the other tends to decrease.

Power or $1 - \beta$ refers to the probability of rejecting a null hypothesis if it is false or the probability of not making a Type 2 error. Similarly, the *confidence level or* $1 - \alpha$ is the probability of not rejecting the null hypothesis when it is true or probability of not making a Type 1 error. Table 12.1 summarizes the type of errors and their significance.

Table 12.1 illustrates the significance level and the rejection region in a two-tailed test. The rejection area is on both sides of the distribution since the alternative hypothesis in this test is non-directional. Here, if $\alpha = 5$ per cent, then the distribution will have two rejection areas of 2.5 per cent ($\alpha/2$) on each side of the distribution (Figure 12.1). On the other hand, a one-tailed test (Figure 12.2) implies that there will be a single rejection region on either side of the distribution, that is, if $\alpha = 5$ per cent, then the distribution has one rejection region of 5 per cent on either side, left or right, depending on the sign of the null value in the alternative hypothesis. A significance level of 5 per cent is the most commonly used for making decisions in the field of finance, economics or medicine. In Excel, the confidence level is, by default, set at 95 per cent, which means a significance level of 5 per cent.

Step 3. Selecting the Suitable Test, Calculating the Test Statistic and Rejection Region

After considering several factors such as population distribution or sample size, we choose the most appropriate statistical test to carry out the hypothesis testing. Based on the statistical test selected, we can calculate the test statistic value. Some common hypothesis tests are *z*-test, *t*-test, *chi-squared* test and *ANOVA*, and their test statistics are *z*-statistics, *t*-statistics, *chi-square* statistics and *F*-statistics, respectively.

FIGURE 12.1

Rejection Areas for a Two-tailed Test

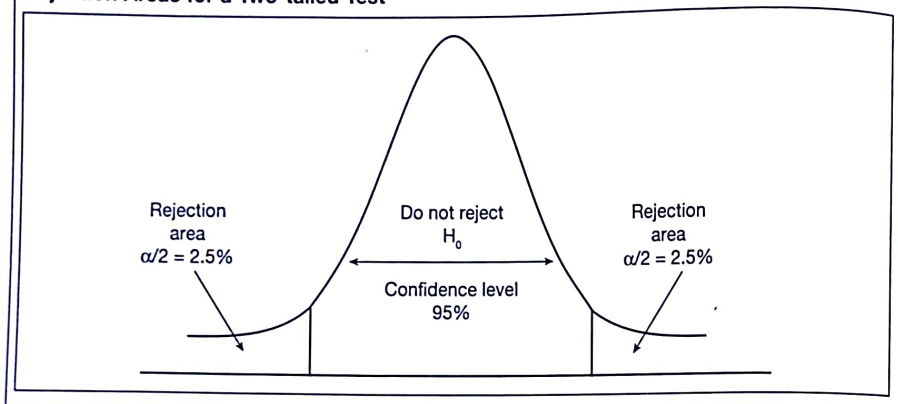
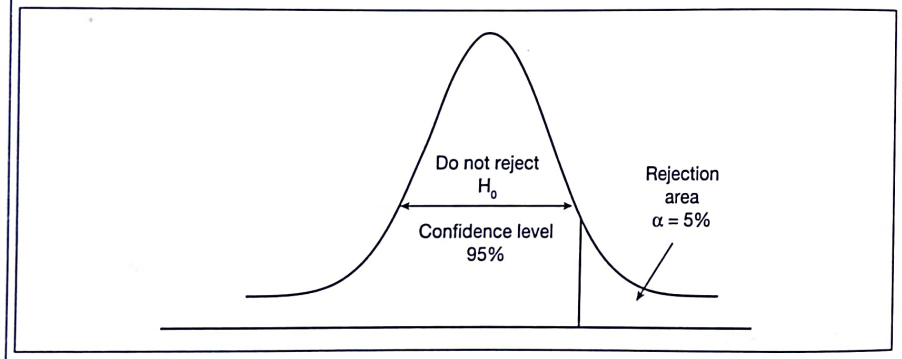


FIGURE 12.2

Rejection Areas for a One-tailed Test



In the case of regression analysis, we translate our problem into t-distribution or z-distribution, based on sample size. A t-distribution is similar to a normal distribution with a bell-shaped curve but having relatively heavier tails. The t-test can be used to compare the means of population and sample means and determine whether the means of the two groups are significantly different from each other. The t-test is used for estimating the statistics when the sample size is small ($n < 30$), and the population standard deviation is unknown.

The formula for calculating the t-statistics is given by:

$$t = \frac{\text{Difference between the sample and the population mean}}{\text{Standard error}} = (\bar{x} - \mu) / (s / n)$$

Here,
$\bar{x}$ = Sample mean
μ = Population mean
s = Sample standard deviation
n = Sample size

It needs to be noted that many samples could be taken from the population. Each sample would have a mean, and it is possible to think of a distribution of the sample mean and standard deviation of the sample mean. If the sample size is small, then the variation would be high, and larger samples would have smaller variation. A measure of how different the sample mean is expected to be from the population mean is provided by the standard error of the mean, which is the standard deviation of the sample mean.

The t-test can also be used to conduct the hypothesis testing on the regression coefficients to determine if the slope coefficient is significantly different from zero. In the regression context, the t-statistic can be calculated as:

$$t = \frac{b_1 - \beta_1}{s_b}$$

Here,
b_1 = Estimated regression coefficient
β_1 = Slope coefficient against which we conduct the hypothesis test or the null value
S_b = Standard error of β_1

The formula for standard error is: $\sqrt{\dfrac{\frac{SS_E}{n-2}}{\sum(X-\bar{X})^2}}$

The following table summarizes the t-statistic and the rejection region.
Null Hypothesis: H_0: $\mu = \mu_0$

Test statistic value: $t = \dfrac{\bar{x} - \mu}{s / \sqrt{n}}$

Alternative Hypothesis	Rejection Region at Level α
H_1: $\mu > \mu_0$	$t \geq t_{\alpha,n-1}$ (upper-tailed)
H_1: $\mu < \mu_0$	$t \leq -t_{\alpha,n-1}$ (lower-tailed)
H_1: $\mu \neq \mu_0$	either $t \geq t_{\alpha,n-1}$ or $t \leq -t_{\alpha,n-1}$ (two-tailed)

The cut-off value or the critical value ($t_{\alpha,n-1}$) for the rejection region is calculated for the given degrees of freedom and the significance level using the t-distribution table. We have already understood the concept of significance level in Step 2, let us now go through the idea of the degree

of freedom. The *degree of freedom* (df) represents how many values in a data set are free to choose. It depends upon the number of parameters to be estimated as we lose one degree of freedom for every parameter estimated. The degree of freedom is calculated by subtracting the number of parameters to be estimated (k) from the sample size (n). Generally, for ordinary t-test, the degree of freedom is taken as $n - 1$. In the case of regression analysis, the degree of freedom is taken as $n - k - 1$. For a simple linear regression (with one independent variable), the df $= n - 1 - 1 = n - 2$.

The degree of freedom is given in the regression summary output obtained using the data analysis option in Excel as a residual degree of freedom (df). As shown in Figure 12.3 below, the residual degree of freedom (df) for $n = 28$ and one X variable is equal to $28 - 1 - 1 = 26$.

Step 4. Check whether the t-statistics fall in the Rejection Region

If the t-statistics calculated in Step 3 above fall in either of the rejection areas, we reject the null hypothesis, and if the t-statistics falls outside the rejection area, we do not reject the null hypothesis. For example, the hypothesis stated as:

$H_0: \beta_1 = 0$ (no significant relationship between the dependent and independent variable)

$H_1: \beta_1 \neq 0$ (a significant relationship between the dependent and independent variable)

The cut-off value for say $\alpha = 0.05$, and df = 26 comes out to be ± 2.056 as obtained from Table 12.2.[1] If the t-statistics calculated using the formula: $t = \dfrac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$ is more than +2.056 or less than −2.056, that is, the value of the t-statistics falls to the left of negative cut-off value or the right of the positive cut-off value, *we reject the null hypothesis* and conclude that there is a significant relationship between the dependent and independent variable. On the other hand, if the t-statistic does not fall in either of the rejection areas, *we do not reject the null hypothesis* and conclude that there is a significant relationship between the dependent and independent variables. It is important to note here that if a critical value is significant at 1 per cent, then it must be significant at 5 per cent also. However, the converse is not true.

