



U.P. Rajarshi Tandon Open
University, Prayagraj

DCESTAT – 106

Basic Knowledge of Statistical Softwares

Unit – 1 : Introduction to Statistical Software's

***Block: 1* Statistics with MS Office**

Unit – 2 : MS Office and its Components

Unit – 3 : Computations with MS Excel I

Unit – 4 : Computations with MS Excel II

Unit – 5 : Computations with MS Excel III

***Block: 2* Statistical Computations with R**

Unit – 6 : Basics of R

Unit – 7 : Statistical Analysis with R

Unit – 8 : Testing of Hypothesis with R

Course Design Committee

Dr. Ashutosh GuptaDirector, School of Sciences
U. P. Rajarshi Tandon Open University, Prayagraj**Chairman****Prof. Anup Chaturvedi**Ex. Head, Department of Statistics
University of Allahabad, Prayagraj**Member****Prof. S. Lalitha**Ex. Head, Department of Statistics
University of Allahabad, Prayagraj**Member****Prof. Himanshu Pandey**Department of Statistics
D. D. U. Gorakhpur University, Gorakhpur.**Member****Prof. Shruti**Professor, School of Sciences
U.P. Rajarshi Tandon Open University, Prayagraj**Member-Secretary**

Course Preparation Committee

Dr. Anuj Kumar SinghSchool of Sciences,
U. P. Rajarshi Tandon Open University, Prayagraj*(Unit – 2 to 8)***Writer****Dr. Anjali Saxena**Department of Mathematical Sciences and Computer Applications
Bundelkhand University, Jhansi*(Unit - 1)***Writer****Dr. P. S. Pundir**Department of Statistics
University of Allahabad, Prayagraj*(Unit – 2 to 8)***Editor****Prof. Shruti**School of Sciences,
U. P. Rajarshi Tandon Open University, Prayagraj*(Unit - 1)***Editor****Prof. Shruti**School of Sciences,
U. P. Rajarshi Tandon Open University, Prayagraj**Course Coordinator**

DCESTAT – 106/ DCESTAT – 106**BASIC KNOWLEDGE OF STATISTICAL SOFTWARES**

©UPRTOU

First Edition: July 2023**ISBN :**

©All Rights are reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from the Uttar Pradesh Rajarshi Tandon Open University, Prayagraj. Printed and Published by Col.. Vinay Kumar, Registrar, Uttar Pradesh Rajarshi Tandon Open University, 2023.

Printed By:

Blocks & Units Introduction

The present SLM on **Basic Knowledge of Statistical Software** consists of eight units with two blocks.

- The **Unit - 1 – Introduction to Statistical Software’s** About Statistical Software’s, its features and the steps for data analysis with related software’s Introduction to system software and application software.

The **Block - 1 – Statistics with MS Office**, is the first block, which is divided into six units.

- In **Unit – 2 – MS office and its Components**, the main emphasis on Statistical Software’s, its features and the steps for data analysis with related software’s Introduction to system software and application software, word processing software – Microsoft office Word, spread sheet (Interface of all the three-application software, file handling, editing, formatting and final output).
- In **Unit – 3 – Computations with MS Excel Part I**, we have focussed mainly on Microsoft Excel Software’s, its features and the steps for data analysis with related software’s Introduction to system software and application software, Microsoft office excel, (Interface of all the three-application software, file handling, editing, formatting and final output). Excel as data base software: cell referencing, concept of list, data sorting and filtering, manipulation of data, naming of cells
- In **Unit – 4 – Computations with MS Excel Part II**, we have focussed mainly on Microsoft Excel Software’s, its advance features and the steps for data analysis Excel using as statistical data analysis tool pack base software.
- In **Unit – 5 –Computations with MS Excel Part III**, we have focussed mainly on Microsoft Excel Software’s, its features and the steps for data analysis. Functions specifically Numeric/Mathematical functions, Statistical Functions, Logical Functions, lookup functions, Statistical Analysis using Excel – Descriptive Statistics, Curve fitting, correlation and regression analysis

The **Block - 2 – Statistical Computations with R**, is the second block with three units.

- In **Unit – 6 – Basics of R**, is being introduced the Terminology and basic Principles of R software, R Studio and R-Commander, creation of data files. Import Export of Data files, Transformation of Data
- In **Unit – 7 – Statistical Analysis with R** is discussed Statistical Analysis using R – Descriptive Statistics, Curve fitting, correlation and regression analysis, graphs
- In **Unit – 8 Testing of Hypothesis with R** has been introduced Studying of Statistical Analysis using R. Studying of general procedure of testing a hypothesis.

At the end of every block/unit the summary, self assessment questions are given.



॥ सरस्वती नः सुभगा मयस्कन्त ॥

**U.P.Rajarshi Tandon Open
University, Prayagraj**

DCESTAT – 106

Basic Knowledge of Statistical Softwares

Unit – 1 : Introduction to Statistical Softwares

Course Design Committee

Dr. Ashutosh Gupta

Director, School of Sciences

U. P. Rajarshi Tandon Open University, Prayagraj

Chairman**Prof. Anup Chaturvedi**

Ex. Head, Department of Statistics

University of Allahabad, Prayagraj

Member**Prof. S. Lalitha**

Ex. Head, Department of Statistics

University of Allahabad, Prayagraj

Member**Prof. Himanshu Pandey**

Department of Statistics

D. D. U. Gorakhpur University, Gorakhpur.

Member**Prof. Shruti**

Professor, School of Sciences

U.P. Rajarshi Tandon Open University, Prayagraj

Member-Secretary

Course Preparation Committee

Dr. Anjali Saxena

Department of Mathematical Sciences and Computer Applications

Bundelkhand University, Jhansi

Writer**Prof. Shruti**

School of Sciences,

U. P. Rajarshi Tandon Open University, Prayagraj

Editor**Prof. Shruti**

School of Sciences,

U. P. Rajarshi Tandon Open University, Prayagraj

Course Coordinator

**DCESTAT – 106/ DCESTAT – 106
SOFTWARES****BASIC KNOWLEDGE OF STATISTICAL**

©UPRTOU

First Edition: July 2023**ISBN :**

©All Rights are reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from the Uttar Pradesh Rajarshi Tandon Open University, Prayagraj. Printed and Published by Col. Vinay Kumar, Registrar, Uttar Pradesh Rajarshi Tandon Open University, 2023.

Printed By:

Unit Introduction

The present SLM on Basic Knowledge of Statistical software has many units. The first unit as:

The *Unit - 1 –An Introduction to Statistical Software*, is the first unit of present self-learning material, which describes about different statistical software with their advantages, disadvantages and applications.

The unit ends with the summary, self-assessment questions and further readings.

UNIT:1 AN INTRODUCTION TO STATISTICAL SOFTWARE

Structure:

- 1.1 Introduction
- 1.2 Objective
- 1.3 SPSS (Statistical Package for the Social Sciences)
 - 1.3.1 Graphical user Interphase
 - 1.3.2 Applications
 - 1.3.3 Advantages
 - 1.3.4 Limitations of using SPSS
- 1.4 STATA
 - 1.4.1 Applications
 - 1.4.2 Advantages
 - 1.4.3 Disadvantages
- 1.5 MATLAB
 - 1.5.1 Basic Operations of MATLAB
 - 1.5.2 Advantages
 - 1.5.3 Disadvantages
- 1.6 SYSTAT (Statistical Data Analysis and Scientific Visualization)
 - 1.6.1 Applications
 - 1.6.2 Advantages
 - 1.6.3 Disadvantages
- 1.7 Summery
- 1.8 Self-Assessment Exercise
- 1.9 References
- 1.10 Further Readings

1.1 Introduction

Statistical software are the best statistical analysis tools for scientists, researchers, business analysts, financial risk analyst, Operational research analyst, Investment analyst, managers, and marketers Actuary, data scientist during the entire analytics process, starting from planning data collection, data analysis, report creation, and deployment. Statistical analysis is the process of collecting and analyzing large volume of data in order to identify trends and develop valuable insight. Statistical analysts take raw data and find relationship between variables to reveal patterns and trends. There are two types of statistical analysis such as descriptive and inferential. Descriptive information provides data visualization in the form of graphs, tables, and charts in an understandable format to guide their decision making going. Inferential statistical analysis is used by businesses to inform company decisions and in scientific research to find relationship between variables. There are so many statistical tools such as SPSS, SYSTAT, STATA, MATLAB, SAS, Minitab, R etc.

1.2 Objective

The learner ought to be able to comprehend the following after finishing this unit

- The brief knowledge of SPSS
- The basic knowledge of STATA
- A basic understanding of MATLAB
- A basic understanding of SYSTAT

1.3 SPSS (Statistical Package for the Social Sciences)

SPSS is powerful best statistical software developed by IBM. It is a user-friendly interface and a robust set of features quickly extract actionable insights from data management to analysis and reporting. SPSS is a widely used program for statistical analysis in social science. SPSS Statistics places constraints on internal file structure, data types, data processing, and matching files, which together considerably simplify programming. All data processing occurs sequentially case-by-case through the file (dataset). Files can be matched one-to-one and one-to-many, but not many-to-many. In addition to that cases-by-variables structure and processing, there is a separate Matrix session where one can process data as matrices using matrix and linear

algebra operations. This may be sufficient for small datasets. Larger datasets such as statistical surveys are more often created in data entry software, or entered during computer-assisted personal interviewing, by scanning and using optical character recognition and optical mark recognition software, or by direct capture from online questionnaires. These datasets are then read into SPSS.

SPSS can read and write data from ASCII text files (including hierarchical files), other statistics packages, spreadsheets and databases. It can also read and write to external relational database tables via ODBC and SQL.

Statistical output is to a proprietary file format (*.spv file, supporting pivot tables). The proprietary output can be exported to text or Microsoft Word, PDF, Excel, and other formats. Alternatively, output can be captured as data (using the OMS command), as text, tab-delimited text, PDF, XLS, HTML, XML, SPSS dataset or a variety of graphic image formats (JPEG, PNG, BMP and EMF).

1.3.1 Graphical User Interphase

The graphical user interface has two views

1. Data view
2. Variable view

The '**Data View**' is like a spreadsheet where rows represents cases and column represents variables unlike spreadsheets, the data cells can only contain numbers or text, and formulas cannot be stored in these cells.

The '**Variable View**' displays the metadata dictionary, where each row represents a variable and shows the variable name, variable label, value label(s), print width, measurement type, and a variety of other characteristics.

. Cells in both views can be manually edited, defining the file structure and allowing data entry without using command syntax.

Variable view

Name:

It is a column field that accepts a unique ID that helps in sorting the data, such as name, gender, sex, educational qualification, designation, etc.

Label:

It gives the label and allows to add special characters.

Type:

It is useful to differentiate the type of data that is being used: numeric and text (or "string").

Width:

The length of the characters.

Decimal:

How to define the digits required after the decimal.

Value:

Enters the value here.

Missing:

Data that is unnecessary for analysis will be ignored.

Align:

It is for alignment-left or right.

Measure:

It measures the data that is being entered in the tools, such as cardinal, ordinal, and nominal.

1.3.2 Applications

It is used by market researchers, health researchers, survey companies, government, education researchers, marketing organizations, retailers, data miners, and others. SPSS is the best tool for market researchers. As there are tones of data generated by businesses, and scanning them manually is not possible to analyze them. It is possible to get accurate information about market trends, forecast, perceptual mapping, preference scaling, predictive analysis, statistical learning, and other advanced tools such as stratified, clustered, and multistage sampling which helps in the decision-making process. The retail industry depends on analytics for everything from initial stock planning to forecasting future trends. Customers are taking their decisions based on the brand's reviews online. Retail businesses generate a lot of data and it needs to be

collected, analyzed, and converted into actionable. The data, come up with customer preferences and give them an analysis of what makes customers turn from casual browsers into shoppers.

1.3.3 Advantages

1. It is reliable and fast statistical software
2. It is easy to draw useful tables and graphs
3. Effective data management
4. It is easy to start the software
5. It is useful for both quantitative and qualitative data
6. The less chances of errors
7. The easiest statistical tools to analyze data

1.3.4 Limitations of using SPSS

1. One of the biggest disadvantages of using SPSS is that we cannot use it to analyze a big data set.
2. Collected data using faulty or biased methods, then the resulting statistical analysis will not give the right answers

1.4 STATA

Stata is an easy-to-use powerful data analysis software package that features strong capabilities for statistical analysis. Stata offers a wide array of statistical tools that include both standard methods and newer, advanced methods, as new releases of Stata are distributed annually. The initial release of Stata was for the DOS operating system. Since then, versions of Stata have been released for systems running Unix variants like Linux distributions, Windows, and MacOS. All Stata files are platform-independent. Hundreds of commands have been added to Stata. Stata has included a graphical user interface based on Qt framework which uses menus and dialog boxes to give access to many built-in commands. The dataset can be viewed or edited in spreadsheet format. From version 11 on, other commands can be executed while the data browser or editor is opened. Data structure and storage Until the release of version 16 Stata could

only open a single dataset at any one time. Stata allows for flexibility with assigning data types to data. Its compress command automatically reassigns data to data types that take up less memory without loss of information. Stata utilizes integer storage types which occupy only one or two bytes rather than four, and single-precision (4bytes) rather than double-precision (8 bytes) is the default for floating point numbers. Stata's data format is always tabular in format. Stata refers to the columns of tabular data as variables. Data format compatibility Stata can import data in a variety of formats. This includes ASCII data formats (such as CSV or databank formats) and spreadsheet formats (including various Excel formats). Stata's proprietary file formats have changed over time, although not every Tata release includes a new dataset format. Every version of Stata can read all older dataset formats, and can write both the current and most recent previous dataset format, using the save old command. Thus, the current Stata release can always open datasets that were created with older versions, but older versions cannot read newer format datasets. Tata can read and write SAS XPORT format datasets natively, using the fdause and fdasave commands. Some other econometric applications, including gretl, can directly import Stat file formats.

1. Data management and manipulation
2. Data visualization

1.4.1 Applications

It is useful in model predictions, joint tests of coefficients or linear combination of statistics, marginal estimates, for graphical representation etc

GRAPHING DATA:

Twoway (scatter mpg weight) // Scatter plot showing relationship between mpg and weight

Twoway (scatter mpg weight), **by**(foreign, **total**) // Three graphs for domestic, foreign, and all cars

LINEAR REGRESSION:

generate wtsq = weight² // Create a new variable for weight squared

regress mpg weight wtsq foreign, **vce**(robust) // Linear regression of mpg on weight, wtsq, and foreign

predict mpghat // Create a new variable contained the predicted values of mpg
twoway (scatter mpg weight) (line mpghat weight, **sort**), **by**(foreign) // Graph data and fitted line

HISTOGRAMS:

Histograms plot distributions of variables by displaying counts of values that fall into various intervals of the variable. Use the option `normal` with histogram to overlay a theoretical normal density. Use the `width()` option to specify interval width.

BOXPLOTS:

Box plots are popular for displaying distributions of continuous variables. The median, the inter quartile range, multiple variables on the same plot and outliers can display through box plot.

SCATTER PLOTS:

Explore the relationship between 2 continuous variables with a scatter plot
The syntax `scatter var1 var2` will create a scatter plot with *var1* on the y-axis and *var2* on the x-axis.

BAR GRAPHS TO VISUALIZE FREQUENCIES

Bar graphs are often used to visualize frequencies. Graph bar produces bar graphs in Stata its syntax is a bit tricky to understand. For displays of frequencies (counts) of each level of a variable. syntax: `graph bar (count), over(variable)`.

ESTIMATING STATISTICS BASED ON A MODEL

Stata provides excellent support for estimating and testing additional statistics after a regression model has been run. Stata refers to these as “postestimation” commands, and they can be used after most regression models To see which commands can be issued as follow-ups to a model estimation command, help `model_command` postestimation, where *model_command* is a Stata model command e.g. for regress, help regress postestimation

DO-FILES ARE SCRIPTS OF COMMANDS:

1. Stata do-files are text files where users can store and run their commands
2. For reuse, rather than retyping the commands into the command window
3. Reproducibility
4. Easier debugging and changing commands

5. We recommend always using a do-file when using Stata
6. The file extension .do is used for do-files

STATA .DTA FILES:

1. Data files stored in Stata's format are known as .dta files
2. Remember that coding files are "do-files" and usually have a .do extension
3. Double clicking on a .dta file in Windows will open up the data in a new instance of Stata (not in the current instance)
4. Be careful of having many Statas open.

LOADING AND SAVING .DTA FILES:

1. The command use loads Stata .dta files
2. Usually these will be stored on a hard drive, but .dta files can also be loaded over the internet (using a web address)
3. Use the command save to save data in Stata's .dta format
4. The replace option will overwrite an existing file with the same name (without replace, Stata won't save if the file exists)
5. The extension .dta can be omitted when using use and save.

CLEARING MEMORY:

1. Stata will only hold one data set in memory at a time, memory must be cleared before new data can be loaded
2. The clear command removes the dataset from memory
3. Data import commands like use will often have a clear option which clears memory before loading the new dataset.

IMPORTING EXCEL DATA SETS:

1. Stata can read in data sets stored in many other formats
2. The command import excel is used to import Excel data
3. An Excel filename is required (with path, if not located in working directory) after the keyword using
4. Use the sheet() option to open a particular sheet
5. Use the firstrow option if variable names are on the first row of the Excel sheet.

PREPARING DATA FOR IMPORT:

To get data into Stata cleanly, make sure the data in your Excel file or .csv file have the following properties

1. Rectangular
2. Each column (variable) should have the same number of rows (observations)
3. No graphs, sums, or averages in the file
4. Missing data should be left as blank fields
5. Missing data codes like -999 are ok too (see command decode)
6. Variable names should contain only alphanumeric characters or _or .
7. Make as many variables numeric as possible
8. Many Stata commands will only accept numeric variables.

HELP FILES:

1. Precede a command name (and certain topic names) with help to access its help file for the summarizes command.

EXPLORE YOUR DATA BEFORE ANALYSIS:

1. Take the time to explore data set before embarking on analysis
2. Get to know your sample with quick summaries of variables
3. Demographics of subjects
4. Distributions of key variables for possible errors in variables.

DATA VISUALIZATION:

1. Data visualization is the representation of data in visual formats such as graphs
2. Graphs help us to gain information about the distributions of variables and relationships among variables quickly through visual inspection
3. Graphs can be used to explore your data, to familiarize yourself with distributions and associations in your data
4. Graphs can also be used to present the results of statistical analysis.

1.4.2 Advantages

1. Command syntax is very compact, saving time
2. Syntax is consistent across commands, so easier to learn.

3. Competitive with other software regarding variety of statistical tools.
4. Excellent documentation
5. Exceptionally strong support for Econometric models and methods
6. Complex survey data analysis tools.

1.4.3 Disadvantages

1. Limited to one dataset in memory at a time
2. Must open another instance of Stata to open another dataset
3. This won't be a problem for most users
4. Community is smaller than R (and maybe SAS)
5. less online help
6. fewer user-written extensions.

1.5 MATLAB

The MATLAB system consists of five main parts:

The MATLAB language.

This is a high-level matrix/array language with control flow statements, functions, data structures, input/output, and object-oriented programming features. It allows both "programming in the small" to rapidly create quick and dirty throw-away programs, and "programming in the large" to create complete large and complex application programs.

The MATLAB working environment.

This is the set of tools and facilities that you work with as the MATLAB user or programmer. It includes facilities for managing the variables in your workspace and importing and exporting data. It also includes tools for developing, managing, debugging, and profiling M-files, MATLAB's applications.

Handle Graphics.

This is the MATLAB graphics system. It includes high-level commands for two-dimensional and three-dimensional data visualization, image processing, animation, and presentation graphics. It also includes low-level commands that allow you to fully customize the

appearance of graphics as well as to build complete Graphical User Interfaces on your MATLAB applications.

The MATLAB mathematical function library.

This is a vast collection of computational algorithms ranging from elementary functions like sum, sine, cosine, and complex arithmetic, to more sophisticated functions like matrix inverse, matrix eigenvalues, Bessel functions, and fast Fourier transforms.

The MATLAB Application Program Interface (API).

This is a library that allows you to write C and Fortran programs that interact with MATLAB. It includes facilities for calling routines from MATLAB (dynamic linking), calling MATLAB as a computational engine, and for reading and writing MAT-files.

1.5.1 Basic Operations of MATLAB

Assignment to variable

```
>> x=3
```

```
x = 3
```

```
>> y=7
```

```
y = 7
```

Addition of variables

```
>> z=x+y
```

```
z = 10
```

```
>> z
```

```
z = 10
```

Subtraction of variables

```
>> y-x
```

```
ans = 4
```

Multiplication and Power of variables

```
>> x*y
```

```
ans = 21
```

```
>> x^2
```

```
ans = 9
```

```
>> y^2
ans = 49
>> y^x
ans = 343
```

Creating matrices

The basic data element in MATLAB is a matrix. A scalar in MATLAB is a 1x1 matrix, and a vector is a 1xn (or nx1) matrix.

Example: Create a 3x3 matrix A that has 1's in the first row, 2's in the second row, and 3's in the third row:

```
>> A = [1 1 1; 2 2 2; 3 3 3]
```

The semicolon is used here to separate rows in the matrix. MATLAB gives:

```
A=   1   1   1
     2   2   2
     3   3   3
```

If we don't want MATLAB to display the result of a command, put a semicolon at the end:

```
>> A = [1 1 1; 2 2 2; 3 3 3];
```

Matrix A has been created but MATLAB doesn't display it.

The semicolon is necessary when you're running long scripts and don't want everything written out to the screen! Suppose you want to access a particular element of matrix A:

```
>> A(1,2)
ans = 1
```

Suppose you want to access a particular row of A:

```
>> A(2,:)
ans = 2 2 2
```

MATLAB has several built-in matrices that can be useful. For example, zeros(n,n) makes an nxn matrix of zeros.

```
>> B = zeros(2,2)
B =   0   0
     0   0
```

A few other useful matrices are:

zeros – create a matrix of zeros
ones – create a matrix of ones
rand – create a matrix of random numbers
eye – create an identity matrix

Matrix operations An important thing to remember is that since MATLAB is matrix-based, the multiplication operator “*” denotes matrix multiplication. Therefore, A*B is not the same as multiplying each of the elements of A times the elements of B. However, you’ll probably find that at some point you want to do element-wise operations (array operations). In MATLAB you denote an array operator by playing a period in front of the operator. The difference between “*” and “.*” is demonstrated in this example:

```
>> A = [1 1 1; 2 2 2; 3 3 3];  
B = ones(3,3);  
A*B  
ans =  
3 3 3  
6 6 6  
9 9 9  
>> A.*B
```

```
ans =  
1 1 1  
2 2 2  
3 3 3
```

Other than the bit about matrix vs. array multiplication, the basic arithmetic operators in MATLAB work pretty much as expect. We can add (+), subtract (-), multiply (*), divide (/), and raise to some power (^).

MATLAB provides many useful functions for working with matrices. It also has many scalar functions that will work element-wise on matrices (e.g., the function sqrt(x) will take the square root of each element of the matrix x). Below is a brief list of useful functions. You’ll

find many, many more in the MATLAB help index, and also in the “Other Resources” listed at the end of this handout.

Useful matrix functions

A' – transpose of matrix A. Also `transpose(A)`.

`det(A)` – determinant of A

`eig(A)` – eigenvalues and eigenvectors

`inv(A)` – inverse of A

`svd(A)` – singular value decomposition

`norm(A)` – matrix or vector norm

`find(A)` – find indices of elements that are nonzero. Can also pass an expression to this function, e.g. `find(A > 1)` finds the indices of elements of A greater than 1.

A few useful math functions:

`sqrt(x)` – square root

`sin(x)` – sine function. See also `cos(x)`, `tan(x)`, etc.

`exp(x)` – exponential

`log(x)` – natural log `log10(x)` – common log

`abs(x)` – absolute value

`mod(x)` – modulus

`factorial(x)` – factorial function

`floor(x)` – round down. See also `ceil(x)`, `round(x)`.

`min(x)` – minimum elements of an array. See also `max(x)`.

`besselj(x)` – Bessel functions of first kind

MATLAB also has a few built-in constants, such as π (π) and i (imaginary number).

Descriptive Statistics:

MATLAB provides a number of commands that we can use to perform basic statistics tasks. When working with *descriptive statistics*, the math quantitatively describes the characteristics of a data collection, such as the largest and smallest values, the mean value of the items, and the average. This form of statistics is commonly used to summarize the data, thus making it easier to understand.

The following steps help you work through some of these tasks:

1. Type `w = 100 * randn(1, 100);` and press Enter.

This command produces 100 pseudo-random numbers that are uniformly distributed between the values 0 and 1. The numbers are then multiplied by 100 to bring them up to the integer values used in Steps 4 and 5.

2. Type `x = 100 * randn(1, 100);` and press Enter.

This command produces 100 pseudo-random numbers that are normally distributed. The numbers can be positive or negative, and multiplying by 100 doesn't necessarily ensure that the numbers are between -100 and 100 (as you see later in the procedure).

3. Type `y = randi(100, 1, 100);` and press Enter.

This command produces 100 pseudo-random integers that are uniformly distributed between the values of 0 and 100.

Of course, you can interact with the vectors in other ways. For example, you can use standard statistical functions on them. Here is a list of the functions.

Function	Usage	Example
<code>corrcoef()</code>	Determines the correlation coefficients between members of a matrix.	<code>corrcoef(AllVals)</code>
<code>cov()</code>	Determines the covariance matrix for either a vector or a matrix.	<code>cov(AllVals)</code>
<code>max()</code>	Specifies the largest element in a vector. When working with a matrix, you see the largest element in each row.	<code>max(w)</code>
<code>mean()</code>	Calculates the average or mean value of a vector. When working with a matrix, you see the mean for each row.	<code>mean(w)</code>
<code>median()</code>	Calculates the median value of a vector. When working with a matrix, you see the median for each row.	<code>median(w)</code>
<code>min()</code>	Specifies the smallest element in a vector. When working with a matrix, you see the smallest element in each row.	<code>min(w)</code>
<code>mode()</code>	Determines the most frequent value in a vector. When working with a matrix, you see the most frequent value for each row.	<code>mode(w)</code>
<code>std()</code>	Calculates the standard deviation for a vector. When working	<code>std(w)</code>

	with a matrix, you see the standard deviation for each row.	
var()	Determines the variance of a vector. When working with a matrix, you see the variance for each row.	var(w)

Example 1 — Calculating Maximum, Mean, and Standard Deviation

This example shows how to use MATLAB functions to calculate the maximum, mean, and standard deviation values for a 24-by-3 matrix called count. MATLAB computes these statistics independently for each column in the matrix.

```
>> % Load the sample data
load count.dat
% Find the maximum value in each column
mx = max(count)
% Calculate the mean of each column
mu = mean(count)
% Calculate the standard deviation of each column
sigma = std(count)
mx =
    114    145    257
mu =
    32.0000    46.5417    65.5833
sigma =
    25.3703    41.4057    68.0281
```

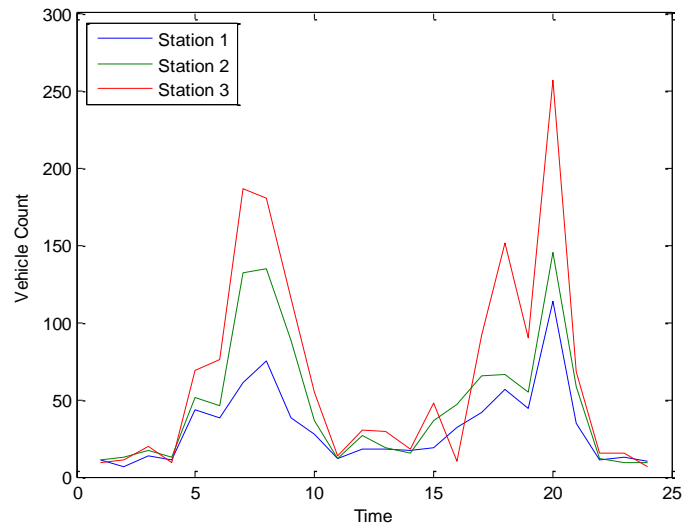
Calculating and Plotting Descriptive Statistics

1. Load and plot the data:

```
>> load count.dat
[n,p] = size(count);
% Define the x-values
t = 1:n;
% Plot the data and annotate the graph
plot(t,count)
legend('Station 1','Station 2','Station 3','Location','northwest')
```

```
xlabel('Time')
```

```
ylabel('Vehicle Count')
```



Ex.

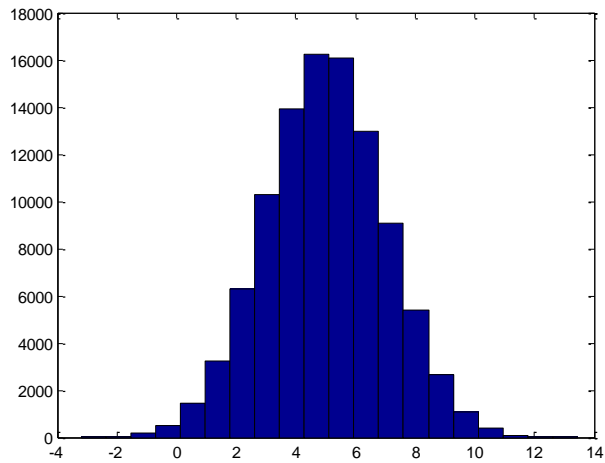
```
>> k=100000;
```

```
sigma =2;
```

```
mu=5;
```

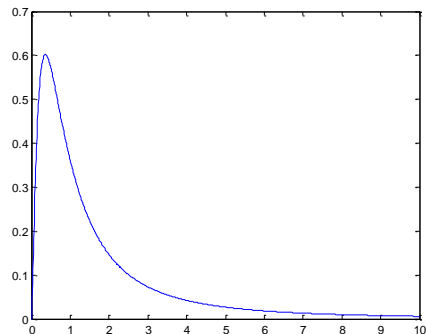
```
y = normrnd( mu , sigma , k , 1 );
```

```
hist( y , 20 );
```



Example and Plot. The most common application of the F distribution is in standard tests of hypotheses in analysis of variance and regression. The plot shows that the F distribution exists on the positive real numbers and is skewed to the right.

```
x = 0:0.01:10;  
y = fpdf(x,5,3);  
plot(x,y)
```



Example and Plot. The following commands generate a plot of the noncentral F pdf.

```
x = (0.01:0.1:10.01)';  
p1 = ncfpdf(x,5,20,10);  
p = fpdf(x,5,20);  
plot(x,p,'- -',x,p1,'-')
```

Correlation Coefficients

Compute the correlation coefficients for a matrix with two normally distributed, random columns and one column that is defined in terms of another. Since the third column of A is a multiple of the second, these two variables are directly correlated, thus the correlation coefficient in the (2,3) and (3,2) entries of R is 1.

Ex.

```
>> x = randn(6,1);  
y = randn(6,1);  
A = [x y 2*y+3];  
R = corrcoef(A)  
R =
```

```
1.0000  0.4973  0.4973
```

```
0.4973  1.0000  1.0000
0.4973  1.0000  1.0000
```

Compute the correlation coefficient matrix between two normally distributed, random vectors of 10 observations each.

```
>> A = randn(10,1);
```

```
B = randn(10,1);
```

```
R = corrcoef(A,B)
```

```
R =
```

```
1.0000  0.3907
0.3907  1.0000
```

Regression

Linear regression

Syntax

```
[r,m,b] = regression(t,y)
```

```
[r,m,b] = regression(t,y,'one')
```

Description

[r,m,b] = regression(t,y) takes these arguments,

t Target matrix or cell array data with a total of N matrix rows

y Output matrix or cell array data of the same size

and returns these outputs,

r Regression values for each of the N matrix rows

m Slope of regression fit for each of the N matrix rows

b Offset of regression fit for each of the N matrix rows

[r,m,b] = regression(t,y,'one') combines all matrix rows before regressing, and returns single scalar regression, slope, and offset values.

Examples

Train a feedforward network, then calculate and plot the regression between its targets and outputs.

```
>> [x,t] = simplefit_dataset;
```

```
net = feedforwardnet(20);
```

```
net = train(net,x,t);  
y = net(x);  
[r,m,b] = regression(t,y)  
plotregression(t,y)
```

Command Summary

The command

```
>> help
```

will give a list of categories for which help is available
(e.g.matlab/generalcovers the topics listed in Table 3.

Further information regarding the commands listed in
this section may then be obtained by using:

```
>> help topic
```

try, for example,

```
>> help help
```

Managing commands and functions.
help On-line documentation. doc Load hypertext documentation. what Directory listing of M-, MAT-and MEX-Files. type List M-File. lookfor Keyword search through the demo Run demos.
Working with les and the operating system.
cd Change current working directory. dir Directory listing. delete Delete File. ! Execute operating system command. unix Execute operating system command & return result. diary Save text of MATLAB session.
Controlling the command window.
cedit Set command line edit/recall facility parameters.

clc Clear command window.

home Send cursor home.

format Set output format.

echo Echo commands inside script files.

more Control paged output in command window.

Quitting from MATLAB.

quit Terminate MATLAB.

Matrix analysis.

cond Matrix condition number.

norm Matrix or vector norm.

rcondLINPACK reciprocal condition estimator.

rank Number of linearly independent rows or columns.

det Determinant.

trace Sum of diagonal elements.

null Null space.

orth Orthogonalization.

rref Reduced row echelon form.

Linear equations.

nand/ Linear equation solution; use \help slash".

chol Cholesky factorization.

lu Factors from Gaussian elimination.

inv Matrix inverse.

qr Orthogonal- triangular decomposition.

qrdelete Delete a column from the QR factorization.

qrinsert Insert a column in the QR factorization.

nls Non{negative least- squares.

pinv Pseudoinverse.

lsq Least squares in the presence of known covariance.

Eigenvalues and singular values.

eig Eigenvalues and eigenvectors.

poly Characteristic polynomial.
polyeig Polynomial eigenvalue problem.
hess Hessenberg form.
qz Generalized eigenvalues.
rsf2csf Real block diagonal form to complex diagonal form.
cdf2rdf Complex diagonal form to realblock diagonal form.
schur Schur decomposition.
balance Diagonal scaling to improve eigenvalue accuracy.
svd Singular value decomposition.

Matrix functions.

expm Matrix exponential.
expm1 M- File implementation of expm.
expm2 Matrix exponential via Taylor se-ries.
expm3 Matrix exponential via eigenval-ues and eigenvectors.
logm Matrix logarithm.
sqrtm Matrix square root.
funm Evaluate general matrix function.

Graphics & plotting.

figure Create Figure (graph window).
clf Clear current gure.
close Close gure.
subplot Create axes in tiled positions.
AxisControl axis scaling and appear-ance.
hold Hold current graph.
figure Create gure window.
text Create text.
print Save graph to le.
plot Linear plot.
loglog Log-log scale plot.
semilogx Semi-log scale plot

semilogy Semi-log scale plot.

1.5.2 Advantages

1. With MATLAB compiler we may distribute MATLAB programs as standalone applications and online apps
2. Accesses big data applications and add-ins Microsoft Excel
3. MATLAB includes a big predefined functions library with tried-and-true options.
4. MATLAB offers a great degree of platform independence supported by Linux, other UNIX version, Windows 2000/XP etc.
5. It is easy to use. The plots and images can be displayed on any type of graphical output device made available by the machine
6. An excellent technical data visualization tool

1.5.3 Disadvantages

1. It is an interpreted language.
2. The MATLAB is slower than compiled language.
3. It is an expensive, efficient for organization not for individual.
4. It requires fast computer with sufficient amount of memory.
5. It is difficult to develop real time applications using MATLAB
6. less online help
7. It was built with the specific goal of enabling rapid prototyping and analysis of scientific data, it is not well suited for the task.

1.6 SYSTAT - Statistical Data Analysis and Scientific Visualization

SYSTAT is a software package for statistical analysis, data visualization, and data management. The software offers a wide range of features, including descriptive statistics, inferential statistics, hypothesis testing, and advanced data visualization tools. It is used by researchers, scientists, and data analysts in various fields, including biology, engineering, social sciences, and business.

SYSTAT is a statistics and statistical graphics software package, developed by Leland Wilkinson in the late 1970s, who was at the time an assistant professor of psychology at the University of Illinois at Chicago.

The user interface of SYSTAT is organized into three spaces:

1. View space
2. Work space
3. Command space

View space has the following tabs

Output Editor: Graphs and statistical results appear in the Output Editor. We can edit, print and save the output displayed in the Output Editor.

Data Editor: The Data Editor displays the data in a row-by-column format. Each row is a case and each column is a variable. we can enter, edit, view, and save data in the Data Editor.