FIGURE 12.3

Regression Analysis Output: Residual Degree of Freedom

ANOVA	df
Regression	1
Residual	26
Total	27

12.2. HYPOTHESIS TESTING USING THE p-VALUE

The concept of p-value plays a vital role in interpreting the regression results and assessing the significance of the estimated regression coefficients. The *p-value is the probability that lies between 0 and 1 and is based on the assumption that the null hypothesis is true*. The p-values for every coefficient

[1] The t-distribution table has degrees of freedom in the rows and the significance level in the columns for both the one-tail and two-tail tests. The intersection cell, in a two-tail test, if the degree of freedom is 26, and the significance level is at 5 per cent, the critical value read from the Table 12.2 is 2.056.

TABLE 12.2

The t-distribution Table

one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707

is automatically produced by Excel in the regression summary output obtained using the data analysis option. This means that Excel conducts a hypothesis testing for every coefficient and produces the results in the form of p-values. To interpret the p-value, it is important to find out which test statistic is used. The first three steps for conducting the hypothesis testing, namely the formulation of hypothesis, setting the significance level and selecting the suitable test remain the same.

Suppose the given test statistic is t-statistic. The p-values can be interpreted as follows. In a two-tailed hypothesis test, the rejection region on both sides is calculated using $\alpha/2$. In Figure 12.4, if the p-value is greater than the significance level, it does not fall in the rejection region. In that case, we do not reject the null hypothesis when $p > \alpha$. Similarly, in Figure 12.5, the t-statistics falls in the rejection region as we observe that the p-value is less than the significance level. Thus, we reject the null hypothesis when $p < \alpha$.

For example, for a two-tail test say $\alpha = 0.05$, residual df = 26 and t-statistics = 24.6545, the p-value is equal to 0.0001. Since the p-value is less than α, that is, $0.0001 < 0.05$, we *reject the null hypothesis* (Figure 12.6). On the other hand, say for another two-tail test, the $\alpha = 0.05$, residual df = 26, and t-statistics is 0.85 then p-value is equal to 0.40. Since the p-value is greater than α, that is, $0.40 > 0.05$, *we do not reject the null hypothesis*.

FIGURE 12.4

The *t*-distribution *p*-value > Alpha (α)

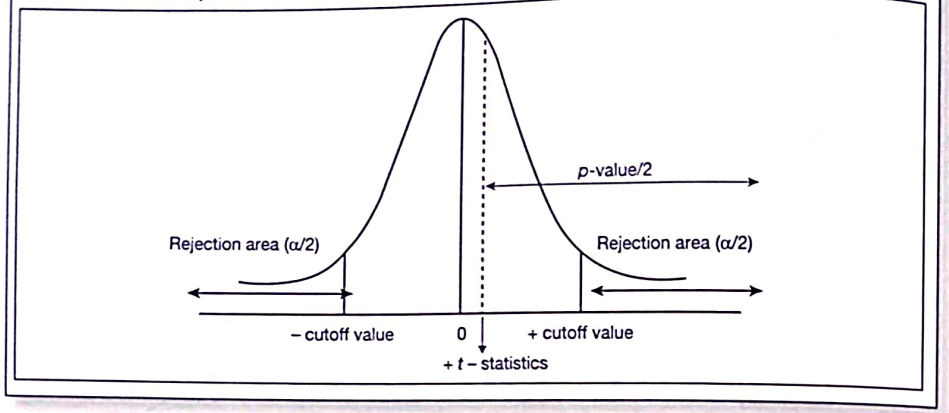


FIGURE 12.5

The *t*-distribution *p*-value < Alpha (α)

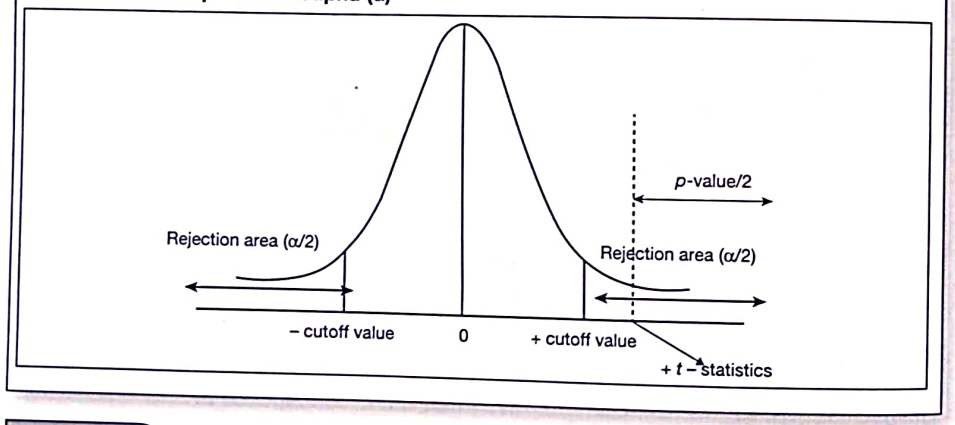


FIGURE 12.6

The *t*-distribution Computed *p*-value < Alpha (α)

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-1648.103786	194.1050806	-8.490781286	5.68922E-09
X Variable 1	306.9271804	12.44910252	24.65456284	1.47084E-19

12.3. HYPOTHESIS TESTING USING THE CONFIDENCE INTERVAL

The third approach used for hypothesis testing is the *confidence interval approach*. It is the most simple and easiest approach to conduct hypothesis testing. It involves interpreting the confidence intervals obtained for the estimated regression coefficients in the regression output in Excel. The confidence level of 0.95 means that if we keep on repeating the event again and again, then in 95 per cent of the events, the true population parameter will lie in the range of confidence interval.

The following steps need to be followed for conducting hypothesis testing using confidence intervals, Step 1 being the same as above.

Step 2. Consider the 95 Per Cent Confidence Interval as Produced for the Slope Coefficient of the *X* Variable by the Regression Output in Excel

In this step, we check whether or not the null value falls in the 95 per cent confidence interval. The lower limit and the upper limit for the 95 per cent confidence interval are specified in the regression output. If the null value indeed falls within the confidence intervals, *we do not reject the null hypothesis*, or we may say that we do not reject the null hypothesis for any value between the two specified limits of the confidence interval. On the other hand, if the value in the null hypothesis does not fall in the confidence intervals, *we reject the null hypothesis.*

We can conclude that:

- If the value specified in H_0 lies outside the interval, we can reject H_0.
- If the value specified in H_0 lies within the interval, we do not reject H_0.

Continuing on the previous example, we need to test the significance of the regression model and interpret the results obtained based on confidence intervals. The highlighted part in the regression output displays the upper and the lower limit for the 95 per cent confidence interval (Figure 12.7). We notice that the reference value in the null hypothesis, that is, 0 does not fall in the specified range; it falls outside the 95 per cent confidence interval. Hence, we reject the null hypothesis, which further implies that there is a significant relationship between the dependent and independent variables.

12.3.1. Illustration

The regression output realized in the example of Chapter 11 (a company incurring promotion expenses to boost their revenue) is reproduced below in Figure 12.8.