Graph Editor: we can edit and save graphs in the Graph Editor.

Start page: Start page window appears in View space as we open SYSTAT. It has five sub-windows.

1. Recent Files
2. Tips
3. Themes
4. Manuals
5. Scratchpad

Workspace has the following tabs

Output Organizer: The Output Organizer tab helps primarily to navigate through the results of statistical analysis. We can quickly navigate to specific portions of output without having to use the Output Editor scrollbars.

Examples: The Examples tab enables us to run the examples given in the user manual with just a click of mouse. The SYSTAT examples tree consists of folders corresponding to different volumes of user manual and nodes. We can also add own example.

Dynamic Explorer: The Dynamic Explorer can be used to rotate 3-D graphs, apply power transformations to values on one or more axes, and change the confidence intervals, ellipses, and kernels in scatter plots.

Command space has the following tabs

Interactive: In the Interactive tab, we can enter commands at the command prompt (>) and issue them by pressing the Enter key.

Untitled: The Untitled tab enables us to run the commands in the batch mode. we can open, edit, submit and save SYSTAT command file (.syc or .cmd)

Log: In the Log tab, we can view the record of the commands issued during the SYSTAT session (through Dialog or in the Interactive mode).

SYSTAT Data, Command and Output files

Data files. We can save data files with (.SYZ) extension.

Command files. A command file is a text file that contains SYSTAT commands. Saving our analyses in a command file allows us to repeat them at a later date. These files are saved with (.SYC) extension.

Output files. SYSTAT displays statistical and graphical output in the output Editor. We can save the output in (.SYO), Rich Text format (.RTF) and Hypertext Markup Language format (*.HTM).

1.6.1 Applications

1. Archeology: Evolution of skull dimension
2. Environmental Sciences: TCE contamination
3. Geology: Estimation of uranium Reserves from groundwater
4. Epidemiology: Tuberculosis incidents

1.6.2 Advantages

1. SYSTAT offers many major performance enhancements for speed and increased ease of use.
2. Simply pointing and clicking the mouse can accomplish most tasks
3. SYSTAT also offers a huge data worksheet for powerful data handling.
4. Windows interface and flexible command language are designed to make research more efficient.

1.6.3 Disadvantages

1. A restricted number of instances, variables, and analysis aspects can be handled by statistical software like SYSTAT.
2. The software provides functions and libraries to read and process data from many different sources, including but not limited to ASCII files, binary files, spreadsheets, and databases.
3. The data can be loaded into SYSTAT as long as the data structure and format are known. There are no restrictions on the size of data other than those caused by the hardware of the computer running the software.

1.7 Summery

Aspects of the scientific field of statistics include data organization, analysis, and extrapolation from samples to the total population. Data collection and presentation are made simpler by statistical analysis software so that researchers can understand it and act on it. Any person can use statistical analysis tools to visualize data and analyze data using mathematical models like regression analysis, multivariate analysis, and statistical simulation. The current state of the market can be accurately determined. Faulty findings may result from the improper use of statistical methodology. Best statistical software is produced by having a proper understanding of the fundamental statistical techniques.

1.8 Self-Assessment Exercise

1. Explain the purpose of statistical software.
2. Name four statistical software.
3. Explain the applications of Systat
4. Is SAS statistical software? If (yes/no), justify your answer.
5. What do you understand by data analysis?

1.9 References

- <https://www.minitab.com/en-us/products/minitab/>

- http://apps.iasri.res.in/ebook/EBADAT/1Computer%20Usage%20and%20Statistical%20Software%20Packages/9-SYSTAT%20TUTORIAL_03feb.pdf
- <https://www.spss-tutorials.com/spss-what-is-it/>
- <https://en.wikipedia.org/wiki/SPSS>
- <https://www.stata.com/>
- <https://en.wikipedia.org/wiki/Stata>
- <https://www.mathworks.com/products/statistics.html>
- <https://sites.google.com/a/nyu.edu/statistical-software-guide/matlab>

1.10 Further Readings

- <https://www.capterra.com/statistical-analysis-software/>
- <https://www.coursera.org/in/articles/data-analysis-software>
- <https://www.publichealthnotes.com/different-types-of-statistical-software/>
- <https://ucsd.libguides.com/data-statistics/matlab>



U.P. Rajarshi Tandon Open
University, Prayagraj

DCESTAT – 106

Basic Knowledge of Statistical Softwares

Block: 1 Statistics with MS Office

Unit – 2 : MS Office and its Components

Unit – 3 : Computations with MS Excel I

Unit – 4 : Computations with MS Excel II

Unit – 5 : Computations with MS Excel III

Course Design Committee

Dr. Ashutosh Gupta

Director, School of Sciences

U. P. Rajarshi Tandon Open University, Prayagraj

Chairman

Prof. Anup Chaturvedi

Ex. Head, Department of Statistics

University of Allahabad, Prayagraj

Member

Prof. S. Lalitha

Ex. Head, Department of Statistics

University of Allahabad, Prayagraj

Member

Prof. Himanshu Pandey

Department of Statistics

D. D. U. Gorakhpur University, Gorakhpur.

Member

Prof. Shruti

Professor, School of Sciences

U.P. Rajarshi Tandon Open University, Prayagraj

Member-Secretary

Course Preparation Committee

Dr. Anuj Kumar Singh

School of Sciences,

U. P. Rajarshi Tandon Open University, Prayagraj

Writer

Dr. P. S. Pundir

Department of Statistics

University of Allahabad, Prayagraj

Editor

Prof. Shruti

School of Sciences,

U. P. Rajarshi Tandon Open University, Prayagraj

Course Coordinator

DCESTAT – 106/ DCESTAT – 106 BASIC KNOWLEDGE OF STATISTICAL SOFTWARES

©UPRTOU

First Edition: July 2023

ISBN :

©All Rights are reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from the Uttar Pradesh Rajarshi Tandon Open University, Prayagraj. Printed and Published by Col.. Vinay Kumar, Registrar, Uttar Pradesh Rajarshi Tandon Open University, 2023.

Printed By:

Block & Units Introduction

The **Block - 1 – Statistics with MS Office**, is the first block, which is divided into six units.

- In **Unit – 2 – MS office and its Components**, the main emphasis on Statistical Software's, its features and the steps for data analysis with related software's Introduction to system software and application software, word processing software – Microsoft office Word, spread sheet (Interface of all the three-application software, file handling, editing, formatting and final output).
- In **Unit – 3 – Computations with MS Excel Part I**, we have focussed mainly on Microsoft Excel Software's, its features and the steps for data analysis with related software's Introduction to system software and application software, Microsoft office excel, (Interface of all the three-application software, file handling, editing, formatting and final output). Excel as data base software: cell referencing, concept of list, data sorting and filtering, manipulation of data, naming of cells
- In **Unit – 4 – Computations with MS Excel Part II**, we have focussed mainly on Microsoft Excel Software's, its advance features and the steps for data analysis Excel using as statistical data analysis tool pack base software.
- In **Unit – 5 –Computations with MS Excel Part III**, we have focussed mainly on Microsoft Excel Software's, its features and the steps for data analysis. Functions specifically Numeric/Mathematical functions, Statistical Functions, Logical Functions, lookup functions, Statistical Analysis using Excel – Descriptive Statistics, Curve fitting, correlation and regression analysis

At the end of every unit the summary, self-assessment questions are given.

Unit -2: MS Office and its Components

Structure

- 2.1 Introduction
- 2.2 Objectives
- 2.3 Basic of Microsoft Word
- 2.4 Typing Features of Microsoft Word
 - 2.4.1 Making a New Blank Document
 - 2.4.2 Editing Documents
 - 2.4.3 Basics of Editing
 - 2.4.4 Pointer Description
- 2.5 Formatting Toolbar of Microsoft Word
 - 2.5.1 World Wrap
 - 2.5.2 Macintosh
 - 2.5.3 Arrow Keys
 - 2.5.4 Using the Undo Command
 - 2.5.5 Using the Undo Command: Menu Option
 - 2.5.6 Using the Undo Command: Toolbar Option
 - 2.5.7 Selecting Text: Lines
 - 2.5.8 Selecting Text: Specific Areas
 - 2.5.9 Using Drag and Drop
 - 2.5.10 Using Drag and Drop: Moving Text
 - 2.5.11 Deleting Text
 - 2.5.12 Deleting Text: Characters
 - 2.5.13 Deleting Text: Type Over
 - 2.5.14 Deleting Text: A Line or Block of Text
- 2.6 Formatting Toolbar
 - 2.6.1 Style
 - 2.6.2 Font

- 2.6.3 Font Size
- 2.6.4 Saving Documents
- 2.6.5 Inserting an Image
- 2.6.6 Create Bulleted and Numbered lists
- 2.6.7 Create a Table
- 2.6.8 Type Special Characters
- 2.6.9 Create a template
- 2.6.10 Hyphenation
- 2.6.11 Headers and Footers
- 2.6.12 Insert Graphs and Pictures in documents
- 2.6.13 Insert Watermark in the document
- 2.7 Mail Merge of Microsoft Word
- 2.8 Summery
- 2.9 Self-Assessment Questions
- 2.10 References

2.1 Introduction

Microsoft Word is an essential tool for the creation of documents. Currently, it is one of the most widely used word processing applications. Microsoft Word is fairly simple software to use for performing simple tasks. There were "advanced features of Microsoft Word which can be used for accomplishing complicated tasks.

2.2 Objectives

After studying this unit, you will be able to understand the following objectives:

- Studying of Open, Save and perform other simple operations on documents.
- Study of Creation documents that include text, graphics, tables, clip art, etc.
- Study creation a variety of documents ranging from simple notes and memos to complex multi-column reports with tables, graphics, table of contents and an index.
- Study of mail-merge documents and labels.

2.3 Basic of Microsoft Word

Opening Microsoft Word To run Microsoft Word on your computer, perform the following instructions: "Start" » "Programs" » "Microsoft Office" » "Microsoft Word". If there is an icon of Microsoft Word available on your desktop (shaped like a square with a "W" in the middle), you can open the program by double-clicking it, as well.

2.4 Typing Features of Microsoft Word

The following typing features of Microsoft word:

2.4.1 Making a New Blank Document

When Microsoft Word is opened, a new blank document should automatically open, if not, then you can begin a new blank document in a variety of ways. First, find the "New Blank Document" icon, which looks like a blank sheet of paper, located underneath the menu bar in Microsoft Word in what is called the "standard toolbar." Click on the icon to bring up a new blank document. Also, you can go to the menu bar and select File » New ... (Shortcut: Ctrl+N). To begin typing, just click the cursor anywhere within the new blank document.

2.4.2 Editing Documents

We have created a Microsoft Word document and typed some text, we may want to edit your work by adding, moving, or deleting text. This document covers the Undo command and the basic editing functions of selecting, moving, and deleting text.

- Basics of editing
- Using the undo command
- Selecting text
- Making multiple selections
- Using Drag and Drop features
- Deleting text

2.4.3 Basics of Editing

The blinking vertical line located in the window is the insertion point. As you type, keyed text will appear to the left of the insertion point. If you move the mouse, it is the pointer that moves on screen. The pointer can appear in several ways: Four of the most common shapes are discussed below.

2.4.4 Pointer Description

When the pointer moves over the page, it looks like an I-beam. When you click the mouse button, the insertion point is placed to the left of the I-beam pointer. When the pointer moves over specific formatting areas, the following icons appear under the insertion point: align left, align right, center, left indent, left text wrap, and right text wrap. The text you type will appear in the format of the corresponding icon. When the pointer moves over the *Menu* bar or the toolbars, it takes the shape of an arrow pointing up and to the left. Clicking the mouse button once over a button or menu option will select that option. When the pointer moves past the left margin of the text on the page, it takes the shape of an arrow pointing up and to the right. Clicking the mouse button at this point will select that line of text. To select the entire paragraph, double click.

2.5 Formatting Toolbar of Microsoft Word

When font designers create fonts, they often add designs for special features. Selected Open Type fonts include some or all of the features below and we can check with the font provider for details. With those fonts, these features are available for us to apply to our text for to make it more polished and easier to read. For example, the fonts in the Microsoft Clear Type Collection Calibri, Cambria, Candara, Consolas, Constantia and Corbel – contain various Open Type including small caps, ligatures, number forms and number spacing. Gabriola, a newer font originally released with Windows 7 includes even richer Open Type feature support, including extensive use of stylistic sets.

To apply Open Type features, do the following:

- (i) On the Home tab, click the Font Dialog Box Launcher.
- (ii) Click the Advanced tab.
- (iii) Under Open Type Features, select the desired options.

We can also disable these Open Type features entirely by opening Word Options, selecting the

Advanced tab and checking Disable Open Type Font Formatting Features under the Layout Options those are right at the bottom.

2.5.1 World Wrap

Text is wrapped at the end of each line and continues on the next line; you do not have to press the [Enter] or [Return] keys as on a typewriter. Delete Character Windows: The [Backspace] key moves the insertion point to the left one space at a time, eliminating text or space. The [Delete] key moves the insertion point to the right one space at a time, eliminating text or space

2.5.2 Macintosh

The [delete] key moves the insertion point to the left one space at a time, eliminating text or space. The [del] key moves the insertion point to the right one space- at a time, eliminating text or space.

2.5.3 Arrow Keys

The arrow keys move the insertion point up or down one line at a time and left or right one space at a time. The arrow keys do not delete. They allow you to position the insertion point exactly where you want it. This is especially helpful for inserting text into different parts of your document.

2.5.4 Using the Undo Command

If text was accidentally deleted or if there was some type of editing mistake, you may be able to reverse the last action using the *Undo* command. If your last action cannot be reversed, the option will read Cannot Undo.

Using the Undo Command: Keyboard Option Microsoft Word

Windows: Press [Ctrl] + [Z]

Macintosh: Press [command] + [Z]

Your last action is reversed.

2.5.5 Using the Undo Command: Menu Option

From the Edit menu, select Undo. The Undo menu option will read Undo Typing, Undo Formatting, or Undo X (where X represents your last action), Your last action is reversed.

2.5.6 Using the Undo Command: Toolbar Option

WARNING: When you undo an action, you also undo all actions above it in the list. Selecting Text is a basic editing skill used in Microsoft Word. In order to format text, it must be selected. Once your text is selected, you can also/cut, copy, or paste your text. For more information, refer to Cutting, Copying, and Pasting Text. For example, by selecting specific text you can change the font size of only the selected text. Several methods are available for selecting text. Use the option that is most convenient for you or use the technique that best fits your task. Keyboard shortcuts can also be used to select text.

2.5.7 Selecting Text: Lines

1. Place the insertion points to the left side of the document until it turns into an arrow
2. To select a single line of text, click the mouse button once.

To select multiple lines of text, click and drag to select the desired lines. The line(s) of text is selected.

2.5.8 Selecting Text: Specific Areas

If the text is near the left margin, it may be easier to start by selecting the last letter of the desired text.

1. Place the I-beam to the left of the beginning of the desired text.
2. Click and hold the mouse button.
3. Drag the mouse over the text to be selected.
4. Release the mouse button.

2.5.9 Using Drag and Drop

Drag and Drop is another option for moving blocks of text. This option is best for moving text short distances. Because you use the mouse, Drag and Drop text is never placed on the Clipboard. As you are dragging the text, a gray insertion point appears. When you let go of the mouse button, the text drops in that location.

2.5.10 Using Drag and Drop: Moving Text

1. Select the text to be moved

NOTE: For more information, refer to Selecting Text.

2. Click on the text and hold the mouse button
3. Drag the text to the desired location

HINT: The insertion line will indicate where the text will be dropped.

4. To drop the text, release the mouse button

The text is moved. HINT: If you dropped the text in the wrong spot, refer to Using the Undo ' Command.

2.5.11 Deleting Text

You can delete anything from a few characters to several pages of text. You can also restore deleted text using the Undo command.

2.5.12 Deleting Text: Characters

1. Place the insertion point to the right of the text to be deleted
2. Windows: Press [**Backspace**] as many times as needed
Macintosh: Press [**delete**] as many times as needed the desired characters) is deleted.

2.5.13 Deleting Text: Type Over

1. Select the text to be replaced

NOTE: For more information, refer to Selecting Text.

2. Begin typing. The selected text is deleted and replaced with what you type.

2.5.14 Deleting Text: A Line or Block of Text

1. Select the text to be deleted

NOTE: For more information, refer to Selecting Text.

2. Press [Backspace] or [Delete]

The selected text is deleted.

2.6 Formatting Toolbar

Microsoft Word allows all toolbars to be customized. So, you may not find all options listed here. There are several buttons that may or may not appear immediately in your version of Microsoft Word. Use the following graphic as a guide to the Formatting Toolbar.

2.6.1 Style

Styles in Microsoft Word are used to quickly format portions of text. For example, you could use the "Normal" or "Default Paragraph Font" for the body text in a document. There are also three preset styles made for headings.

2.6.2 Font

Font is a simple but important factor in Microsoft Word documents. The choice of font (the style of the text itself) can influence the way others view Documents either on the screen or

in print. For example, Arial font looks better on screen, while Times New Roman is clearer in print. To apply a font to text, select desired text with your cursor, and chose a font from the font drop down menu.

2.6.3 Font Size

You may encounter times in which you need to display some text larger or smaller than other text. Selecting desired text with the cursor and choosing a font size from the drop down menu changes the size of text.

4. Bold: Places the text in bold.
5. Italic: Places the text in *italics*
6. Underline: Underlines the text.
7. Align Left: Aligns the selection to the left of the screen/paper.
8. Center: Aligns the selection to the center of the screen/paper.
9. Ling Right: Aligns the selection to the righ t of the screen/paper.
10. Justify: Aligns the selection to both the left and right of the screen/paper
11. Line Spacing: Adjust the line spacing (single-spaced, double-spaced, etc.)
12. Numbering: Create a numbered list.
13. Bullets: Create an unordered, bulleted list.
14. Decrease Indent: Decreases the indentation of the current selection (to the left).
15. Increase Indent: Increases the indentation of the current selection (to the right).
16. Outside Border: Places a border around the current selection; click the drop-down for a wide selection of bordering options.
17. Highlight: Highlight the current selection; default color is yellow.
18. Font Color: color the font

2.6.4 Saving Documents

When you are working with any sort of media in any software, you should be sure to save your work often. In Microsoft Word, there are numerous options for saving documents in a variety of file types. Figure 1.3 depicts one of the methods of saving a document. . To save a

new, unsaved document, you can click on the Save icon, shaped like a disk located on the standard toolbar. Or, you can go to the menu bar and select File » Save ... (shortcut: Ctrl+S).

A dialogue box will appear, offering you a number of options. To save the document in the desired location on your computer, locate and select the folder on, your computer. Give your document a name in the file name text box. While you can give your id cument long names, make sure you save it with a name you can remember.

Please note that it's good practice not to use spaces or special characters in file names. For example, a long file name may look like this: sample_paper1.doc to save a completely new document using previously existing (and opened) text, you use the Save As option. Open the document that you wish to save as an entirely new file, go to the menu bar, and click on File » Save as. In the file name text box, give your document a new name. Using this option allows you to save multiple versions (with different file names) of a document based on one original file

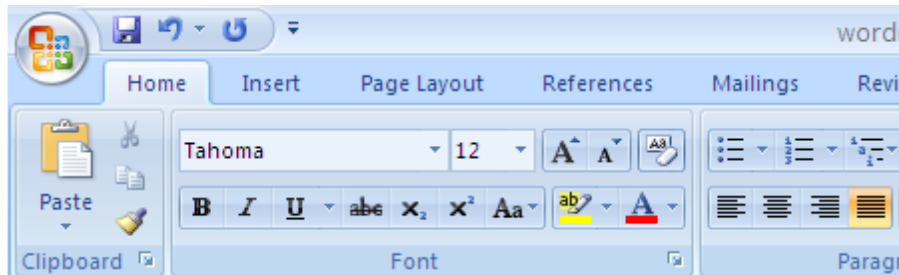
2.6.5 Inserting an Image

In Microsoft Word, it's possible to add clipart or other images to a document.

Click the cursor in your document where you wish to place an image. Then go to the menu bar and select "Insert" » "Picture." From there, you will find a number of options to choose from. "Clipart" searches through your computer's Clipart library. "From File" will allow you to insert an image saved elsewhere on your computer. Other options include "AutoShapes" and "WordArt."

2.6.6 Create Bulleted and Numbered Lists

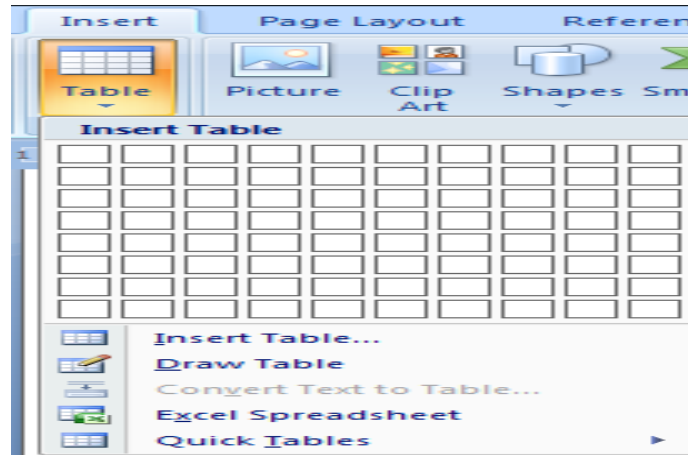
You can improve display of list by adding bullets or numbers in the starting of each item of the list. Bulleted list have bullets or icons in the beginning of the every entry of the list, similarly numbered list have number or alphabets in ascending order at the starting of each entry of the list. The outline list combines bullets and numbers formats to display list in the document. The outline list is also known as multilevel list. Option for creation and editing of various types of list are available under the Paragraph sub-group of the Home tab of the ribbon.



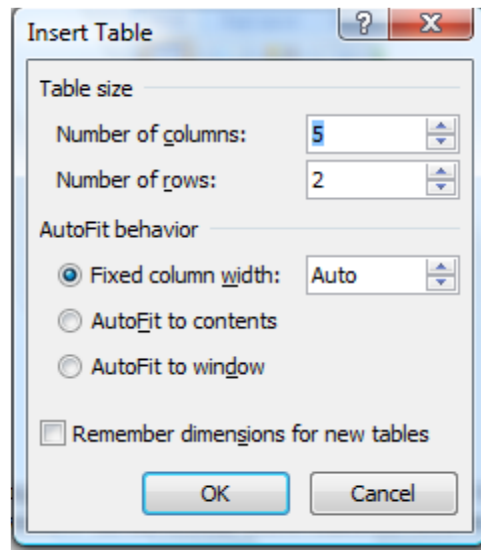
You can add bullet or number at the starting of each entry while typing a new list or these can be added to the existing lists. Just click on Bulleted list, Numbered list or Multilevel list icons before typing the list, now bullet or number is automatically entered as you press enter to add new item to the list. Once you finish the typing of the list, click again on Numbered List or Bulleted List icons to finish the numbering/bulleted of the list. Now typed contents will be added as a paragraph. You can add numbers or bullets to existing list by selecting the entire list to be numbered or bulleted and then click on one of the options i.e. Numbered list, Bulleted list or Multi level list. We can edit a list at any time. First select the portion of the list to be edited and then type the revised entries. When you click over the icon of Numbered, Bulleted or Multilayered list options, the list is displayed with default bullets or number style. If default style does not meet your requirement, then click on arrow key (□) next to icons of Numbered list, bulleted list or multi-levelled list to select one of the styles to make the list more attractive.

2.6.7 Create a Table

First insert amply blank space at appropriate in the document to accommodate a table. Take cursor to the position where you want to insert a table in your document. Click on the Insert tab from the ribbon then click of Table icon () under Tables sub group. You get a window like this on your desktop.



You can draw a table by dragging mouse pointer on the boxes listed in the form of grid under the Insert Table label. Drag mouse on desired numbers of rows and columns. A table with number of rows and columns dragged over by you is inserted in the document as you release the mouse button. You can also click on Insert Table icon. You will get the following dialogue window to insert a table in the document. Enter number of rows and columns needed in table in appropriate textboxes. Click on OK button to insert table in the document.




Enter Data in a Table:

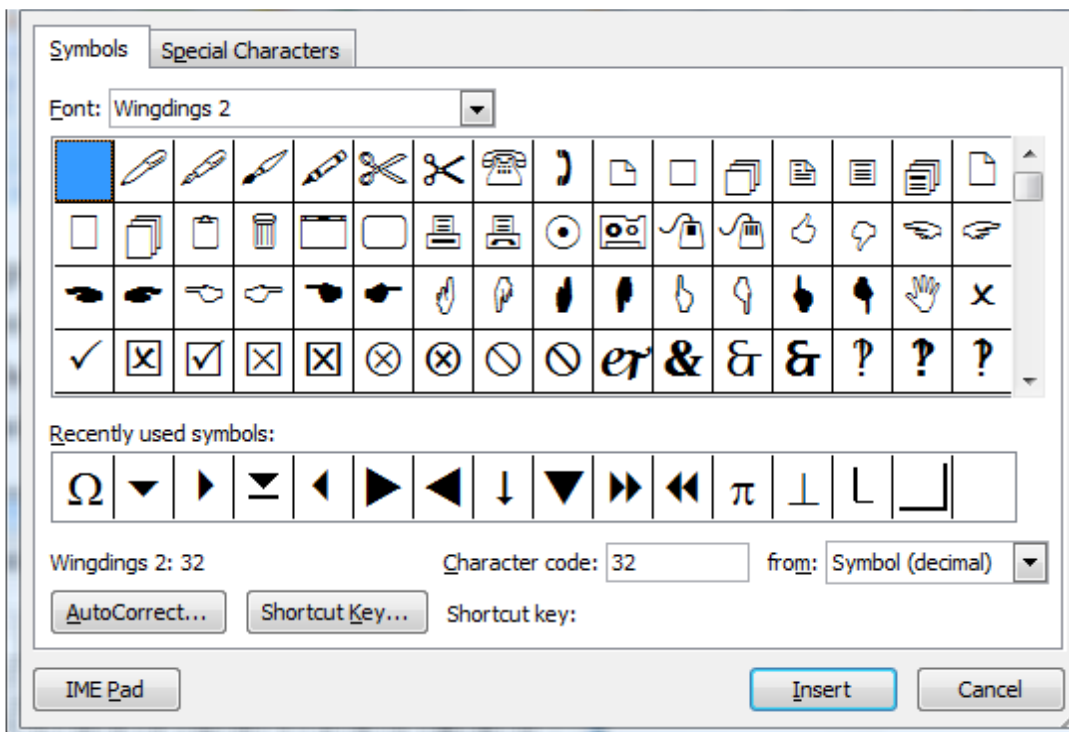
Place the cursor in the cell, where you wish to enter the information, cursor blinks in the left corner of that cell. Typed text is inserted at the position of the cursor.

Once you reach right end of the cell, text automatically moves to next line in the same cell. When you are typing in a table or have focus on table, notice two new tabs Table Design

and Table Layout appeared on the ribbon. Commands and functions of this table modify the structure and format of the table. Commands and functions of the Table Design tab are grouped into three major subgroups i.e. Table Style Options, Table Styles and Draw Borders.

2.6.8 Type Special Characters

Alphabets of foreign languages, mathematical symbols, other popular signs and symbols cannot be entered in document with the help of keyboard. You can insert these alphabets, signs and symbols in document by clicking on  Symbols icon of the Symbols sub-group of the Insert tab of ribbon. You will get a list of special characters/symbols used recently by you under the Symbol label, click on symbol to be inserted at the position of cursor. If symbol desired by you is not listed in recently used symbols then click on More Symbols ... option, special symbols are listed in a separate window as shown below.



Click on the symbol to be inserted, and then click on Insert button of the dialogue box. The selected symbol is inserted at the position of the cursor. If symbol desired by you is not listed then look for the symbol in different fonts by selecting new font style from the combo box of the Font: label

2.6.9 Create a Template

Template is special document that serve as starting point to draft a new document. A template under the MS Word 2007 is identified by its extensions i.e. dotx or dotm (a .dotm file type allows you to enable macros in the file). Templates are pre-formatted documents. You can design a template for business correspondence with name and logo of the company in the header of document, predefined page layout and dimension as per office stationery. Sections, fonts, margins, and styles of documents are directed by business ethics and practices. You need not define structure and format of business letters, every time you draft a business correspondence.

You can find pre-defined Word's templates by clicking on the Office Button, click on new option, various template categories are listed in the window on the left pane. Select one of the template categories, templates from that group are displaced in the middle pane and right side pane display preview of the template highlighted in the middle pane. Select the appropriate template and click on Create command button. New document adopts the structure and format of the template opted by you.

When predefined templates failed to meet your requirement then define your own template. Create a new document set its header, footer, margins, paragraph, line spacing, and company name and logo in header/footer, watermark logo and other styles options. Once you are satisfied with the style of the document then save it. Click on Office Button, take focus to Save as option, click on Word Template, define the name of the template and click on save option to store as document template.

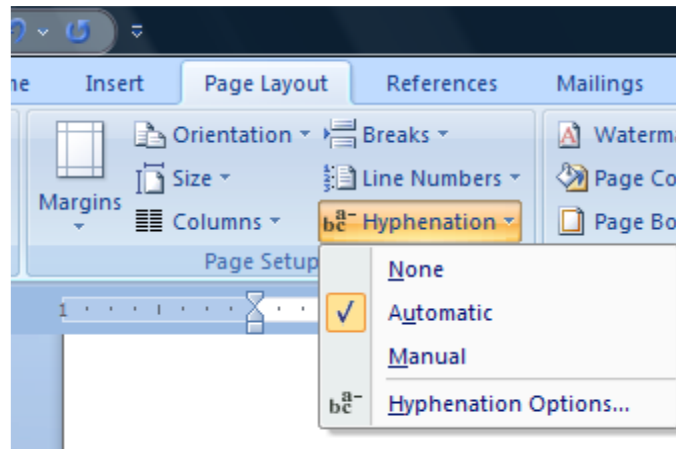
Spelling and Grammar Checker Spell and grammar checker tool is a powerful feature of the MS-Word. This tool checks spelling and grammar of the entire documents including hidden texts, text in boxes, header and footer. Click on Review tab and then select Spelling & Grammar tool () from the Proofing subgroup to initiate checking of the spelling and grammar of the documents. The checking starts from the position of cursor, goes to the end of the document, and starts again from the beginning of the document till the position of the cursor in the document, where you have initiated this tool. Later it checks contents of the header, footer, and textboxes and hidden texts. The following dialogue window appears over the screen while Spelling and Grammar tool is functioning and it notices any error in the document.

You can decide whether you want to check both i.e. spelling and grammar of the document or just the spelling in the document. If you are not interested in checking the grammar of the document then clear the check box of Check grammar option. If this box is checked then the Checker tool will check both grammar and spelling of the document. Whenever, the Spell Checker tool encounters any spelling and grammar errors in the document, it highlights the text with mistake in the document in the upper window of the spell-checker tool labeled as Not in Dictionary and makes suggestions from the dictionary, to rectify error, in the lower window labeled as Suggestions. Select the appropriate correct word from the various suggestions, listed in lower window, click on Change button to replace incorrect text with word highlighted by you. The Spell checker also offers alternatives to Change button such as Ignore Once, Ignore All, Add to Dictionary and Change All.

These options are self-explained. The Change All option rectifies all the repetition of the highlighted wrong word with words suggested by you, in a single stroke throughout the entire document. The Ignore once option retains the wrong word highlighted by the Spell Checker only for this instance. The Ignore all option retains the word highlighted by spell checker as wrong, throughout the document, later on this word is not highlighted as wrong in the document. You can exclude specific words from checking by specifying these words as not to be checked by the spell checkers. The spell checker treats these words correct in future. This facility is useful when your document includes foreign words or proper noun reflecting names of places, animal or person in the documents. These words are not listed in dictionary maintained by MS-Word. The Add to Dictionary option adds the highlighted word to dictionary, treats it correct in future, and never highlights it again wrong. You should be careful in adding in a new word to dictionary, if you add word with wrong spelling to dictionary than spell-checker would treat it as correct and do not flag it wrong in future. The spell check has limited capacity; it could not understand the context in which specific words have been used in sentences. During the grammar check, the tool identifies sentences having grammatical or stylistic errors and it makes suggestions to rectify error and improves the meaning of sentences. If you could not understand, how to rectify grammar errors highlighted by the Spell Check and need further explanation and assistance to understand the error then click on Explain option. The Spell Checker offers detailed assistance on highlighted grammar error in a separate window.

2.6.10 Hyphenation

The Hyphenation feature allows you to break long words into lines. This facility is frequently used by publishers to make the spacing between words uniform and keep the ends of lines on the margins of the page. Hyphenated break of word is reflected through hyphenation symbols, (-) at the position of break. The hyphenation mode can be switched on by clicking on the Hyphenation option under the Page Setup subgroup of the Page Layout tab of the ribbon. MS Word allows Hyphenation of words either manually or automatically. When you click over the Hyphenation icon, you have to decide whether hyphenation will be added manually or automatically. A tick mark appears against the hyphenation label on the ribbon, when it is on. Now long words are broken between lines.



2.6.11 Headers and Footers

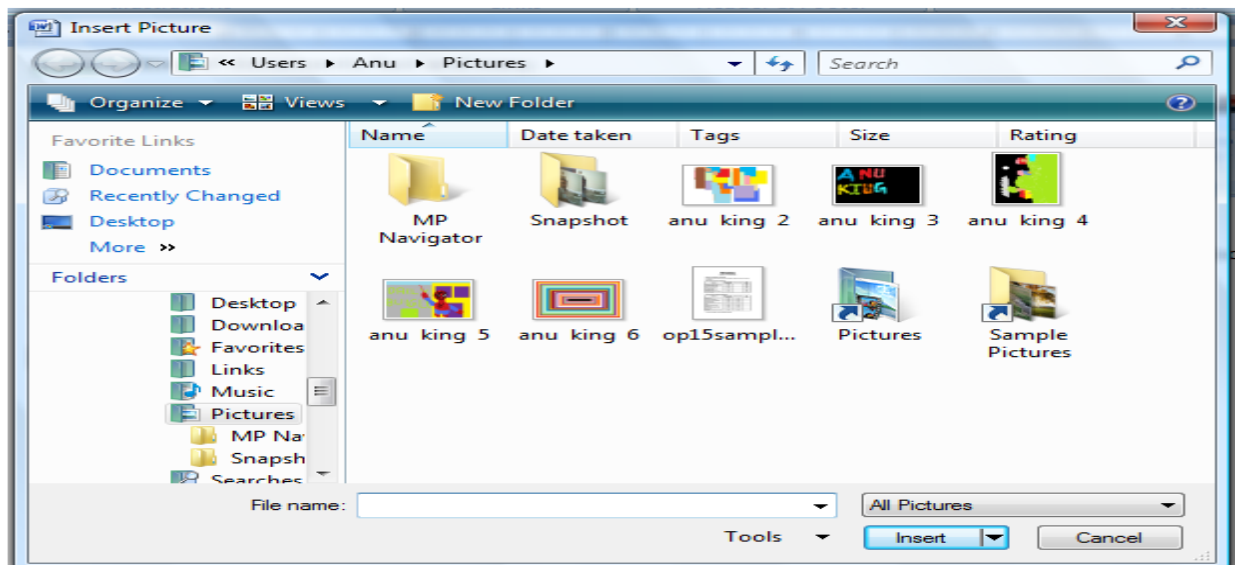
The top and bottom parts in a document beyond its contents are called header and footer of a document respectively. The information placed in header and footer is repeated on each page of the document, in the printed copy of the document. Contents of header and footer are not reflected while you are drafting the document. The header is primarily used to display name and logo of organization. The header may also include other information such as location of folder, where document is saved, page number, date/time etc. The contents of header and footer are independent of each other and designed independently. You can either have the same information in header/footer areas of each page of the document or have separate header and footer for each section of the document, or have separate header and footer for odd and even

pages. Certain components of header and footer change on every consecutive page for example page number changes in incremental order on each page.

You can add contents to header or footer in a document by selecting corresponding icons from Header & Footer sub group from the Insert tab of ribbon. When you click on Header or Footer icon, it shows predefined text and formats, which can be used as header or footer. You can select one of predefined header/footer or you can define your own header formats by clicking on Edit header icon. Similarly, you can define footer for your document. You can either adopt predefined footer or design your footer by clicking on Edit Footer icon. Customize the display of ribbon and behavior of MS word. You can customize the features and behavior of the MS Word software through the Word Option command under the Office Button menu. Click on the Office Button, and then click on the Word Option in the last line next to Exit Word option. Features of MS-Word, which can be customized by users, are grouped under the following headings on the left vertical pane of window.