FIGURE 12.7

Regression Analysis Output (95% Confidence Intervals)

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1648.103786	194.1050806	-8.49078129	5.69E-09	-2047.092494	-1249.115079	-2047.092494	-1249.115079
X Variable 1	306.9271804	12.44910252	24.65456284	1.47084E-19	281.3376836	332.5166771	281.3376836	332.5166771

FIGURE 12.8

Test Results of the Regression Equation Displayed in the Excel Sheet

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.809705674
R Square	0.655623278
Adjusted R Square	0.639969179
Standard Error	317.483346
Observations	24

ANOVA

	df	SS	MS	F	Significance F
Regression	1	4221678.483	4221678.483	41.88352807	1.6435E-06
Residual	22	2217504.85	100795.675		
Total	23	6439183.333			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	3183.357994	450.0836347	7.07281436	4.2821E-07	2249.941666	4116.774322	2249.941666	4116.774322
Promotion Expenses(X)	5.101726264	0.788307256	6.471748456	1.6435E-06	3.466877077	6.73657545	3.466877077	6.73657545

We have already developed and estimated the regression model for the promotion expenses as the independent variable and the revenue as the dependent variable. The estimated regression model is:

$$\text{Revenue} = 3{,}183.3579 + 5.101726 \times \text{Promotion expenses}$$

Now, the sales manager wishes to test the significance of the regression model developed and assess the importance of the promotion expenses in explaining the variation in the revenues.

Solution and Interpretation

First Method (the *t*-value Approach)

The null hypothesis assumes that the beta coefficient is equal to zero (null value), which means that there is no significant relationship between the promotion expenses and revenues. The hypotheses can be stated as:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

The significance level (α) is set at 5 per cent.

Excel provides the *t*-statistic (*t*-value) as 6.4717.

The critical *t*-value is calculated using the table. The rejection region for this two-tailed test is either $t \geq t_{\alpha, n-1}$ or $t \leq -t_{\alpha, n-1}$. The table value of $t_{0.025,23}$ is 2.074. Since 6.4717 > 2.074, we reject the null hypothesis. This implies that there is a significant relationship between the revenue earned by the company (dependent variable) and promotion expense incurred (independent variable).

Second Method (the *p*-value Approach)

The *p*-value that Excel provides us is 1.6435E-06 which denotes a very small number equal to 0.000001643. We observe that this *p*-value is less than the significance level of 0.05, that is, $p < \alpha$, *therefore, we reject the null hypothesis* and conclude that there is a significant relationship between the revenues earned by the company (dependent variable) and promotion expense incurred (independent variable).

Third Method (Confidence Interval Approach)

The lower limit and the upper limit for the 95 per cent confidence interval are specified in the regression output are 3.466877 and 6.73657545, respectively (Figure 12.8). Since our null value $\beta_1 = 0$ does not fall within the confidence intervals, *we reject the null hypothesis*, which further implies that there is a significant relationship between the revenues earned by the company (dependent variable) and promotion expense incurred (independent variable).

It is important to note that the outcomes in all these approaches remain the same, which is to reject the null hypothesis.

To sum up, this chapter introduced a very important concept of hypothesis testing, which plays a crucial role in making inferences about the true population estimates. The estimated coefficients are checked, whether there is a significant linear relationship between the dependent and independent variables, using different approaches.

PRACTICE QUESTIONS

1. The area of the house (in sq. ft) and the electricity bill (in ₹) is given below. Test the significance of the regression coefficients estimated using the *t*-statistics approach. (Significance level = 5%).

Area of the House (in sq. ft) (X)	*Electricity Bill (Y)*
1,716	10,800
1,886	18,000
2,056	16,000
2,057	20,000
2,095	21,000
2,403	23,000
3,291	24,500
3,333	27,000
3,747	26,000
3,838	28,500
4,726	29,500
4,956	31,000
5,561	31,600
7,550	32,000
7,815	31,900
7,863	32,500
7,899	36,000
8,128	34,000

Area of the House (in sq. ft) (X)	*Electricity Bill* (Y)
8,215	39,000
8,500	38,311

Ans. $H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$
t-statistics = 9.403576788, residual df = 18 and critical value = 2.101
Since the value of the *t*-statistics falls to the right of the positive cut-off value, *we reject the null hypothesis* and conclude that there is a significant relationship between the area of the house and electricity bill.

2. The average annual income of 25 people and the taxes paid by them are given in the table below. Develop the regression model and test the significance of the regression model developed using the *p*-values approach. (Significance level = 5%).

Individual Annual Income (X)	*Taxes Paid* (Y)
700,000	70,000
750,000	90,000
770,000	123,200
850,000	85,000
854,000	128,100
870,000	69,600
1,030,000	206,000
1,030,000	206,000
1,080,000	216,000
1,100,000	220,000
1,120,000	240,000
1,130,000	226,000
1,160,000	230,000
1,240,000	248,000
1,300,000	260,000
1,360,000	272,000
1,460,000	280,000
1,500,000	300,000
1,650,000	495,000
1,780,000	534,000
1,800,000	548,000
1,825,000	550,000
1,970,000	591,000
1,980,000	594,000
2,042,000	612,600

Ans. $H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$
t-statistics = 22.95098679, *p*-value = 0.0000
The *p*-value of 0.00 is less than the significance level of 0.05, that is, $p < \alpha$. Therefore, we *reject the null hypothesis* and conclude that there is a significant relationship between the taxes paid (dependent variable) and average annual income earned by the individual (independent variable).

3. Abhinav uploads videos on YouTube on statistics and economics concepts. The views and likes on 20 of his videos are given below. Abhinav believes that the number of views has a significant impact on the number of likes on the videos. Conduct the significance test to determine if the views (the X variable) are useful to explain the variation in the number of likes (Y variable) using the *p*-values approach. Is the belief of Abhinav correct? (Significance level = 5%)

Views (X)	*Likes* (Y)
212,000	106,000
139,000	55,600
298,000	89,400
216,000	9,000
301,000	15,000
315,000	157,500
350,000	140,000
370,000	111,000
460,000	25,000
480,000	10,000
475,000	237,500
500,000	200,000
520,000	156,000
562,000	112,400
630,000	20,000
600,000	300,000
860,000	344,000
880,000	264,000
950,000	50,000
930,000	70,000

Ans. $H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$
t-statistics = 1.617637
p-value = 0.1231
$\alpha = 5\%$

The p-value of 0.1231 is more than the significance level of 0.05, that is, $p > \alpha$. Therefore, we *do not reject the null hypothesis* and conclude that there is no significant relationship between the likes (dependent variable) and the views (independent variable) on a video. The belief of Abhinav that the number of views has a significant impact on the number of likes on the videos is incorrect.

4. The literacy level (in %) and the number of people below the poverty line (as % of the total population) for 20 states are given below. Develop the regression model and test the significance of the regression model developed using a 95 per cent confidence interval approach. (Significance level = 5%). Is your answer the same if the test of significance is done using a 99 per cent confidence interval approach? (Significance level = 1%).

Literacy Level (%; X)	*People below the Poverty Line (%; Y)*
94	5
92	7
89	8
86	9
82	9
81	12
78	15
78	17
76	19
73	21
72	22
68	26
67	28
65	25
62	29
61	32
59	35
52	27
51	36
50	39

Ans. $H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$
Regression output (95 per cent confidence interval)

	Coefficients	*Standard Error*	*t-stat*	*p-value*	*Lower 95%*	*Upper 95%*
Intercept	74.99175186	3.703232912	20.2503471	7.76988E-14	67.21154821	82.7719555
Literacy level	–0.751277881	0.050736835	–14.80734613	1.59998E-11	–0.857872016	–0.64468375

The lower limit and the upper limit for the 95 per cent confidence interval are specified in the regression output are (–0.8578) and (–0.6446) respectively. Since our hypothesized value $\beta_1 = 0$ does not fall within the confidence intervals, *we reject the null hypothesis*, which further implies that there is a significant relationship between the people below the poverty line (dependent variable) and literacy rate (independent variable).

Regression output (99 per cent confidence interval)

	Coefficients	*Standard Error*	*t-stat*	*p-value*	*Lower 99.0%*	*Upper 99.0%*
Intercept	74.99175186	3.703232912	20.2503471	7.76988E-14	64.33221637	85.65128735
Literacy level	–0.751277881	0.050736835	–14.80734613	1.59998E-11	–0.897320841	–0.605234921

The lower limit and the upper limit for the 99 per cent confidence interval are specified in the regression output are (– 0.8973) and (– 0.6052), respectively. Since our hypothesized value $\beta_1 = 0$ does not fall within the confidence intervals, *we reject the null hypothesis*, which further implies that there is a significant relationship between the people below the poverty line (dependent variable) and literacy rate (independent variable).