2.6.12 Insert Graphs and Pictures in Documents

You can insert graphs, pictures, drawings and watermarks into your documents with the help of Insert tab of the ribbon. You can insert picture by taking cursor to the position where picture is to be inserted and then clicking on Picture icon of the Illustrations sub-group, you will get the following dialogue window to insert picture. Browse folder and select picture to be inserted by clicking over it and then click on Insert command button.

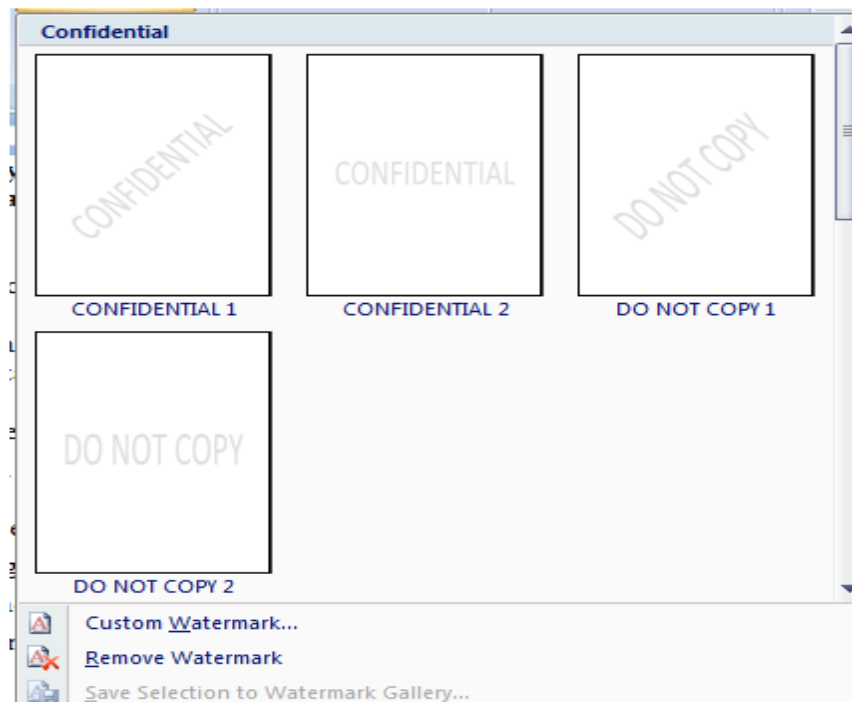


The Clip Art () option under Illustrations sub-group of the Insert tab is used to insert images, graphics and drawing. The Clip Art offers a gallery of images and drawings; it is an integral component of the MS-Office. The gallery archive frequently used drawings and icons. Some of the components of the gallery are installed as an integral component of the MS Office, for other component of gallery; you have to get it over the internet. Click on ClipArt Icon on the Illustrations sub-group under the Insert tab, you will get the following window on the right side of the document window. This window is titled as Clip Art. The collection of Images under the ClipArt is classified into various groups for convenient and fast surfing of clipart gallery. You can view images from a specific group by selecting that group from the listing of Search in combo box. Insert a specific image by clicking over it. The image is inserted at the position of the cursor.

2.6.13 Insert Watermark in the Document

A watermark is a translucent image that appears behind the primary text in a document. It is mainly used to insert logo of the company. To insert a watermark

Click the Page Layout Tab in the Ribbon, Click the Watermark Button in the Page Background sub group and get the following Window.

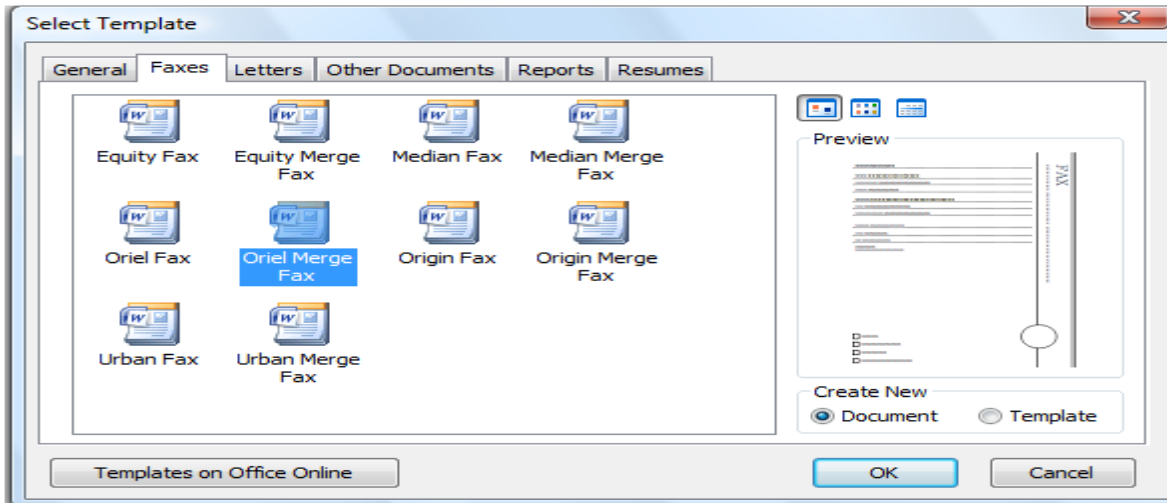


Select one of the five predefined layouts to set watermark in the document. If predefined watermarks do not meet your requirements than click on Custom Watermark and create your own watermark Picture can be inserted as watermark by selecting on Picture watermark radio button and click on select picture icon. Select the picture for the watermark. The Scale: option zoom the picture to specific % of original size of picture. Select the Washout check box to lighten the picture so that it does not interfere with text, on the foreground. You can remove a watermark by clicking over Remove water mark option. There are options to set colors (Page Color) and border of the page (Page Borders) bellow the Watermark option in the Page background sub group. Colored pages are useful, when you upload pages on web or take printout, of voluminous documents over colored printer. Printouts with Colored background can be easily identified in a voluminous document.

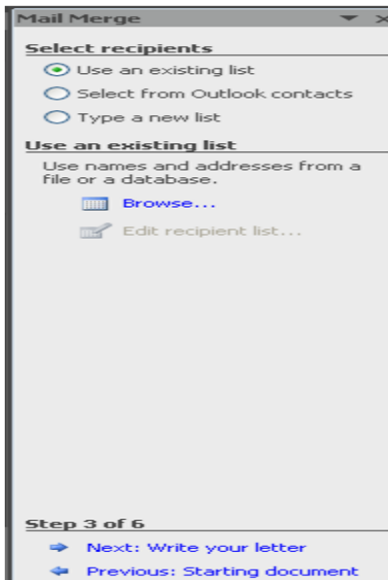
2.7 Mail Merge of Microsoft Word

The Mail Merge is a quick and easy way to send the same letter to a number of people, without typing it again and again for each person. You can draft personalized letter for individuals by merging content of letter with address database. The address database store information about address of individual receivers. Mail Merge can also be used to prepare address envelopes, mailing labels, phone directory apart from individual letter. The Mail Merge procedure is explained with the help Mail Merge Wizard. This procedure is very helpful for students who are new to mail merge. The Step-by-Step Mail Merge Wizard offers immediate assistance at each step and makes suggestions for the next step. Once you are well conversant with mail-merge wizard, you can easily create mail merge without the wizard. Click on Mailing tab, of the ribbon click on arrow key () next to Start Mail Merge icon, click on Step by Step Mail Merge Wizard option to initiate the wizard. This procedure takes six steps to complete the mail merge. In the first step, computer will ask what type of documents you would like to prepare selected document may be letters, e-mail messages, envelopes, label or Directory. Once you selected type of document, click on Next: Starting document option at the lower corner of dialogue box to move to next step. The next step will ask how do you want to design your letter One of the options is to use the current document, you have been working. If you have already designed a template to customized mail merge then select Start from a template option and select

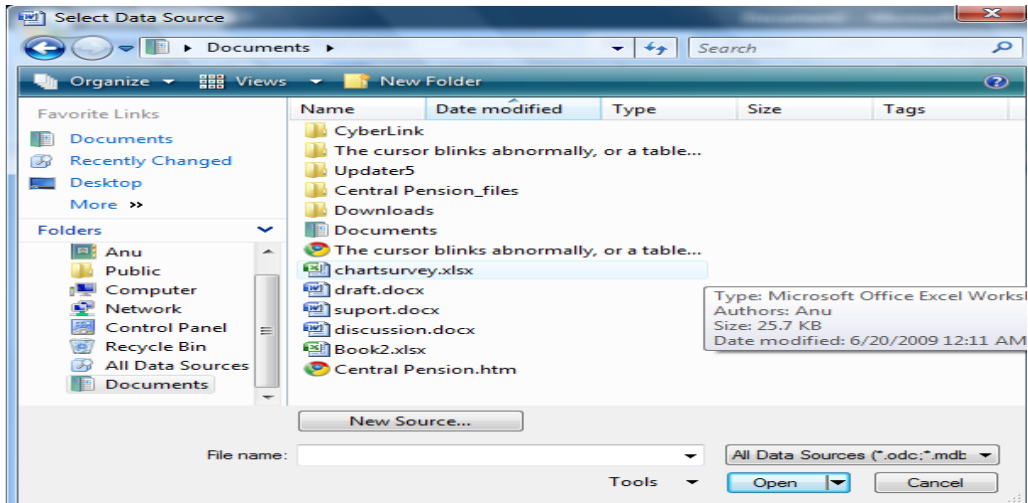
appropriate template by browsing through templates supplied with the MS office or designed by you. For current exercise, select Start from a template option. This brings up a link Select template in the Window.



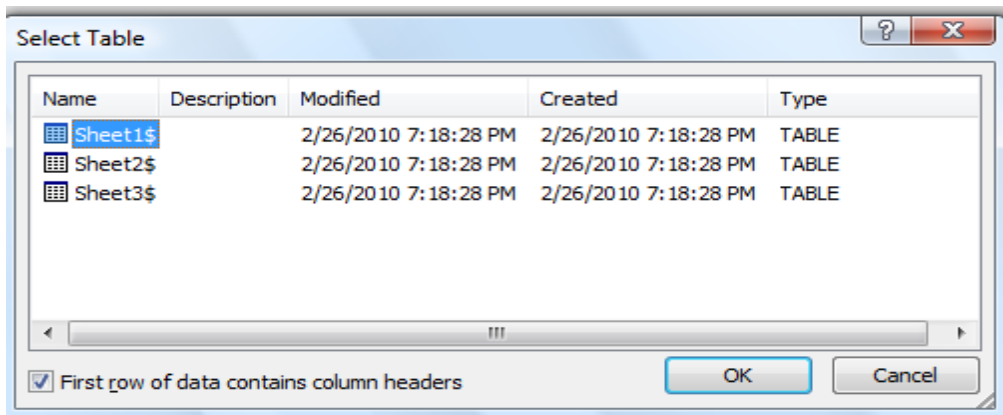
Once you select a template of your interest, click on Next: Select recipient's link, the next step is to decide recipient of the mail merge letters.



You can opt to use existing database or create a new database of recipients. You can also use contact list maintain by the MS Outlook. If source data is already available on computer then select Use an existing list option, and click on Browse link, a window will appear on the screen to choose the source of data for the mail merge.



Decide the type of source file by clicking over All Data Source combo box, a list of different formats supported by mail merge will appear, select one of the formats. The window displays all the sources from the selected folder in the format specified by you. Here, we have used an excel file to do mail merge, you can also use database, text, word documents etc. as source for mail merge. We click on the source excel file and then click on Open Command button. The Excel file selected by us has three worksheets named as Sheet1\$, Sheet 2\$ and Sheet3\$, these are listed in next window titled as select table and prompt us to select one of the worksheet as source of information about recipient.



Select the sheet and click on OK command button. You will get next window displaying all records from the selected worksheet. A single row of the excel sheet represents record of single recipient. All records are displayed with tick mark in the check box for each record. Tick mark in check box indicates that specific records are included in the mail merge.

2.8 Summery

This unit describes the essential features of Microsoft Word. After studying this unit, we will be able to create documents, open the documents that were earlier created save the documents as well as perform simple operations. The unit also describes operations such as including text, graphics, tables, clip artistic. Also, the unit introduces performing some advanced operations on the documents. It also describes the process of creation of different kinds of documents such as simple notes and memos as well as multi-column reports with tables, graphics, etc. This unit also describes Mail Merge facility in Microsoft Word. The unit also describes several other features of Microsoft Word. the purpose of a dashboard is to effectively display the necessary and sufficient data with added visual impact as required by the potential audience. The layout of the dashboard and its components vary across the different viewers based on their preferences Microsoft Word is a great tool for creating interactive and dynamic dashboard.

2.9 Self-Assessment Exercise

1. Create a document as Read-Only and then change the AutoSave duration of MS Word to 3minutes
2. Create a new document and then password-protect it by setting a password.
3. The basic formatting buttons provided by MS Words are changing document styles, font styles, type styles, paragraph styles, alignments, listing formats, indenting levels and borders.
4. The various components of a title bar are control menu, document title, minimize button, maximize button and close button.

2.10 References

- www.support.office.com
- www.tutorialspoint.com
- www.chandoo.org
- www.ignou.ac.in

UNIT:3**COMPUTATIONS WITH MS EXCEL - PART I**

Structure

- 3.1 Introduction
- 3.2 Objectives
- 3.3 How to Open MS Excel
 - 3.3.1 Starting to MS Excel
- 3.4 Components of the MS Excel Window
 - 3.4.1 Office Button
 - 3.4.2 Title Bar
 - 3.4.3 Ribbon Bar
 - 3.4.4 Status Bar
 - 3.4.5 Sheet Tabs
 - 3.4.6 Formula Bar
 - 3.4.7 Worksheet
 - 3.4.8 Create a New Workbook
 - 3.4.9 Entering Data
 - 3.4.10 Formatting and Styling Data
 - 3.4.11 Setting and Clearing the Cells' Styles
- 3.5 Type of Data Supported by MS-Excel
 - 3.5.1 Numeric
 - 3.5.2 Text
 - 3.5.3 Date and Time
 - 3.5.4 Data in Series
 - 3.5.5 Boolean or Logical Data
 - 3.5.6 Deletion of data
 - 3.5.7 Cell Formats
 - 3.5.8 General Format
 - 3.5.9 Number Format
 - 3.5.10 Currency Format

- 3.5.11 Account Format
- 3.5.12 Date Format
- 3.5.13 Percentage Format
- 3.5.14 Fraction Format
- 3.5.15 Scientific Format
- 3.5.16 Text Format
- 3.5.17 Special Format
- 3.5.18 Custom Format
- 3.5.19 Print Preview
- 3.6 The Specification and Limitation of Worksheet Designed with MS Excel
 - 3.6.1 Calculation Specifications and Limits
 - 3.6.2 Inserting Rows or Columns
 - 3.6.3 Auto fills
 - 3.6.4 Sorting data
- 3.7 Summery
- 3.8 Self-Assessment Questions
- 3.9 References

3.1 Introduction

In this unit will cover all topics from MS Excel basic to advance. There are specialized functions to perform financial calculations and statistical analysis. The terms worksheet, spreadsheet and workbook are used interchangeably by different books and authors for calculation worksheet prepared with MS-Excel. Here is some of main usage of the MS Excel.

1. Create charts for visual presentation of data,
2. Do statistical analysis and survey,
3. Organize large volume of data and records, and Automatic computation of complex calculations.

Electronic worksheet is like finance ledgers, maintained manually having entries in tabular form (rows and columns). The MS Excel workbook consists of grids of cells, arrange in tabular

form as crossing of rows and columns. Each cell is an independent unit for keeping records. Data store in cells may be numeric, string, date or time.

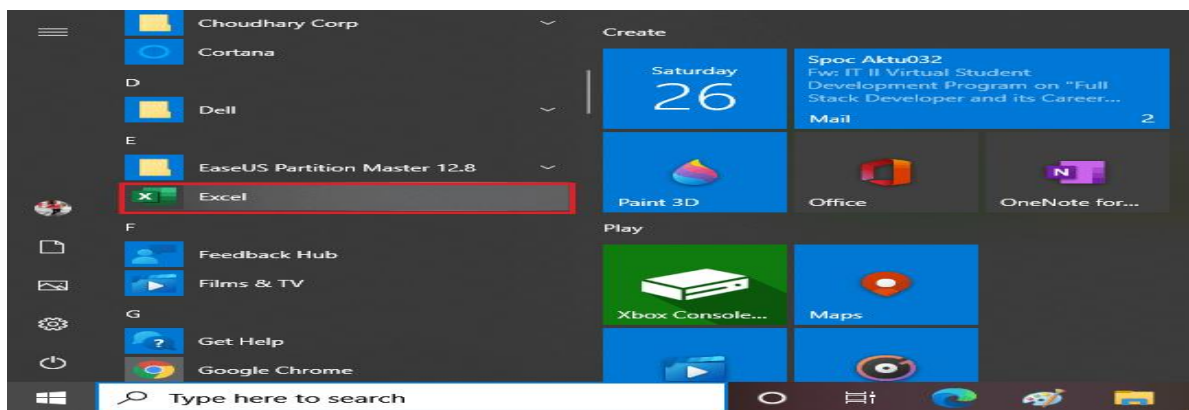
3.2 Objectives

MS Excel is an integral component of the MS Office. It is used for preparation of calculation worksheets, conduct data analysis and present data in the form of charts. After reading this unit, you will be able:

- to understand basic features of MS Excel
- to feed data or modify entries in worksheet
- to get print out of worksheet and set footer/header

3.3 How to Open MS Excel

In Windows 10 operating system, click on the Start button and search for the MS Excel application. If it is already installed in your system, it will appear here like this



Double-tap on this icon to open the Excel, When the Excel opens; an interface will appear like this. From here, we can create a new workbook, choose a template, and access your recently edited workbooks

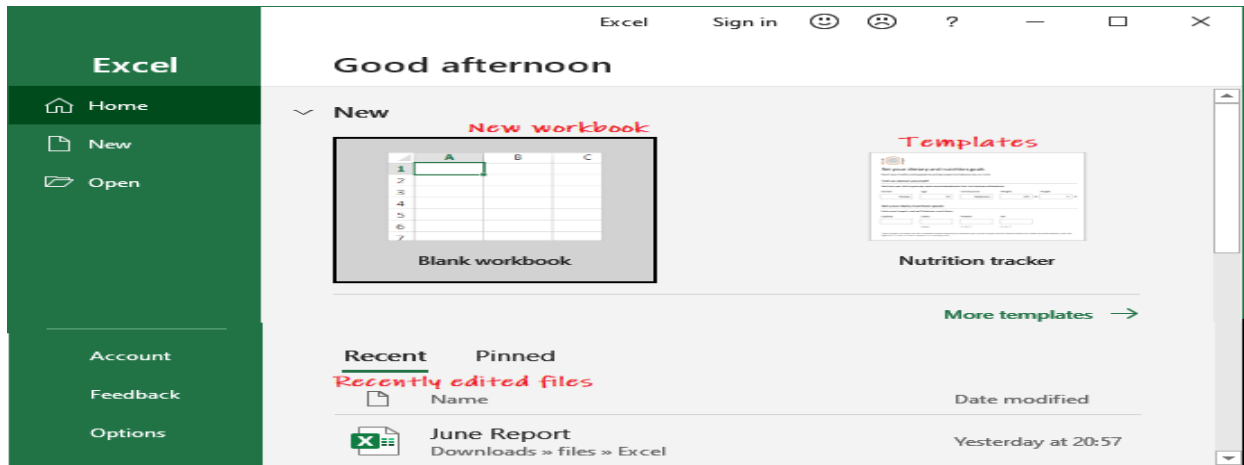


Figure A

3.3.1 Starting MS Excel

Click on the Start button, takes pointer over Programs on the start menu, wait for next cascade menu. Point to Microsoft Office entry of the Programs menu wait for next menu to appear, double click on Microsoft Office Excel to start it or click on Microsoft Office Excel and press enter key. We can also open MS Excel software by double clicking on files prepared with the MS Excel software. The files prepared by the MS Excel can be identified by their extension i.e. xlsx (workbook prepared with MS Excel 2007) and xls (Workbook prepared with the earlier version of MS Excel). The first window of MS –Excel looks like as shown below in the figure A.

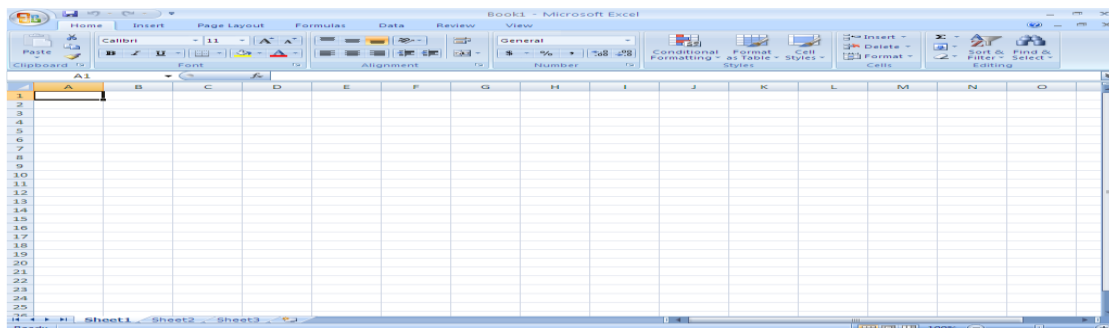


Figure A

Data in a worksheet is presented in the tabular form. The cross-section of rows and columns forms cells. Cell is an independent unit for inputting data. Cells can contain texts, number, a combination of text and number, date, time or mathematical formula. The headings of columns are designated by English alphabet on the top of the first row of the worksheet. Rows of

the worksheet are designated by numbers in the left side of the first column of the worksheet. The headings of rows and columns are displayed in sky blue colour. The first column in a worksheet is identified as A followed by columns designated as B, C, so on up-to Z, these are followed by columns designated as AA, AB and so on. Similarly, first row of the worksheet is identified as 1; it is followed by rows identified as 2, 3, 4 so on. Each cell is identified and referred by the unique address. Cells are identified uniquely by their column and row numbers. The cell address is defined as its column label (A, B, C etc.) followed by its row label (1, 2, 3 etc.). Cell under focus is known as active cell. An active cell is identified by its dark and bold border. For example, the active cell in the above figure is A1. When you open a worksheet, the focus is always on the A1 cell. When we type, while working with a worksheet, it is entered in the active cell. You can use arrow keys or mouse pointer to take focus to cells of your interest in the worksheet. Tab key can also be used to move focus horizontally one cell at a time in a row. Scroll bars are situated at the right and bottom corners of the worksheet windows; these allow you to see components of the sheet, which are not displayed in worksheet window. Scrollbars“ sliders can be dragged to take focus to contents beyond size of window. If we can select a single cell, a row, a column, and group of rows, group of columns or the entire worksheet. To select the entire row/rows or column/columns, click on the heading of rows or columns, to be selected. When pointer is over the heading of rows/columns, it changes to, now click on mouse button. The selected rows or columns are highlighted and surrounded by dark lines as shown below. You can select individual cell by clicking on the specific cell. To select a range of cells, click on the cell situated at one of the corner of region to be selected, now press the left mouse button and without releasing button, drag the mouse pointer in the direction of cells to be selected, until the last cell of the range, to be selected, is covered. You can select more than one rows or columns by drag action. The selected cells have dark borders and displayed in different background colour. We can select the entire worksheet by clicking on the icon at the crossing of rows and columns headings.

3.4 Components of the MS Excel Window

The Window of the Microsoft excel have the following components:

3.4.1 Office Button

The Office Button of the MS Excel has similar commands as being made available in the MS-Word. We have discussed in detail about the commands of the Office Button menu in the units1. The common features of different components of the office button are not discussed again to avoid the repetition.

3.4.2 Title Bar

The title bar is situated at the top of the window; it shows the name of worksheet on the left side of the Microsoft Excel label. Both entries are written exactly at the center of title bar. The Office button is situated at the left most corner of the title bar. The title bar has three icons at the rightmost corner i.e. minimize, maximize and close buttons (). There is Quick Access Toolbar next of the Office Button on the title bar. This toolbar consist of commands used frequently. You can add or remove commands from the Quick Access Toolbar by click on arrow key (▾) next to it. The commands visible in this toolbar are listed with tick mark. You can add/delete commands by clicking of specific command from the list.

3.4.3 Ribbon Bar

The ribbon of MS Excel has seven tabs i.e. Home, Insert, Page Layout, Formulas, Data, Review and View. Commands under Home, Insert and Review tabs of the MS Excel are almost similar to corresponding tabs of the MS Word software. The entire ribbon is displayed on the screen when it is displayed with resolution of 1024×786 pixels and MS Excel window is maximized on the screen. When you reduce the size of the Excel window, the size of Ribbon is reduced; some of the tabs may not be visible. Commands of the active tab of the ribbon in this situation shrink horizontally and display as a single icon. The most frequently used commands or features are still visible in the shrunk ribbon. You have noticed that a large part of MS-Excel window is occupied by ribbon. You can minimize the size of ribbon by clicking on the Down Arrow icon to the right of the Quick Access toolbar and click on Minimize the Ribbon option,

use the shortcut Ctrl + F1, or right click in empty space of the ribbon and choosing Minimize the Ribbon option from the pop-up menu. The Quick Access Toolbar, standard commands for editing and formatting data and their shortcuts in all the subcomponents of the MS office are alike so discussion about these commands is avoided in other units of the course to eliminate repetition.

3.4.4 Status Bar

The status bar is situated at the bottom of the excel window. This bar displays information about the current worksheet. The leftmost part of the status bar display the processing status of the worksheet, whether worksheet is processing data, ready to get instruction from you or editing data as typed by you. The right side of the status bar has icons to display the active worksheets in different visual layouts. The MS Excel supports different display layouts to improve the visual of the worksheet and make data entry and editing convenient. There are three icons to set the layout of the worksheet i.e. Normal, Page layout and Page Break Preview (). Zoom slider is used to magnify or shrink the view of the worksheet over the screen; the actual size of the worksheet remains intact.

3.4.5 Sheet Tabs

When you open a new workbook, by default, there are three worksheets in a workbook titled as sheet1, sheet2 and sheet3. You can add more worksheets or delete the existing one as per your needs. Each worksheet is shown, as a tab is the status bar. You can jump across worksheets by clicking the tabs of corresponding work sheets. If you have added large number of worksheets in a workbook tab for only few worksheets are visible at a time on the windows. If tab of the worksheet, which you want to open is not visible then use the work sheet-tab scroll button situated next to tabs of worksheet, to see tabs of other worksheets. Keep on scrolling worksheet scroll button until tab of the worksheet of your interest is not visible, click over it, to open it.

3.4.6 Formula Bar

There is a formula bar situated under the ribbon. The text box on the left side of formula bar displays the address of cell, where formula will be applied. This textbox is recognized as Name box. The text box on the right side of the Name box is used to enter and edit formulas. A formula can have up to 8192 characters. If entire formula is not visible in formula bar then drag lower boarder of the formula bar to increase its size.

3.4.7 Worksheet

A worksheet is made of rows and columns that intersect each other to form cells where data is entered. It is capable of performing multiple tasks like calculations, data analysis, and integrating data.

3.4.8 Create a New Workbook

A workbook is defined as a collection of related worksheets saved together under a single name. When you open a new workbook, by default it has three worksheets. We can add new work sheets in your workbook as per your requirements. The maximum numbers of worksheet in a workbook depends on memory capacity of your computer. To start a new workbook, click on the Office Button and select new option. Excel opens a blank sheet with mouse pointer situated in the first cell of the first row of the worksheet. When we open a new workbook, it is assigned a default name book1 and underlines three default worksheets, are named as sheet1, sheet 2 and sheet 3. Consecutive new workbooks are assigned name such as book2, book3, and so on i.e. the workbook is prefixed a number, which increase each time you create a new workbook. However, you can rename default titles of both worksheets and workbook later. The window of MS –Excel worksheets looks like as shown below in the figure B.

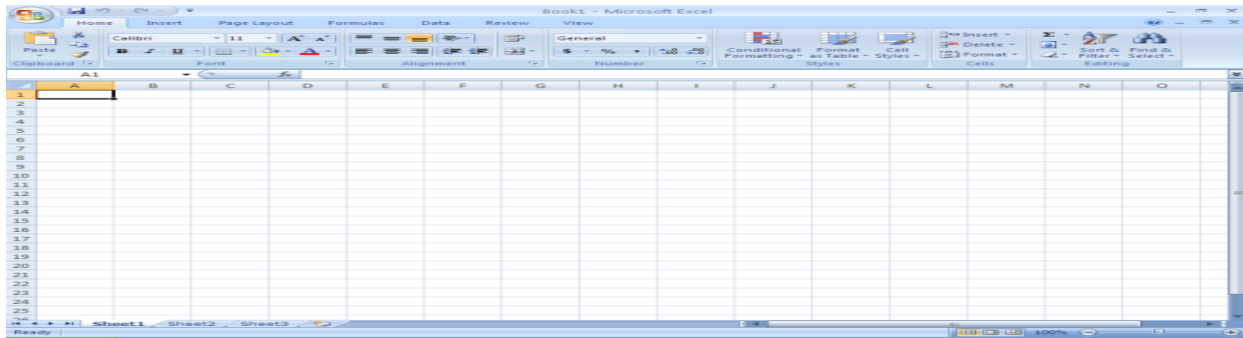


Figure B

3.4.9 Entering Data

First, take focus to cell, in which you want to enter data by clicking over that cell or by using arrow keys to move to specific cell. Up and down keys are used to move cross rows. Left and right keys are used to move across columns. We can use tab key to take focus to next column on the right side in the same row. If there are contents in the active cell then its contents are displayed in the formula bar. Whatever is typed now, it is displayed in the formula bar, and it overwrites the existing data. We can move data from the formula bar to active cell by pressing

Enter key or by taking focus to another cell in the worksheet. When you press the Enter key, the focus moves to cell in the next row downside but in the same column. If you do not want to move texts typed in formula bar to active cell then press the Escape key without shifting focus to another cell. We can edit existing entries in cells, by taking focus to specific cell. The existing entry of the active cell appears in the formula bar. Click at specific position in the contents of the formula bar and make changes in the text displayed in the formula bar, and press the Enter key after making changes. The revised text from the formula bar appears in the active cell of the worksheet as focus is moved to another cell active in the worksheet

	A	B	C	D
1	Items	Quantity		
2	Pen	15		
3	Pencil	24		
4	Copies	13		
5	Glue	12		
6	scale	5		

Figure C

We can insert multiple lines in a single cell. The positions, where you want to insert a new line while entering data in a cell press Alt +Enter shortcut. Texts typed afterward appear in a new line. If you need more line break then use this shortcut again, when a cell displays ##### through you have typed number or text in that cell, it indicates that display of this data requires a wider column to display its content. Take mouse pointer to the boundary of this cell, the pointer takes the shape of I, now drag the boundary of the cell to widen it until we see the information typed by you in the cell.

3.4.10 Formatting and Styling Data

First, select the text to be styled either by drag action or by using Shift (arrow key) shortcut, when focus is on the top most corner of the region to be selected or by using Shift+(arrow key) shortcut, when focus is on the bottom most corner of the region to be selected. Select appropriate font size, font style and other formatting features to improve the display of information in the highlighted area. You can also use different colours for fonts and color fill effect by using commands and options available under Font subgroup of the Home Tab. Colors of texts and background fill effect help in highlighting cells with important characters/features in a large size worksheet. You are familiar with different features and functions offered by the MS Word for editing and formatting texts.

3.4.11 Setting and Clearing the Cells' Styles

Several steps are needed to apply specific style and format to cells. We can reduce these steps involved in setting style of cell and can perform it in one-step with the help of Cell Styles function. You can also keep consistence in the cells style across the worksheet/workbook by using the Cell Styles option. MS Office Excel has several pre-defined cell styles that specify format and style of cells. These predefined cell styles can be applied to selected cells of the worksheet. Select the cells to be styled with specific features. Click on Cell styles under the Styles sub group of the home tab you will get the following dialogue box

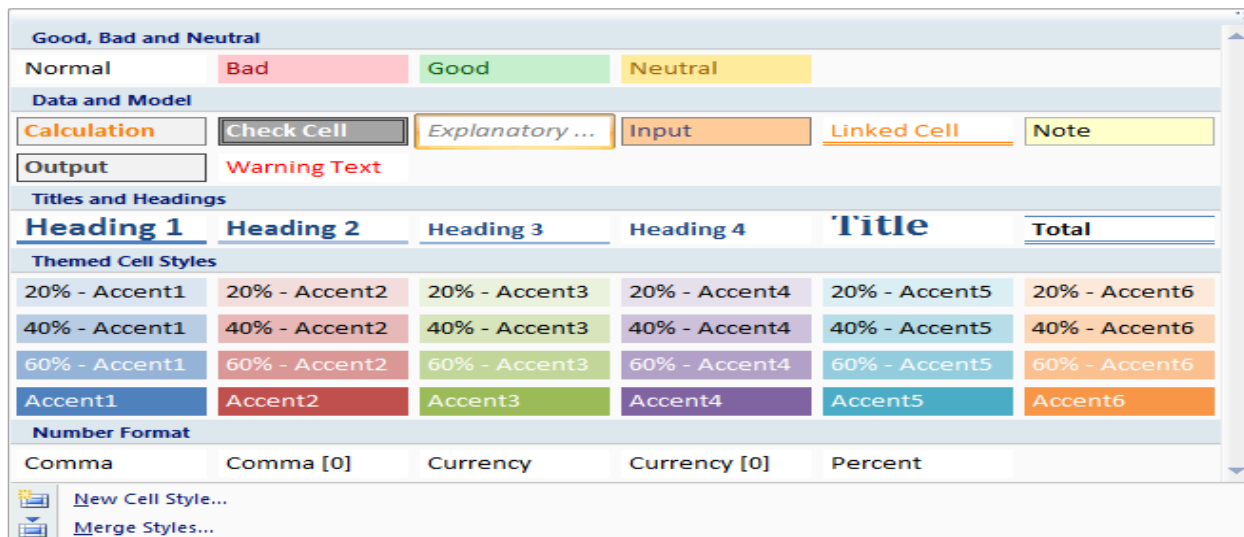


Figure D

Click on the Cell style that you want to apply to highlighted area of the worksheet. If listed cell styles fail to meet your requirement, then click on New Cell Style option, you will get following window to define a new style. Specify the name of new style, click on Format button, you will get Format Cells dialogue box. Use different tabs of the Format Cells dialogue window to design in new style, click on OK command button to bring focus to earlier window (Style), now again click on OK button to finish the designing of style and apply it to selected cells. Select the cells where you want to clear the cell styles, Click on Arrow key (□) of Cell Styles under the styles sub-group of the Home tab of ribbon and select Normal option from Good, Bad, Normal and Neutral options under the Cell styles option. The contents of the cells do not change with clearing of the cell styles.

3.5 Type of Data Supported by MS-Excel

A cell in a worksheet can have following type of data

3.5.1 Numeric

Numeric data may be whole numbers, such as 3, 5, or 7, may be decimal number such as 45.15 or it may be in scientific form such as 0.2567E+2. MS Excel displays number in scientific notation automatically, when you enter a number too long to be viewed in a cell. You can increase or decrease numbers of digits after decimal place by clicking on icons under Number subgroup of the Home tab of the ribbon. First select the cells to be modified (number of digits after decimal place are popularly known as significant digits) and then use icon. The numeric data is right aligned automatically in the cell, as we do in writing numeric figures manually.

3.5.2 Text

We can type string, number or combination of both as text data. Certain symbols such as @, +, and = are treated as an integral component of formula. When you type text starting with these symbols, Excel may treat that text as formula. For example when you type =MOP in a cell and press Enter key then Excel will display this text as #name?. If your text start with these reserved symbols then put apostrophe (,) before the text so that MS Excel treats the reserved symbol as a part of text not as the reserved symbol. If you have numeric data those are to be treated as text, such as phone number and pin code number then use apostrophe (,) before the number. You cannot perform any calculations on numbers, which are entered as text. Text data is automatically aligned to left in the cell as we do in writing text.

3.5.3 Date and Time

When you enter data reflecting date or time, Excel converts these entries into serial numbers and keeps these serial numbers as background information. The serial numbers are again converted into date and time formats, when you retrieve data from computer. Dates and time are displayed in the worksheet as per the format set by you, under the Regional setting of the Control panel. You can still modify the display the data in a specific format; when Excel failed to recognize a date or time-based data entered by you then it is treated as text.