5. The rainfall (in inches) and production of rice (in thousand million tonnes) are given below. Interpret the regression output as provide by Excel to test the significance of the regression coefficients. The agriculture officer claims that when rainfall increases by 1 inch, the rice production increase by 1,400 thousand million tonnes. Check the validity of this claim? (Significance level = 5%).

Rainfall (in Inches; X)	*Rice Production (in Thousand Million Tonnes; Y)*
30	117,939
28	116,480
29	112,760
25	109,698
23	106,646
16	105,482
22	105,301
20	105,241
17	104,408
15	99,172
13	96,682
12	95,970
11	93,345
12	91,785
14	89,083

Ans. Summary output.

Regression Statistics	
R^2	0.868283369
Adjusted R^2	0.858151321
Standard error	3,325.932545
Observations	15

Coefficients	*Standard Error*	*t-stat*	*p-value*	*Lower 95%*	*Upper 95%*
79,592.30045	2,704.490341	29.4296856	2.77836E-13	73,749.60429	85,434.99662
1,240.792659	134.0346561	9.257252531	4.37177E-07	951.2283893	1,530.356929

Rainfall (X variable) coefficient: $\beta_1 = 1{,}240.7926$ indicates that if the rainfall increases by one inch (a unit of the independent variable) the rice production increases by 1,240.7926 thousand million tonnes (dependent variable):

H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$
t-statistics = 9.2572 and cut-off value = 2.160
p-value = 0.1231
$\alpha = 5\%$

a. Since the value of the *t*-statistics falls to the right of the positive cut-off value, *we reject the null hypothesis* and conclude that there is a significant relationship between the rice production (dependent variable) and the rainfall (independent variable).
b. The *p*-value of 0.00 is less than the significance level of 0.05, that is, $p < \alpha$. Therefore, we *reject the null hypothesis* and conclude that there is a significant relationship between rice production (dependent variable) and the rainfall (independent variable).
c. Since our null value $\beta_1 = 0$ does not fall within the confidence intervals, *we reject the null hypothesis* which further implies that there is a significant relationship between the rice production (dependent variable) and the rainfall (independent variable).

The claim made by agriculture officer is that if rainfall increases by 1 inch, the rice production increase by 1,400 thousand million tonnes.
The hypothesis can be stated as:

H_0: $\beta_1 = 1400$
H_1: $\beta_1 \neq 1400$

Using the 95 per cent confidence interval approach, we observe that the hypothesized value of 1,400 falls within the lower and upper limit of the interval, that is, 951.2283893 and 1,530.356929. Hence, we do not reject the null hypothesis. It is possible that the true value of the increase in production can be 1,400 thousand million tonnes. The claim of the officer may be true.

Appendix 1
Present Value and Future Value Tables

Table A-1

Future Value Interest Factors for One Rupee Compounded at *r* Percent for *n* periods $FVIF_m = (1+r)^n$

Period	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%	11%	12%	13%	14%	15%	16%	20%	24%	25%	30%
1	1.0100	1.0200	1.0300	1.0400	1.0500	1.0600	1.0700	1.0800	1.0900	1.1000	1.1100	1.1200	1.1300	1.1400	1.1500	1.1600	1.2000	1.2400	1.2500	1.3000
2	1.0201	1.0404	1.0609	1.0816	1.1025	1.1236	1.1449	1.1664	1.1881	1.2100	1.2321	1.2544	1.2769	1.2996	1.3225	1.3456	1.4400	1.5376	1.5625	1.6900
3	1.0303	1.0612	1.0927	1.1249	1.1576	1.1910	1.2250	1.2597	1.2950	1.3310	1.3676	1.4049	1.4429	1.4815	1.5209	1.5609	1.7280	1.9066	1.9531	2.1970
4	1.0406	1.0824	1.1255	1.1699	1.2155	1.2625	1.3108	1.3605	1.4116	1.4641	1.5181	1.5735	1.6305	1.6890	1.7490	1.8106	2.0736	2.3642	2.4414	2.8561
5	1.0510	1.1041	1.1593	1.2167	1.2763	1.3382	1.4026	1.4693	1.5386	1.6105	1.6851	1.7623	1.8424	1.9254	2.0114	2.1003	2.4883	2.9316	3.0518	3.7129
6	1.0615	1.1262	1.1941	1.2653	1.3401	1.4185	1.5007	1.5869	1.6771	1.7716	1.8704	1.9738	2.0820	2.1950	2.3131	2.4364	2.9860	3.6352	3.8147	4.8268
7	1.0721	1.1487	1.2299	1.3159	1.4071	1.5036	1.6058	1.7138	1.8280	1.9487	2.0762	2.2107	2.3526	2.5023	2.6600	2.8262	3.5832	4.5077	4.7684	6.2749
8	1.0829	1.1717	1.2668	1.3686	1.4775	1.5938	1.7182	1.8509	1.9926	2.1436	2.3045	2.4760	2.6584	2.8526	3.0590	3.2784	4.2998	5.5895	5.9605	8.1573
9	1.0937	1.1951	1.3048	1.4233	1.5513	1.6895	1.8385	1.9990	2.1719	2.3579	2.5580	2.7731	3.0040	3.2519	3.5179	3.8030	5.1598	6.9310	7.4506	10.604
10	1.1046	1.2190	1.3439	1.4802	1.6289	1.7908	1.9672	2.1589	2.3674	2.5937	2.8394	3.1058	3.3946	3.7072	4.0456	4.4114	6.1917	8.5944	9.3132	13.786
11	1.1157	1.2434	1.3842	1.5395	1.7103	1.8983	2.1049	2.3316	2.5804	2.8531	3.1518	3.4785	3.8359	4.2262	4.6524	5.1173	7.4301	10.657	11.642	17.922
12	1.1268	1.2682	1.4258	1.6010	1.7959	2.0122	2.2522	2.5182	2.8127	3.1384	3.4985	3.8960	4.3345	4.8179	5.3503	5.9360	8.9161	13.215	14.552	23.298
13	1.1381	1.2936	1.4685	1.6651	1.8856	2.1329	2.4098	2.7196	3.0658	3.4523	3.8833	4.3635	4.8980	5.4924	6.1528	6.8858	10.699	16.386	18.190	30.288
14	1.1495	1.3195	1.5126	1.7317	1.9799	2.2609	2.5785	2.9372	3.3417	3.7975	4.3104	4.8871	5.5348	6.2613	7.0757	7.9875	12.839	20.319	22.737	39.374
15	1.1610	1.3459	1.5580	1.8009	2.0789	2.3966	2.7590	3.1722	3.6425	4.1772	4.7846	5.4736	6.2543	7.1379	8.1371	9.2655	15.407	25.196	28.422	51.186
16	1.1726	1.3728	1.6047	1.8730	2.1829	2.5404	2.9522	3.4259	3.9703	4.5950	5.3109	6.1304	7.0673	8.1372	9.3576	10.748	18.488	31.243	35.527	66.542
17	1.1843	1.4002	1.6528	1.9479	2.2920	2.6928	3.1588	3.7000	4.3276	5.0545	5.8951	6.8660	7.9861	9.2765	10.761	12.468	22.186	38.741	44.409	86.504
18	1.1961	1.4282	1.7024	2.0258	2.4066	2.8543	3.3799	3.9960	4.7171	5.5599	6.5436	7.6900	9.0243	10.575	12.375	14.463	26.623	48.039	55.511	112.455
19	1.2081	1.4568	1.7535	2.1068	2.5270	3.0256	3.6165	4.3157	5.1417	6.1159	7.2633	8.6128	10.197	12.056	14.232	16.777	31.948	59.568	69.389	146.192
20	1.2202	1.4859	1.8061	2.1911	2.6533	3.2071	3.8697	4.6610	5.6044	6.7275	8.0623	9.6463	11.523	13.743	16.367	19.461	38.338	73.864	86.736	190.050
21	1.2324	1.5157	1.8603	2.2788	2.7860	3.3996	4.1406	5.0338	6.1088	7.4002	8.9492	10.804	13.021	15.668	18.822	22.574	46.005	91.592	108.420	247.065
22	1.2447	1.5460	1.9161	2.3699	2.9253	3.6035	4.4304	5.4365	6.6586	8.1403	9.9336	12.100	14.714	17.861	21.645	26.186	55.206	113.574	135.525	321.184
23	1.2572	1.5769	1.9736	2.4647	3.0715	3.8197	4.7405	5.8715	7.2579	8.9543	11.026	13.552	16.627	20.362	24.891	30.376	66.247	140.831	169.407	417.539
24	1.2697	1.6084	2.0328	2.5633	3.2251	4.0489	5.0724	6.3412	7.9111	9.8497	12.239	15.179	18.788	23.212	28.625	35.236	79.497	174.631	211.758	542.801
25	1.2824	1.6406	2.0938	2.6658	3.3864	4.2919	5.4274	6.8485	8.6231	10.835	13.585	17.000	21.231	26.462	32.919	40.874	95.396	216.542	264.698	705.641
30	1.3478	1.8114	2.4273	3.2434	4.3219	5.7435	7.6123	10.063	13.268	17.449	22.892	29.960	39.116	50.950	66.212	85.850	237.376	634.820	807.794	•
35	1.4166	1.9999	2.8139	3.9461	5.5160	7.6861	10.677	14.785	20.414	28.102	38.575	52.800	72.069	98.100	133.176	180.314	590.668	•	•	•
36	1.4308	2.0399	2.8983	4.1039	5.7918	8.1473	11.424	15.968	22.251	30.913	42.818	59.136	81.437	111.834	153.152	209.164	708.802	•	•	•
40	1.4889	2.2080	3.2620	4.8010	7.0400	10.286	14.974	21.725	31.409	45.259	65.001	93.051	132.782	188.884	267.864	378.721	•	•	•	•
50	1.6446	2.6916	4.3839	7.1067	11.467	18.420	29.457	46.902	74.358	117.391	184.565	289.002	450.736	700.233	•	•	•	•	•	•

Table A-2

Future Value Interest Factors for One Rupee Annuity Compounded at *r* Percent for *n* periods $FVIFA_{rn} = [(1+r)^n - 1]/r$

Period	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%	11%	12%	13%	14%	15%	16%	20%	24%	25%	30%
1	1.0000	1.0200	1.0300	1.0400	1.0500	1.0600	1.0700	1.0800	1.0900	1.1000	1.1100	1.1200	1.1300	1.1400	1.1500	1.1600	1.2000	1.2400	1.2500	1.3000
2	2.0100	2.0200	2.0300	2.0400	2.0500	2.0600	2.0700	2.0800	2.0900	2.1000	2.1100	2.1200	2.1300	2.1400	2.1500	2.1600	2.2000	2.2400	2.2500	2.3000
3	3.0301	3.0604	3.0909	3.1216	3.1525	3.1836	3.2149	3.2464	3.2781	3.3100	3.3421	3.3744	3.4069	3.4396	3.4725	3.5056	3.6400	3.7776	3.8125	3.9900
4	4.0604	4.1216	4.1836	4.2465	4.3101	4.3746	4.4399	4.5061	4.5731	4.6410	4.7097	4.7793	4.8498	4.9211	4.9934	5.0665	5.3680	5.6842	5.7656	6.1870
5	5.1010	5.2040	5.3091	5.4163	5.5256	5.6371	5.7507	5.8666	5.9847	6.1051	6.2278	6.3528	6.4803	6.6101	6.7424	6.8771	7.4416	8.0484	8.2070	9.0431
6	6.1520	6.3081	6.4684	6.6330	6.8019	6.9753	7.1533	7.3359	7.5233	7.7156	7.9129	8.1152	8.3227	8.5355	8.7537	8.9775	9.9299	10.980	11.259	12.756
7	7.2135	7.4343	7.6625	7.8983	8.1420	8.3938	8.6540	8.9228	9.2004	9.4872	9.7833	10.089	10.405	10.730	11.067	11.414	12.916	14.615	15.073	17.583
8	8.2857	8.5830	8.8923	9.2142	9.5491	9.8975	10.260	10.637	11.028	11.436	11.859	12.300	12.757	13.233	13.727	14.240	16.499	19.123	19.842	23.858
9	9.3685	9.7546	10.159	10.583	11.027	11.491	11.978	12.488	13.021	13.579	14.164	14.776	15.416	16.085	16.786	17.519	20.799	24.712	25.802	32.015
10	10.462	10.950	11.464	12.006	12.578	13.181	13.816	14.487	15.193	15.937	16.722	17.549	18.420	19.337	20.304	21.321	25.959	31.643	33.253	42.619
11	11.567	12.169	12.808	13.486	14.207	14.972	15.784	16.645	17.560	18.531	19.561	20.655	21.814	23.045	24.349	25.733	32.150	40.238	42.566	56.405
12	12.683	13.412	14.192	15.026	15.917	16.870	17.888	18.977	20.141	21.384	22.713	24.133	25.650	27.271	29.002	30.850	39.581	50.895	54.208	74.327
13	13.809	14.680	15.618	16.627	17.713	18.882	20.141	21.495	22.953	24.523	26.212	28.029	29.985	32.089	34.352	36.786	48.497	64.110	68.760	97.625
14	14.947	15.974	17.086	18.292	19.599	21.015	22.550	24.215	26.019	27.975	30.095	32.393	34.883	37.581	40.505	43.672	59.196	80.496	86.949	127.913
15	16.097	17.293	18.599	20.024	21.579	23.276	25.129	27.152	29.361	31.772	34.405	37.280	40.417	43.842	47.580	51.660	72.035	100.815	109.687	167.286
16	17.258	18.639	20.157	21.825	23.657	25.673	27.888	30.324	33.003	35.950	39.190	42.753	46.672	50.980	55.717	60.925	87.442	126.011	138.109	218.472
17	18.430	20.012	21.762	23.698	25.840	28.213	30.840	33.750	36.974	40.545	44.501	48.884	53.739	59.118	65.075	71.673	105.931	157.253	173.636	285.014
18	19.615	21.412	23.414	25.645	28.132	30.906	33.999	37.450	41.301	45.599	50.396	55.750	61.725	68.394	75.836	84.141	128.117	195.994	218.045	371.518
19	20.811	22.841	25.117	27.671	30.539	33.760	37.379	41.446	46.018	51.159	56.939	63.440	70.749	78.969	88.212	98.603	154.740	244.033	273.556	483.973
20	22.019	24.297	26.870	29.778	33.066	36.786	40.995	45.762	51.160	57.275	64.203	72.052	80.947	91.025	102.444	115.380	186.688	303.601	342.945	630.165
21	23.239	25.783	28.676	31.969	35.719	39.993	44.865	50.423	56.765	64.002	72.265	81.699	92.470	104.768	118.810	134.841	225.026	377.465	429.681	820.215
22	24.472	27.299	30.537	34.248	38.505	43.392	49.006	55.457	62.873	71.403	81.214	92.503	105.491	120.436	137.632	157.415	271.031	469.056	538.101	*
23	25.716	28.845	32.453	36.618	41.430	46.996	53.436	60.893	69.532	79.543	91.148	104.603	120.205	138.297	159.276	183.601	326.237	582.630	673.626	*
24	26.973	30.422	34.426	39.083	44.502	50.816	58.177	66.765	76.790	88.497	102.174	118.155	136.831	158.659	184.168	213.978	392.484	723.461	843.033	*
25	28.243	32.030	36.459	41.646	47.727	54.865	63.249	73.106	84.701	98.347	114.413	133.334	155.620	181.871	212.793	249.214	471.981	898.092	*	*
30	34.785	40.568	47.575	56.085	66.439	79.058	94.461	113.283	136.308	164.494	199.021	241.333	293.199	356.787	434.745	530.312	*	*	*	*
35	41.660	49.994	60.462	73.652	90.320	111.435	138.237	172.317	215.711	271.024	341.590	431.663	546.681	693.573	881.170	*	*	*	*	*
36	43.077	51.994	63.276	77.598	95.836	119.121	148.913	187.102	236.125	299.127	380.164	484.463	618.749	791.673	*	*	*	*	*	*
40	48.886	60.402	75.401	95.026	120.800	154.762	199.635	259.057	337.882	442.593	581.826	767.091	*	*	*	*	*	*	*	*
50	64.463	84.579	112.797	152.667	209.348	290.336	406.529	573.770	815.084	*	*	*	*	*	*	*	*	*	*	*