3.5.4 Data in Series

We can fill a range of cells either with the same value or with a series of values either increasing or decreasing in systematic manner, with the help of AutoFill facility of the MS Excel.

3.5.5 Boolean or Logical Data

This data type is used for entering data having one of the two possible values. The data value may be either TRUE or FALSE, in terms of numbers, 0 represents false value and any other number represent true value. For example, gender of employers can be defined as Boolean data i.e. male and female.

3.5.6 Deletion of Data

To delete contents from specific cell, first take focus to the cell to be deleted than press Backspace or Del keys. You can delete a range of cells by selecting cells and press Del key. The selected cells can be deleted by clicking on the Delete icon under the Cells subgroup of the Home tab. If you have deleted certain data accidentally then you can revert-back the changes committed accidentally, either by clicking on Undo button in the Quick Access tool bar or use shortcuts Ctrl +Z. Changes are reverted-back one at a time in chronological order with latest changes are reverted first.

3.5.7 Cell Formats

The format feature of a cell controls the display of data in that cell. It only modifies the display of data on monitor or printer, underlying data of the worksheet remains intact. The format option improves the readability of the worksheet and highlights the important data in the worksheet. MS Excel supports the following formats to display data.

3.5.8 General Format

This is the default format of cells. Excel displays text and number under the General format. This format style has no specific style.

3.5.9 Number Format

Number can be displayed as integer, fixed decimal or punctuation formats.

You can set number of digits after decimal place. However, currency cannot be represented in number format.

3.5.10 Currency Format

We can add currency symbol, to improve display of money. MS Excel 2007 supports more than 250 different currency symbols and Indian Rupees (old one) is one of them.

3.5.11 Account Format

This format is used to create accounting worksheet. This format is popular among accounting professional. Numbers with accounting format aligns decimal points of all numbers in a column. Comma is used at predefined position in number to making it easy to read large numbers.

3.5.12 Date Format

This format improves the display of date and time. People in different regions of the globe have adopted different notions to display date and time. If you include * symbol before date, then format for display are in accordance with regional setting of the operating system.

3.5.13 Percentage Format

This format displays numbers as percentage. The decimal point of the percentage number moves two places to the right and percentage sign appears at the end of the number

3.5.14 Fraction Format

This format displays digits after decimal place as fraction rather than digits.

3.5.15 Scientific Format

This format displays numbers as exponential notation. This format is mainly used in engineering and scientific research. When number is too large to fit in the wide of cell, then it is automatically displayed in scientific format.

3.5.16 Text Format

The cell with texts format will treat each data as texts even when you have typed only number, date or time in the cell. Contents in text cell are by default left aligned. The contents of the cell are displayed exactly as these are typed. You cannot perform calculation on data entered in worksheet with text format.

3.5.17 Special Format

Certain data consist of combination of text and number in systemic manner. For example, number plate of a car registered in Delhi may be DL 6C H 4159. Here DL, C, and H alphabets are placed at predefined position and format. This format is used mainly to display Pin code, phone number, Driving license number and vehicle registration number etc. You can quickly type numbers without having to enter the punctuation characters in the data. Punctuation marks are automatically inserted into data after a fix length. This format is also very helpful when you import data from other applications compactable with MS Excel.

3.5.18 Custom Format

When predefined formats of the MS Excel fail to meet your requirements then you can set your own format with custom format option. We can set the format of cells by first selecting affected cell, right click to get pop-up menu and click on Format Cells you will get the following dialogue window. Different type of formats supported by the MS Excel is listed under the Category list. When you move across this list, the Sample frame displays format under the

highlighted category and a short description about that format is listed below the Sample frame. Select the appropriate format to get desired effect. The last option in the list is Custom; use this option to set your own format, when predefined formats fail to meet your requirements.

3.5.19 Print Preview

The Print Preview display mode allows you to view the layout of the worksheet, when it is printed. The display of worksheet under the print preview mode is placed under the What You See is what you Get (WYSIWYG) category. Thus, the print preview of worksheet is very close to its display, when printed. The option for print preview is available in Print submenu of the Office button menu. Point the mouse pointer to Print submenu and wait for next cascade submenu to pop-up. Select Print Preview command, to preview worksheet. The preview is displayed in a separate window, commands displaced in the ribbon, while worksheet is in the print preview mode is shown below. This tab is recognized as the Print Preview tab.

The impact of the various options of the Page setup dialogue box as defined in the last section is visible only under the print preview mode or printed worksheet. Click on the Next Page or Previous Pages buttons at display other than current pages of the worksheet and click the Zoom icon to increase or decrease the view size of worksheet. The zooming effect on the worksheet takes place only on its display on screen; there is no change in the size of worksheet, when it is printed on paper. If you want to change dimension of paper and other printing features then click on Page setup option. Options of the Page setup command are already discussed in the last section. Click on Close Print Preview to return the worksheet back to normal position.

3.6 The Specification and Limitation of Worksheet Designed with MS Excel

Maximum number of rows and columns- in	1,048,576 rows and 16,384 columns in a Worksheet
Column width	255 characters
Row height	409 points (72 points = 1 Inch)
Number of Pages	1,026 horizontally and vertically
Total number of characters that a cell can	32,767 characters

contain	
Characters in a header or footer	255
Number of sheets in a workbook	Limited by available memory in computer is(default 3 sheets)

3.6.1 Calculation Specifications and Limits

Number precision	15 digits
Largest number can to be typed into a cell	9.999999999999999E+307
Largest positive number allowed	1.79769313486231E+308
Smallest negative number allowed	2.2251E-308
Smallest positive number allowed	2.229E-308
Largest negative number allowed	-1E-307
Length of formula contents	8,192 characters
Earliest date allowed for calculation	January 1, 1900 (January 1, 1904, if 1904 date system is used)
Latest date allowed for calculation	December 31, 9999

3.6.2 Inserting Rows or Columns

Take pointer to cell, row or column where you want to insert cells, rows or columns. Select specific number of rows, columns and cell to be inserted into worksheet by dragging mouse pointer over existing entries and then right click to get pop-up menu. Select Insert option from the pop-up menu. If you have selected specific number of rows before clicking on Insert option then specific number of new rows get inserted above the selected rows and existing rows below the selected rows go further down to make space for new rows. Similarly, existing columns right to selected columns move further right to make space for newly inserted columns. When you have selected specific cells before selecting Insert option then you will get the following dialogue window to decide displacement of existing columns, rows or cell after inserting specific number of cells. The window reflects that you can insert entire blank row, columns, individual cell either on left or side of the faced cells. Select on the option of the above

window and click on OK button to interest rows/columns/cells in the worksheet. You can also use Insert option from the Cells subgroup of the Home tab to insert rows, columns and cells in a worksheet

3.6.3 Auto Fills

MS Excel can insert a series of numbers automatically in worksheet. A series may consist of same number repetitively or numbers in increasing or decreasing in systematic manner. All series shown below are generated automatically in a worksheet by the MS Excel First fill two or three entries of the series manually and then select these entries. Once you point to the bottom right corner of the selected cell, the pointer changes to a Plus sign. Now, Drag the pointer to any number of cells, you wish to fill by series as per number pattern feed by you manually in first two/three entries. You get a new series of numbers in the selected cells as you release button of mouse. This function of Excel is known as auto fill.

3.6.4 Sorting Data

MS Excel can sort alphabetical, date, time and number data. Data can be arranged in ascending or descending order. First, highlight all the data including any headings of the data to be sorted. Click on Sort& Filter icon from the Editing subsection of the Home tab of the ribbon. MS Excel decides the ordering of data on the basis of attribute of data to be sorted. If selected data is numeric then it can be arranged in descending (largest to smallest) or in ascending (smallest to largest) value of data. If data consist of alphabets, then it can be arranged in ascending (A to Z order) or in descending order (Z to A order). If data consist of date or time then it can be arranged in oldest to newest or newest to oldest order. This is the easiest approach for sorting data, when sorting is based on single field. You can perform complex sorting of data in a worksheet by sorting it on more than one column at a time. When data is sorted on more than one field then data is first sorted on main field (first level) in ascending or descending order as instructed by you. The sorted entries of the first level are further sorted on second field in the order specified by you. Data can be sorted up to 64 levels. For example, if sorting is done as per the criteria specified in the sort dialogue box shown below. Data is first sorted on the basis of

Names of candidates in ascending order (A to Z order) and then sorted records on names are further sorted on the basis of age of candidate in descending order. Candidate with same names are arranged on ages of the candidates. Candidates with higher age are placed first then other candidate with same name

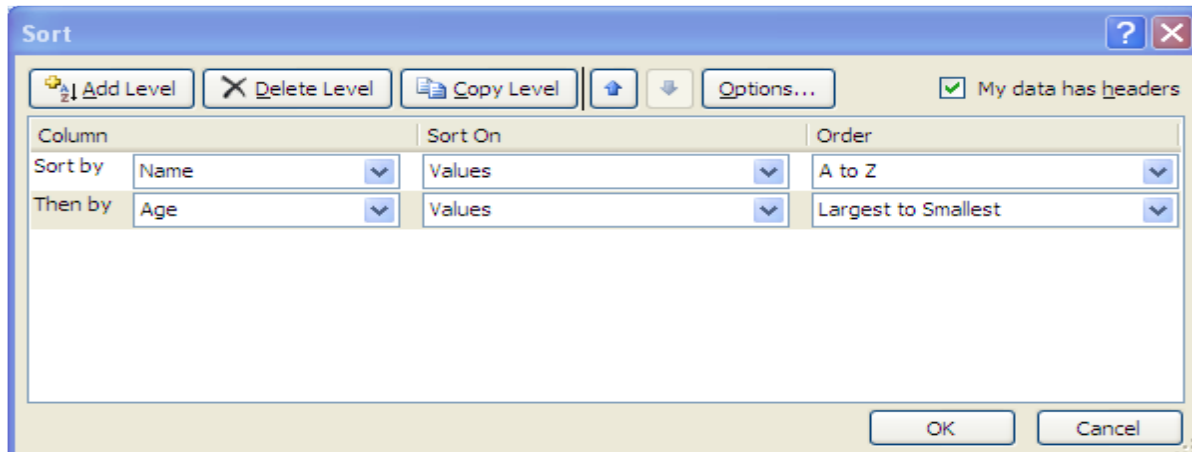


Figure E

Select a range of cells to be sorted, click on arrow key next to Sort & Filter in Editing sub-group on Home tab ribbon, and select Custom Sort from the cascade menu. The following sort dialog box is appeared on the screen.

1. Select column on the basis of which data is to be sorted in the Sort by combo box.
2. Under Sort On, select the type of sorting. The data is sorted on its value by default. However, data can be sorted on the format of field chosen in Sort by combo box, by selecting one of the options Cell Color, Font Color, or Cell Icon from the sort on combo box.
3. Under Order, select how to do sorting. Select one of the following:
 - For text values, select A to Z or Z to A.
 - For number values, select Smallest to Largest or Largest to Smallest.
 - For date or time values, select Oldest to Newest or Newest to Oldest.
 - To set other than previously mentioned criteria, select Custom List.
 - To add another column on the basis of which data will be further sorted. Click on Add Level command and then repeat early mentioned steps. The sorting at the second level changes order of only data, having same ranking on the first sorting criteria.
4. To copy shorting criteria specified at any level to new level of sort window, select entry at that level, and then click Copy Level.

5. To delete a sorting criterion, select the specific sorting criteria (level) and then click Delete Level.

3.7 Summery

The science, which deals with numbers, is statistics. It crunches the numbers and organizes them in a meaningful way so that information is generated. Computer can help immensely in the statistical analysis. There exists numerous statistical tools available and the need is to identify their actual usage. Most of the Statistical methods are based on the quantitative data. One can find different statistical packages for applications to different disciplines.

In this Unit you have read about package MS Excel. The Unit has discussed some of their applications in details. You have also gone through a brief introduction of some of the popular statistical software's. This Unit was intended to make you familiar with the basic statistical functions that can be performed with the help of computer and to arouse your interest in the beautiful and huge world of statistical computing.

3.8 Self-Assessment Questions

1. What are the various components of a Title bar?
2. What basic formatting buttons does MS Word provides
3. Create a new document and then password-protect it by setting a password

3.9 References

- www.support.office.com
- www.tutorialspoint.com
- www.chandoo.org
- www.ignou.ac.in
- www.contextures.com
- www.exinfm.com
- www.chicopee.mec.edu

- www.lacher.com
- www.jaxworks.com

Structure

- 4.1 Introduction
- 4.2 Objectives
- 4.3 Chart Prepare to MS Excel
 - 4.3.1 Column Charts
 - 4.3.2 Line Charts
 - 4.3.3 Bar Charts
 - 4.3.4 Area Charts
 - 4.3.5 Pie Charts
 - 4.3.6 Surface Charts
 - 4.3.7 Choose your Charts Wisely
 - 4.3.8 How to Insert a Chart
 - 4.3.9 Read a Chart
 - 4.3.10 Options Under the Chart Tools Design
- 4.4 Summery
- 4.5 Self-Assessment Questions
- 4.6 References

4.1 Introduction

The MS Excel is used for designing of electronic calculation worksheets. These worksheets are used to organize, compute, analyze and evaluate data and to do financial calculations. The Electronic workbooks are easy to prepare, manipulate and update. Electronic spreadsheet is better than manual spreadsheet. The amendment of electronic worksheet is convenient as you have to write a new manual worksheet when there are too many amendments. A number of calculations can be programmed and carried out automatically in the MS Excel. There are specialized functions to perform financial calculations and statistical analysis. The

terms worksheet, spreadsheet and workbook are used interchangeably by different books and authors for calculation worksheet prepared with MS-Excel.

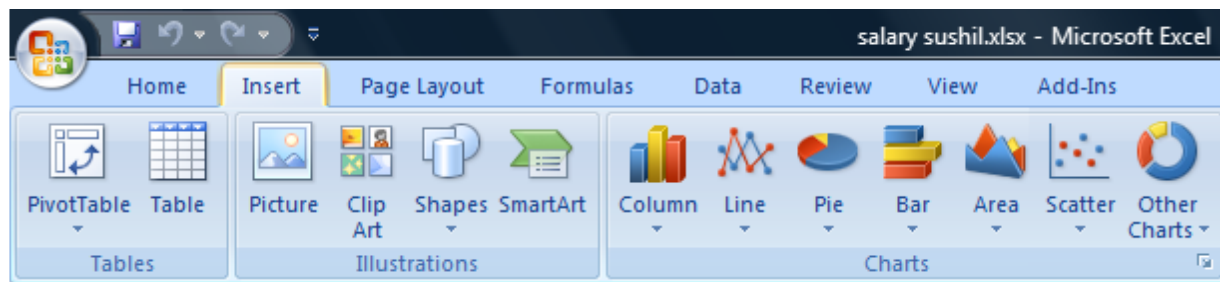
4.2 Objectives

MS Excel is an integral component of the MS Office. It is used for preparation of calculation worksheets, conduct data analysis and present data in the form of charts. After reading this unit, you will be able:

- Prepare Chart to MS Excel
- Formula and Functions of Statistics by MS Excel
- Describe Statistical Analysis with Solver by MS Excel
- Explain the Feature of some Statistical Packages in MS Excel

4.3 Chart Prepare to MS Excel

Charts are being used to create visual presentation of the data for the spreadsheet. Graphical view of data is very effective in understanding trends in continuous data over a long period without going in detail, even a non-professional can easily grasp message conveyed by our data through charts. You can create charts in MS Excel by using following procedure. First, enter all the data in a worksheet that will be displayed through of chart. Select the data to be displayed through chart through drag action and click on the Insert Tabor the ribbon, decide type of chart is to be prepared by clicking on arrow key of corresponding group icon from the Charts sub-group. Various kinds of charts are displayed as the drop down list. Choose one of chart by clicking over its icon.



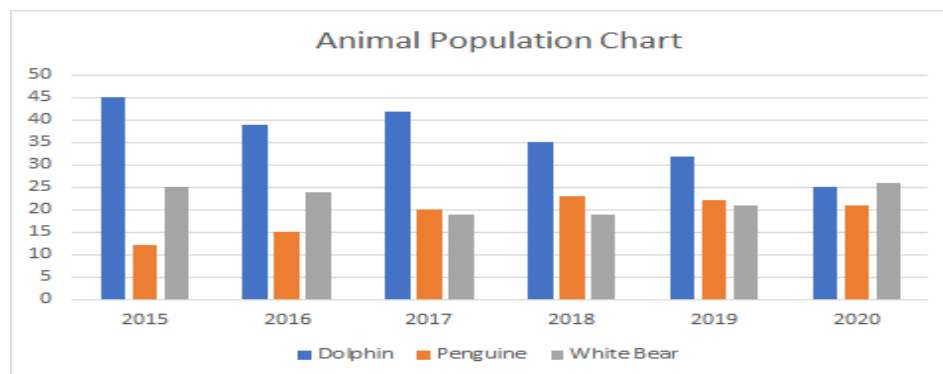
Excel offers many charts to represent the data in different manners, such as - Pie charts, Bar charts, Line charts, Stock charts, Surface charts, Radar charts, and many more. You can use them according to your data and analysis. All these charts there are a list of basic and advanced level of charts used for different purposes to interpret the data.

1. Column Chart
2. Line Chart
3. Bar Chart
4. Area chart
5. Pie chart or Doughnut chart
6. Surface chart

These are the most used charts of Excel that an Excel user usually requires. Microsoft Excel introduces one newer chart called Tree map chart for 2016 and newer version. It has come with some advanced features and representation styles. We will illustrate each chart and its functionality with an example in this chapter. Learn carefully and use them accordingly.