Table A-3

Present Value Interest Factors for One Rupee Discounted at *r* Percent for *n* periods $PVIF_{rn} = 1/(1+r)^n$

Period	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%	11%	12%	13%	14%	15%	16%	20%	24%	25%	30%
1	0.9901	0.9804	0.9709	0.9615	0.9524	0.9434	0.9346	0.9259	0.9174	0.9091	0.9009	0.8929	0.8850	0.8772	0.8696	0.8621	0.8333	0.8065	0.8000	0.7692
2	0.9803	0.9612	0.9426	0.9246	0.9070	0.8900	0.8734	0.8573	0.8417	0.8264	0.8116	0.7972	0.7831	0.7695	0.7561	0.7432	0.6944	0.6504	0.6400	0.5917
3	0.9706	0.9423	0.9151	0.8890	0.8638	0.8396	0.8163	0.7938	0.7722	0.7513	0.7312	0.7118	0.6931	0.6750	0.6575	0.6407	0.5787	0.5245	0.5120	0.4552
4	0.9610	0.9238	0.8885	0.8548	0.8227	0.7921	0.7629	0.7350	0.7084	0.6830	0.6587	0.6355	0.6133	0.5921	0.5718	0.5523	0.4823	0.4230	0.4096	0.3501
5	0.9515	0.9057	0.8626	0.8219	0.7835	0.7473	0.7130	0.6806	0.6499	0.6209	0.5935	0.5674	0.5428	0.5194	0.4972	0.4761	0.4019	0.3411	0.3277	0.2693
6	0.9420	0.8880	0.8375	0.7903	0.7462	0.7050	0.6663	0.6302	0.5963	0.5645	0.5346	0.5066	0.4803	0.4556	0.4323	0.4104	0.3349	0.2751	0.2621	0.2072
7	0.9327	0.8706	0.8131	0.7599	0.7107	0.6651	0.6227	0.5835	0.5470	0.5132	0.4817	0.4523	0.4251	0.3996	0.3759	0.3538	0.2791	0.2218	0.2097	0.1594
8	0.9235	0.8535	0.7894	0.7307	0.6768	0.6274	0.5820	0.5403	0.5019	0.4665	0.4339	0.4039	0.3762	0.3506	0.3269	0.3050	0.2326	0.1789	0.1678	0.1226
9	0.9143	0.8368	0.7664	0.7026	0.6446	0.5919	0.5439	0.5002	0.4604	0.4241	0.3909	0.3606	0.3329	0.3075	0.2843	0.2630	0.1938	0.1443	0.1342	0.0943
10	0.9053	0.8203	0.7441	0.6756	0.6139	0.5584	0.5083	0.4632	0.4224	0.3855	0.3522	0.3220	0.2946	0.2697	0.2472	0.2267	0.1615	0.1164	0.1074	0.0725
11	0.8963	0.8043	0.7224	0.6496	0.5847	0.5268	0.4751	0.4289	0.3875	0.3505	0.3173	0.2875	0.2607	0.2366	0.2149	0.1954	0.1346	0.0938	0.0859	0.0558
12	0.8874	0.7885	0.7014	0.6246	0.5568	0.4970	0.4440	0.3971	0.3555	0.3186	0.2858	0.2567	0.2307	0.2076	0.1869	0.1685	0.1122	0.0757	0.0687	0.0429
13	0.8787	0.7730	0.6810	0.6006	0.5303	0.4688	0.4150	0.3677	0.3262	0.2897	0.2575	0.2292	0.2042	0.1821	0.1625	0.1452	0.0935	0.0610	0.0550	0.0330
14	0.8700	0.7579	0.6611	0.5775	0.5051	0.4423	0.3878	0.3405	0.2992	0.2633	0.2320	0.2046	0.1807	0.1597	0.1413	0.1252	0.0779	0.0492	0.0440	0.0254
15	0.8613	0.7430	0.6419	0.5553	0.4810	0.4173	0.3624	0.3152	0.2745	0.2394	0.2090	0.1827	0.1599	0.1401	0.1229	0.1079	0.0649	0.0397	0.0352	0.0195
16	0.8528	0.7284	0.6232	0.5339	0.4581	0.3936	0.3387	0.2919	0.2519	0.2176	0.1883	0.1631	0.1415	0.1229	0.1069	0.0930	0.0541	0.0320	0.0281	0.0150
17	0.8444	0.7142	0.6050	0.5134	0.4363	0.3714	0.3166	0.2703	0.2311	0.1978	0.1696	0.1456	0.1252	0.1078	0.0929	0.0802	0.0451	0.0258	0.0225	0.0116
18	0.8360	0.7002	0.5874	0.4936	0.4155	0.3503	0.2959	0.2502	0.2120	0.1799	0.1528	0.1300	0.1108	0.0946	0.0808	0.0691	0.0376	0.0208	0.0180	0.0089
19	0.8277	0.6864	0.5703	0.4746	0.3957	0.3305	0.2765	0.2317	0.1945	0.1635	0.1377	0.1161	0.0981	0.0829	0.0703	0.0596	0.0313	0.0168	0.0144	0.0068
20	0.8195	0.6730	0.5537	0.4564	0.3769	0.3118	0.2584	0.2145	0.1784	0.1486	0.1240	0.1037	0.0868	0.0728	0.0611	0.0514	0.0261	0.0135	0.0115	0.0053
21	0.8114	0.6598	0.5375	0.4388	0.3589	0.2942	0.2415	0.1987	0.1637	0.1351	0.1117	0.0926	0.0768	0.0638	0.0531	0.0443	0.0217	0.0109	0.0092	0.0040
22	0.8034	0.6468	0.5219	0.4220	0.3418	0.2775	0.2257	0.1839	0.1502	0.1228	0.1007	0.0826	0.0680	0.0560	0.0462	0.0382	0.0181	0.0088	0.0074	0.0031
23	0.7954	0.6342	0.5067	0.4057	0.3256	0.2618	0.2109	0.1703	0.1378	0.1117	0.0907	0.0738	0.0601	0.0491	0.0402	0.0329	0.0151	0.0071	0.0059	0.0024
24	0.7876	0.6217	0.4919	0.3901	0.3101	0.2470	0.1971	0.1577	0.1264	0.1015	0.0817	0.0659	0.0532	0.0431	0.0349	0.0284	0.0126	0.0057	0.0047	0.0018
25	0.7798	0.6095	0.4776	0.3751	0.2953	0.2330	0.1842	0.1460	0.1160	0.0923	0.0736	0.0588	0.0471	0.0378	0.0304	0.0245	0.0105	0.0046	0.0038	0.0014
30	0.7419	0.5521	0.4120	0.3083	0.2314	0.1741	0.1314	0.0994	0.0754	0.0573	0.0437	0.0334	0.0256	0.0196	0.0151	0.0116	0.0042	0.0016	0.0012	•
35	0.7059	0.5000	0.3554	0.2534	0.1813	0.1301	0.0937	0.0676	0.0490	0.0356	0.0259	0.0189	0.0139	0.0102	0.0075	0.0055	0.0017	0.0005	•	•
36	0.6989	0.4902	0.3450	0.2437	0.1727	0.1227	0.0875	0.0626	0.0449	0.0323	0.0234	0.0169	0.0123	0.0089	0.0065	0.0048	0.0014	•	•	•
40	0.6717	0.4529	0.3066	0.2083	0.1420	0.0972	0.0668	0.0460	0.0318	0.0221	0.0154	0.0107	0.0075	0.0053	0.0037	0.0026	0.0007	•	•	•
50	0.6080	0.3715	0.2281	0.1407	0.0872	0.0543	0.0339	0.0213	0.0134	0.0085	0.0054	0.0035	0.0022	0.0014	0.0009	0.0006	•	•	•	•

Table A-4

Present Value Interest Factors for One Rupee Annuity Discounted at *r* Percent for *n* periods $PVIFA_{rn} = [1 - 1/(1+r)^n]/r$

Period	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%	11%	12%	13%	14%	15%	16%	20%	24%	25%	30%
1	0.9901	0.9804	0.9709	0.9615	0.9524	0.9434	0.9346	0.9259	0.9174	0.9091	0.9009	0.8929	0.8850	0.8772	0.8696	0.8621	0.8333	0.8065	0.8000	0.7692
2	1.9704	1.9416	1.9135	1.8861	1.8594	1.8334	1.8080	1.7833	1.7591	1.7355	1.7125	1.6901	1.6681	1.6467	1.6257	1.6052	1.5278	1.4568	1.4400	1.3609
3	2.9410	2.8839	2.8286	2.7751	2.7232	2.6730	2.6243	2.5771	2.5313	2.4869	2.4437	2.4018	2.3612	2.3216	2.2832	2.2459	2.1065	1.9813	1.9520	1.8161
4	3.9020	3.8077	3.7171	3.6299	3.5460	3.4651	3.3872	3.3121	3.2397	3.1699	3.1024	3.0373	2.9745	2.9137	2.8550	2.7982	2.5887	2.4043	2.3616	2.1662
5	4.8534	4.7135	4.5797	4.4518	4.3295	4.2124	4.1002	3.9927	3.8897	3.7908	3.6959	3.6048	3.5172	3.4331	3.3522	3.2743	2.9906	2.7454	2.6893	2.4356
6	5.7955	5.6014	5.4172	5.2421	5.0757	4.9173	4.7665	4.6229	4.4859	4.3553	4.2305	4.1114	3.9975	3.8887	3.7845	3.6847	3.3255	3.0205	2.9514	2.6427
7	6.7282	6.4720	6.2303	6.0021	5.7864	5.5824	5.3893	5.2064	5.0330	4.8684	4.7122	4.5638	4.4226	4.2883	4.1604	4.0386	3.6046	3.2423	3.1611	2.8021
8	7.6517	7.3255	7.0197	6.7327	6.4632	6.2098	5.9713	5.7466	5.5348	5.3349	5.1461	4.9676	4.7988	4.6389	4.4873	4.3436	3.8372	3.4212	3.3289	2.9247
9	8.5660	8.1622	7.7861	7.4353	7.1078	6.8017	6.5152	6.2469	5.9952	5.7590	5.5370	5.3282	5.1317	4.9464	4.7716	4.6065	4.0310	3.5655	3.4631	3.0190
10	9.4713	8.9826	8.5302	8.1109	7.7217	7.3601	7.0236	6.7101	6.4177	6.1446	5.8892	5.6502	5.4262	5.2161	5.0188	4.8332	4.1925	3.6819	3.5705	3.0915
11	10.368	9.7868	9.2526	8.7605	8.3064	7.8869	7.4987	7.1390	6.8052	6.4951	6.2065	5.9377	5.6869	5.4527	5.2337	5.0286	4.3271	3.7757	3.6564	3.1473
12	11.255	10.575	9.9540	9.3851	8.8633	8.3838	7.9427	7.5361	7.1607	6.8137	6.4924	6.1944	5.9176	5.6603	5.4206	5.1971	4.4392	3.8514	3.7251	3.1903
13	12.134	11.348	10.635	9.9856	9.3936	8.8527	8.3577	7.9038	7.4869	7.1034	6.7499	6.4235	6.1218	5.8424	5.5831	5.3423	4.5327	3.9124	3.7801	3.2233
14	13.004	12.106	11.296	10.563	9.8986	9.2950	8.7455	8.2442	7.7862	7.3667	6.9819	6.6282	6.3025	6.0021	5.7245	5.4675	4.6106	3.9616	3.8241	3.2487
15	13.865	12.849	11.938	11.118	10.380	9.7122	9.1079	8.5595	8.0607	7.6061	7.1909	6.8109	6.4624	6.1422	5.8474	5.5755	4.6755	4.0013	3.8593	3.2682
16	14.718	13.578	12.561	11.652	10.838	10.106	9.4466	8.8514	8.3126	7.8237	7.3792	6.9740	6.6039	6.2651	5.9542	5.6685	4.7296	4.0333	3.8874	3.2832
17	15.562	14.292	13.166	12.166	11.274	10.477	9.7632	9.1216	8.5436	8.0216	7.5488	7.1196	6.7291	6.3729	6.0472	5.7487	4.7746	4.0591	3.9099	3.2948
18	16.398	14.992	13.754	12.659	11.690	10.828	10.059	9.3719	8.7556	8.2014	7.7016	7.2497	6.8399	6.4674	6.1280	5.8178	4.8122	4.0799	3.9279	3.3037
19	17.226	15.678	14.324	13.134	12.085	11.158	10.336	9.6036	8.9501	8.3649	7.8393	7.3658	6.9380	6.5504	6.1982	5.8775	4.8435	4.0967	3.9424	3.3105
20	18.046	16.351	14.877	13.590	12.462	11.470	10.594	9.8181	9.1285	8.5136	7.9633	7.4694	7.0248	6.6231	6.2593	5.9288	4.8696	4.1103	3.9539	3.3158
21	18.857	17.011	15.415	14.029	12.821	11.764	10.836	10.017	9.2922	8.6487	8.0751	7.5620	7.1016	6.6870	6.3125	5.9731	4.8913	4.1212	3.9631	3.3198
22	19.660	17.658	15.937	14.451	13.163	12.042	11.061	10.201	9.4424	8.7715	8.1757	7.6446	7.1695	6.7429	6.3587	6.0113	4.9094	4.1300	3.9705	3.3230
23	20.456	18.292	16.444	14.857	13.489	12.303	11.272	10.371	9.5802	8.8832	8.2664	7.7184	7.2297	6.7921	6.3988	6.0442	4.9245	4.1371	3.9764	3.3254
24	21.243	18.914	16.936	15.247	13.799	12.550	11.469	10.529	9.7066	8.9847	8.3481	7.7843	7.2829	6.8351	6.4338	6.0726	4.9371	4.1428	3.9811	3.3272
25	22.023	19.523	17.413	15.622	14.094	12.783	11.654	10.675	9.8226	9.0770	8.4217	7.8431	7.3300	6.8729	6.4641	6.0971	4.9476	4.1474	3.9849	3.3286
30	25.808	22.396	19.600	17.292	15.372	13.765	12.409	11.258	10.274	9.4269	8.6938	8.0552	7.4957	7.0027	6.5660	6.1772	4.9789	4.1601	3.9950	3.3321
35	29.409	24.999	21.487	18.665	16.374	14.498	12.948	11.655	10.567	9.6442	8.8552	8.1755	7.5856	7.0700	6.6166	6.2153	4.9915	4.1644	3.9984	3.3330
36	30.108	25.489	21.832	18.908	16.547	14.621	13.035	11.717	10.612	9.6765	8.8786	8.1924	7.5979	7.0790	6.6231	6.2201	4.9929	4.1649	3.9987	3.3331
40	32.835	27.355	23.115	19.793	17.159	15.046	13.332	11.925	10.757	9.7791	8.9511	8.2438	7.6344	7.1050	6.6418	6.2335	4.9966	4.1659	3.9995	3.3332
50	39.196	31.424	25.730	21.482	18.256	15.762	13.801	12.233	10.962	9.9148	9.0417	8.3045	7.6752	7.1327	6.6605	6.2463	4.9995	4.1666	3.9999	3.3333

APPENDIX 2
TABLE OF THE STUDENT'S *t*-DISTRIBUTION

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

INDEX

annuity
calculation, 146
first (direct) method, 144
second method (using future value annuity table), 145
third method (using excel), 145
summation of future value, 144
timeline, 143
arithmetic mean, 81
calculating, 84
calculation
first method (direct method), 83
second method (using excel), 83–84
demerits, 85–86
formula, 82
merits, 85

basic arithmetic operations in excel, 15
absolute cell reference, 22–23
addition, 16–18
autosum function, 23
cell reference, 22
count and countif function, 25–27
division, 21
MAX and MIN functions, 25
multiplication, 18–20
Pivot tables, 28–30
data filtering, 31–32
summarizing of values, 33–34
power function, 27–28
relative cell reference, 22
subtraction, 18
sumif function, 23–24

charts. *See* graphs
coefficient of variation, 119
calculation, 120
compound interest (CI), 139
calculation
first (direct) method, 141
second method, 141
third method (using excel), 142
future value interest factor table, 142, 143
impact of compounding, 140
confidence interval, 239
calculation
first method (the *t*-value approach), 240
second method (the *p*-value approach), 240–241
third method (confidence interval approach), 241
correlation, 185
negative correlation, 186
positive correlation, 186
zero correlation, 186
correlation coefficient, 186
perfect positive correlation, 187
positive correlation, 187
covariance, 181
calculation
first method (direct method), 183–184
second method (using excel), 184
interpretation of results, 185
negative covariance, 182
positive covariance, 182
zero covariance, 182

data
adding form, 39
Bombay Stock Exchange (BSE)
downloading historical data, 55–57
data entry form, 40
data series, 42–43
data sets, 41
defined, 37
Directorate of Economics and Statistics, 50
National Data Bank, 51
entry in excel, methods, 38–40
exporting data (sector-wise), 44
importing into excel, 40–41
Ministry of Statistics and Programme Implementation (MoSPI)
downloading of reports, 46
downloading of tables data, 48
microdata catalogue, 48–50
press release, 47
statistical publications, 47
website, 45
National Stock Exchange (NSE), 51

downloading historical reports, 52–55
primary and secondary, 38
qualitative data, 38
quantitative data, 38
UN data
downloading of data series, 62–63
website, 61
World Bank
data catalogue, 58
Development Data Group, 57
downloading data, 59–60
free and open access, 58
microdata, 58
descriptive statistics, 80
statistical measure, 81–101
arithmetic mean, 81–86
geometric mean, 97–98
harmonic mean, 99–101
median, 86–90
mode, 90–94
weighted arithmetic mean, 95–97

excel, defined, 1
excel 2019
adjustment of column width, 3
data analysis, 2
data manipulation, 3–4
filtering data through key commands, 10–13
finding number of rows and columns, 4
formatting decimal places, 4–5
formula bar, 3
freezing panes, 14
layout of, 2
name box, 3
quick access toolbar, 3
sorting data, 6–9
tabs, 3
unhide and hide rows and columns, 9
work area, 3
worksheet, 3
zoom in-zoom out, 3

future value interest factors, 247–248
future values of annuity. *See* annuity

geometric mean, 97
calculation, 98
graphs, 64
column chart, 68–69
formatting the axes, 67–68
histogram, 69–73
labelling, 66–67
line graph, 65–66
pie chart, 73–76
scatter plot, 76–77

harmonic mean, 99
calculation
first method (direct method), 100–101
second method (using excel), 101
hypothesis testing using confidence interval, 239
calculation
first method (the *t*-value approach), 240
second method (the *p*-value approach), 240–241
third method (confidence interval approach), 241
hypothesis testing using *p-value*, 236–237
t-distribution, 238
hypothesis testing using *t*-statistics
cut-off value, 235
degree of freedom (df), 236
formula, 232–236
null hypothesis, 235
one-tailed test, rejection areas, 234
two-tailed test, rejection areas, 234
type 1 and type 2 errors, 233

internal rate of return (IRR), 153
applications
endowment plans with moneyback at maturity, 157
moneyback policies with moneyback in between, 157–158
simple pension plan, 158–159
calculation, 154
first method, 155–156
second method (using excel), 156–157
demerits, 160–162
extended internal rate of return (XIRR), 159
merits, 160
two discount rate, 154

Karl Pearson's correlation coefficient, 187
calculation
first method (direct method), 188–189
fourth method (using data analysis in excel), 191
interpretation of results, 194
second method (using excel), 189
third method (using different function in excel), 191
formula, 188

kurtosis, 123
calculation, 127–128
excess, 126
histogram, 125
leptokurtic distribution, 125
mesokurtic distribution, 125
platykurtic distribution, 126

loan amortization
amount of loan after *n* periods, 169
arithmetic derivation of payment (PMT)/EMI in reducing balance method, 166–167
calculating payment, 167–168
calculation outstanding loan amount using PV method, 171
CUMIPMT, 174
CUMPRINC, 172–173
cumulative principal and interest amount using excel, 172–173
effective annual interest rate, 177–178
EMI using present value interest factor of annuity (PVIFA), 167–168
equated monthly instalment (EMI), 164
fixed interest and reducing balance, 165–166
interest, 164
IPMT function, 176
monthly EMI table, 170–171
payment or EMI using excel, 171–172
PPMT function, 174–176
principal amount, 164
table, 168–169

measure of dispersion, 105
coefficient of variation, 119–120
data analysis using excel, 128–131
kurtosis, 123–128
range, 107
calculation, 109–110
first method (direct method), 108
merits and demerits, 109
second method (excel), 108
scatter plot, 106–107
skewness, 120–123
standard deviation, 110
demerits, 115
first method (direct method), 112–113
merits, 114–115
rule of thumb, 111
second method (using excel), 113–114
variance, 115–119
measures of association, 181
median
calculation, 86, 89
demerits, 90
first method (direct method), 87–88
formula, 87
merits, 89
second method (using excel), 88–89
mode
calculation, 91–92, 93
first method (direct method), 90, 93
frequency distribution, 92–94
merits and demerits, 94
second method (graphical method), 93
second method (using excel), 90
third method (using excel), 94

net present value (NPV)
accept/reject criteria, 149–150
calculation, 150
cash flows, 149
demerits, 153
discount rate, 149
first method, 151
merits, 152–153
second method (using PVF table), 151
third method (using excel), 151–152

Pivot tables, 28–30
data filtering, 31–32
summarizing of values, 33–34
present value interest factors, 249–250
p-value, 236–237
t-distribution, 238

regression analysis, 206–207
calculation
first method, 222–223
first method (direct method), 212
interpretation of results, 224–225
second method of calculating parameters (using excel), 213–214
second method (using excel), 223–224
coefficient of determination, 220–225
estimation of the coefficients, 210–211
forecast the values of the dependent variable, 215–218
graphical representation of regression line, 208
interpretation of the estimated coefficients, 215

model, 207–209
OLS method, 219
residuals or errors, 219
slope coefficients, 208
constant slope, 209
negative slope, 209
positive slope, 209
three types of variations, graphical representation, 220

simple interest (SI)
calculation, 138–139
skewness
asymmetrical distribution, 120–121
calculation, 124
histogram, 123
Karl Pearson coefficient, 123
normal distribution, 120
second method (using excel), 123
symmetrical distribution, 120
slope coefficients, 208
constant slope, 209
negative slope, 209
positive slope, 209
Spearman correlation coefficient, 195–196
calculation, 199–201
first method (direct method), 197
hours of work, 202–203
interpretation of results, 199
linear relationships, 201
nonlinear relationship, 201
second method (using excel), 198
standard deviation, 110
demerits, 115
first method (direct method), 112–113
merits, 114–115
rule of thumb, 111
second method (using excel), 113–114
statistical measure, 81–101
arithmetic mean, 81–86
geometric mean, 97–98
harmonic mean, 99–101
median, 86–90
mode, 90–94
weighted arithmetic mean, 95–97

t-distribution
table, 250
time value of money, 137
t-statistics
cut-off value, 235
degree of freedom (df), 236
formula, 232–236
null hypothesis, 235
one-tailed test, rejection areas, 234
two-tailed test, rejection areas, 234
type 1 and type 2 errors, 233

variance
calculation, 118
demerits, 119
first method, 116–117
merits, 118
second method (using excel), 117

weighted arithmetic mean
first method, 96
formula, 95
putting values in formula, 95
second method, 96–97

52643