4.3.1 Column Charts

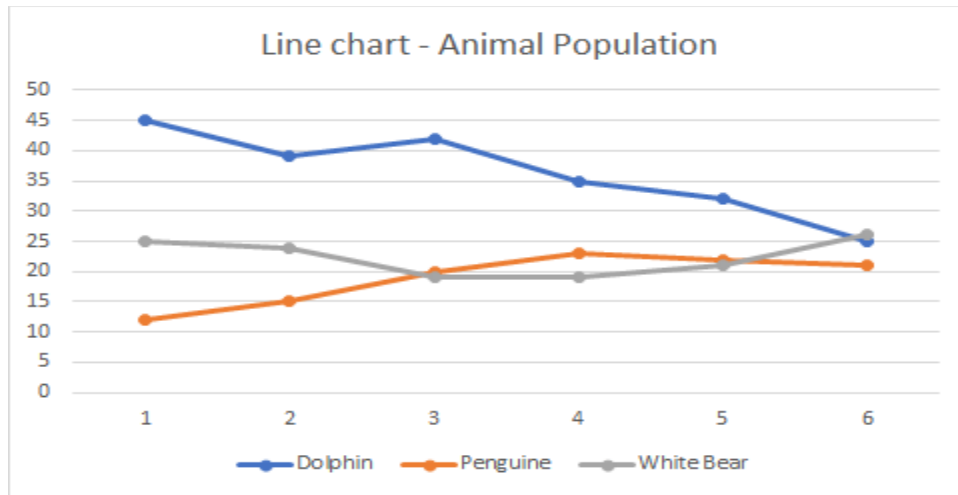
A column chart is basically a vertical chart that is used to represent the data in vertical bars. It works efficiently with different types of data, but it is usually used for comparing the information. For example, a company wants to see each month sell graphically and also wants to compare them. Column charts are best for it that helps to analyze and compare each month's data with each other.



2D and 3D Column Charts

4.3.2 Line Chart

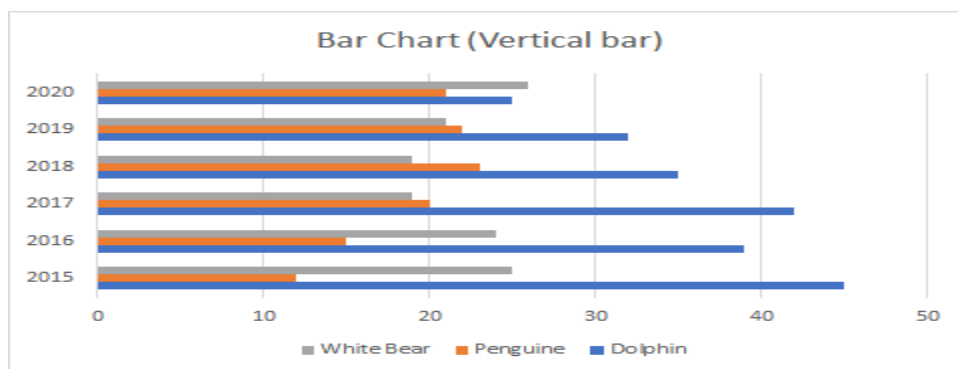
Line charts are most useful for showing trends. Using this chart, you can easily analyze the ups and downs in your data over time. In this chart, data points are connected with lines. For example, a company wants to analyze the cell of products for the last five years graphically. Additionally, it also wants to analyze the ups and downs of each year product sell.



2D and 3D line charts.

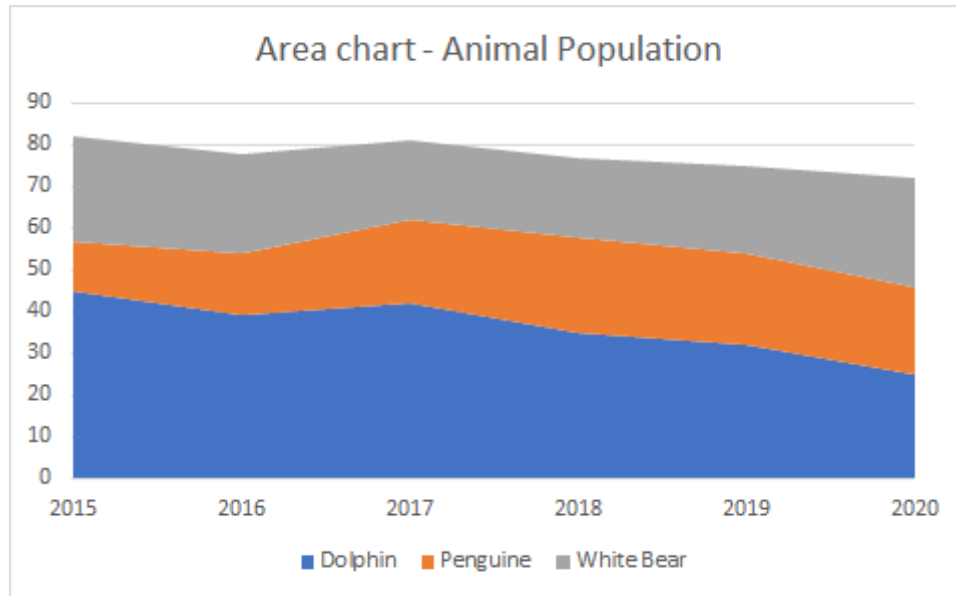
4.3.3 Bar chart

Bar charts are horizontal bars that work like column charts. Unlike column charts, Bar charts are horizontally plotted. Or you can say that bar charts and column charts are just opposite to each other. For example, a company uses the bar chart to analyze the data through vertical bars to represent the data graphically. You can see as well as compare the values to each other, respective to data.



4.3.4 Area Chart

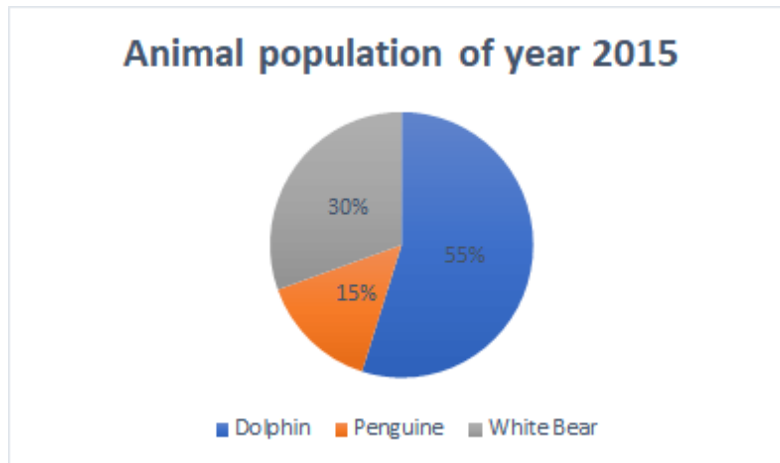
Area charts are just like line charts. Unlike the line charts, gaps are filled with color in area charts. Area charts are easy to analyze the growth in business as its shows ups and downs through line.



Similar to the line charts, data points in area charts are connected with lines.

4.3.5 Pie Chart

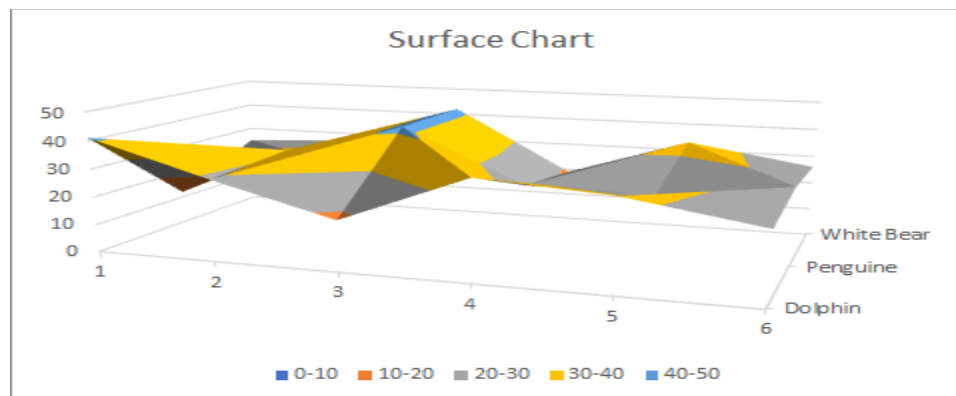
A pie chart is a rounded shape graph that is divided into slices of pie. Using this chart, you can easily analyze data that is divided into slices. It makes the data easy to compare the proportion.



Pie charts make it easy to analyze which values make up the percentage of whole. Pie chart is also known as Doughnut chart. Excel offers 2D and 3D pie charts.

4.3.6 Surface Chart

Surface chart is actually a 3D chart that helps to represent the data into a 3D landscape. These charts are best to use with a large dataset. This chart allows to displaying a variety of data at the same time.



A large dataset is not easy to represent using other charts. Surface chart solve this problem that allows displaying large datasets using this 3D chart.

4.3.7 Choose your Charts wisely

Excel offers too many charts as well as their 2D and 3D type. You can use any of them but choose them wisely according to your data. Different scenario requires different charts.

Though, it can display all and correct information. We have a list of some points for each type of chart that helps you to choose the chart wisely. Read them carefully -

Chart Type		When to choose this chart
1.	Column Chart	Use the column chart when you want to compare the multiple values across a few categories. The values are shown through vertical bars.
2.	Line Chart	Choose this chart when you want to show the trends (ups and downs) over a period of time, like for months or years.
3.	Bar Chart	Like the column chart, use this chart to compare the values across a few categories. In this chart, values are displayed in the horizontal bar.
4.	Area chart	Area chart has the same pattern as the line chart. This chart is best to use for indicating a change among different sets.
5.	Pie or Doughnut chart	Pie chart is best to use when you want to quantify the values and show them as percentage.
6.	Surface chart	Surface chart is different than other charts. Use it when you need to analyze the optimum combination between two sets of data.

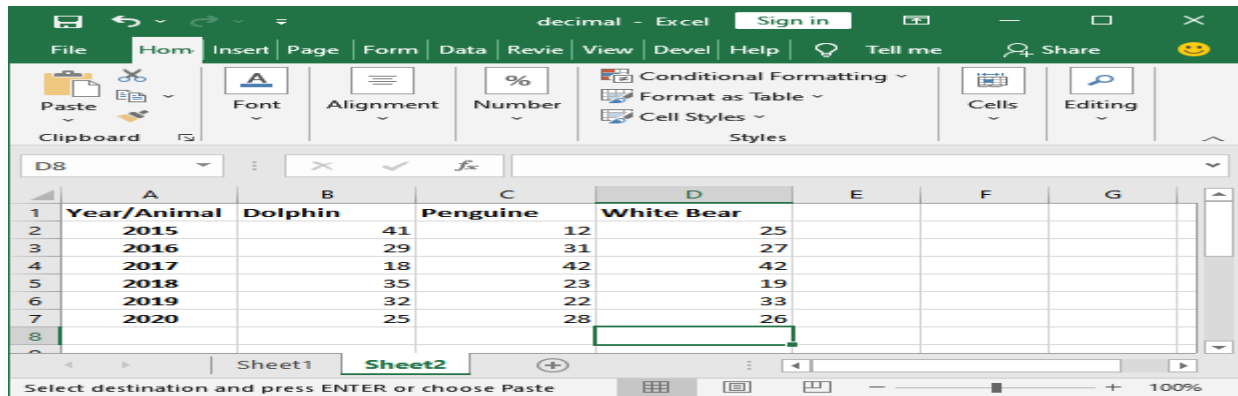
4.3.8 How to Insert a Chart?

Excel enables easy to use user interface using which you can easily insert a required chart for your data.

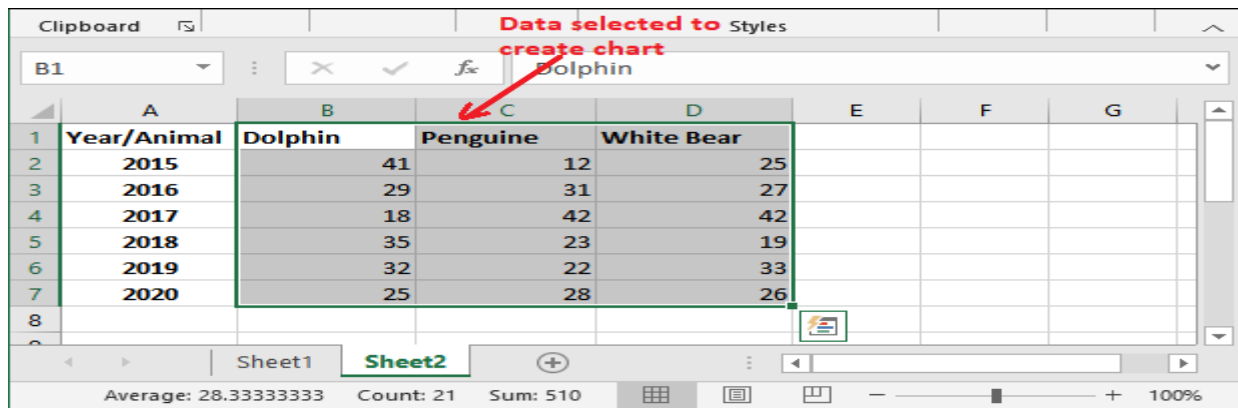
We need to follow few simple steps, Excel > Insert tab > Chart section > choose a chart.

We will follow these steps from start to end for creating the chart for Excel data. Following are the steps to insert a chart in Excel.

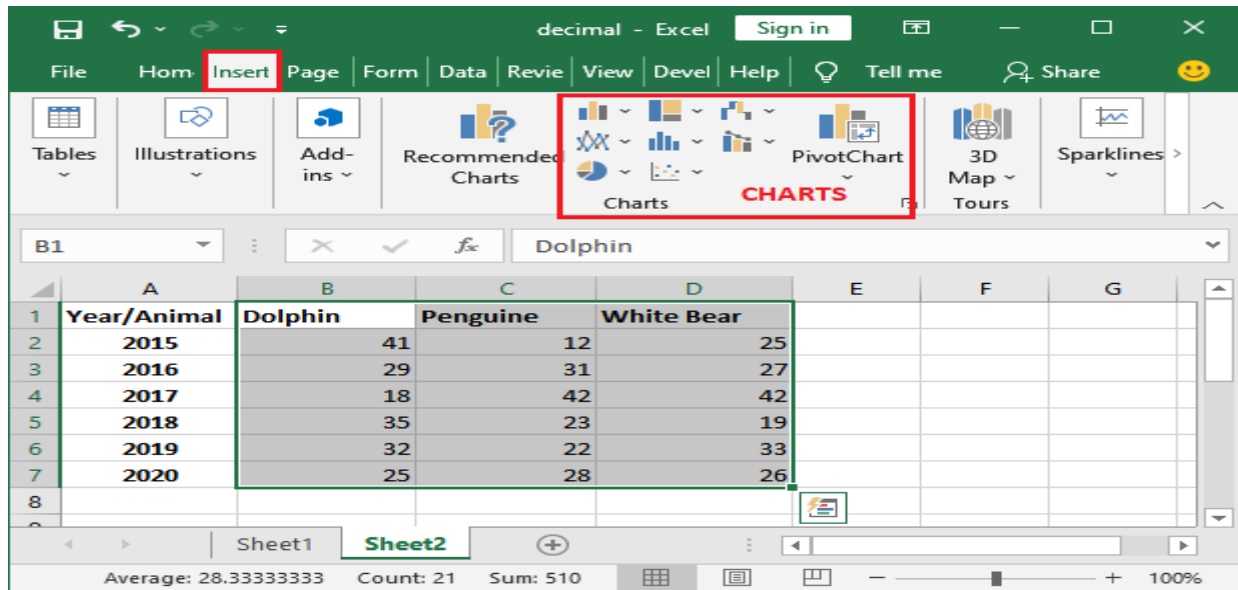
Step 1: We have the following dataset (Animal population rate for six years from 2015-2020) for which you want to create a chart in Excel.



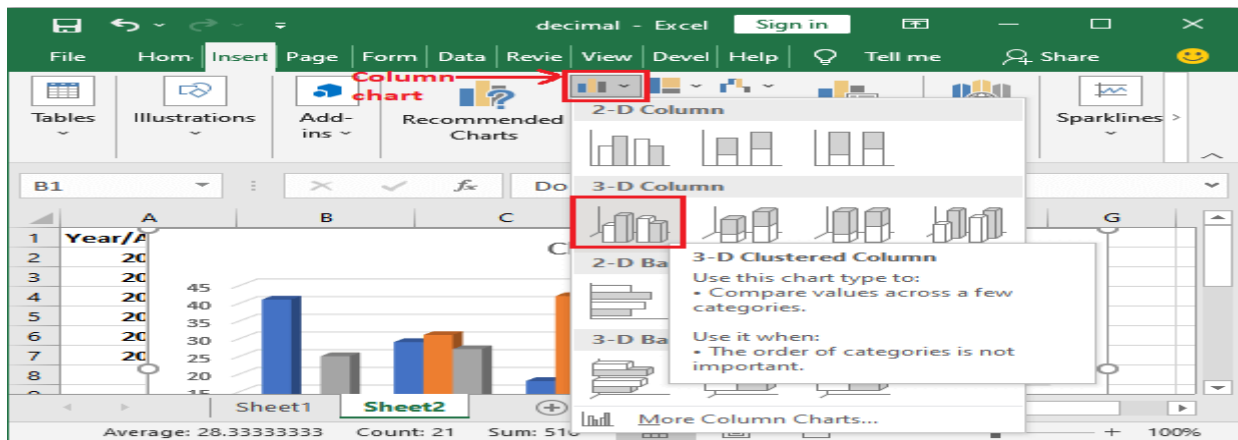
Step 2: Select the data, including column header and row label for which you want to create a chart. This data will be the source data for your chart.



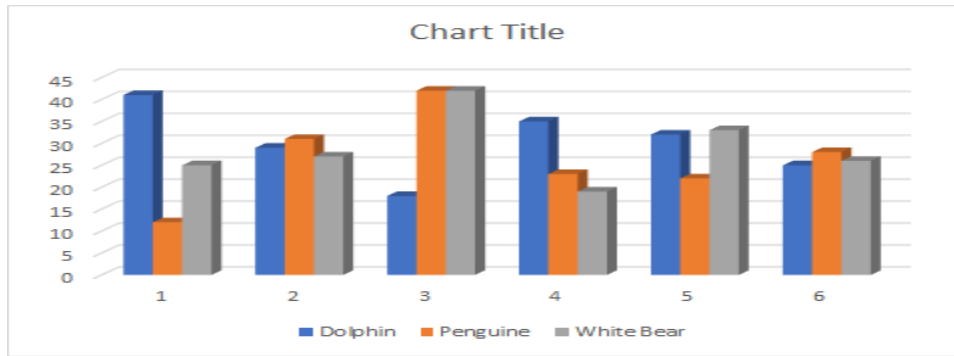
Step 3: Navigate to the Insert tab in the Excel header, where you will see a charts section that contains a list of all these charts.



Step 4: Choose a chart from here according to your data. We have chosen a 3D Column chart containing vertical bars for your data.

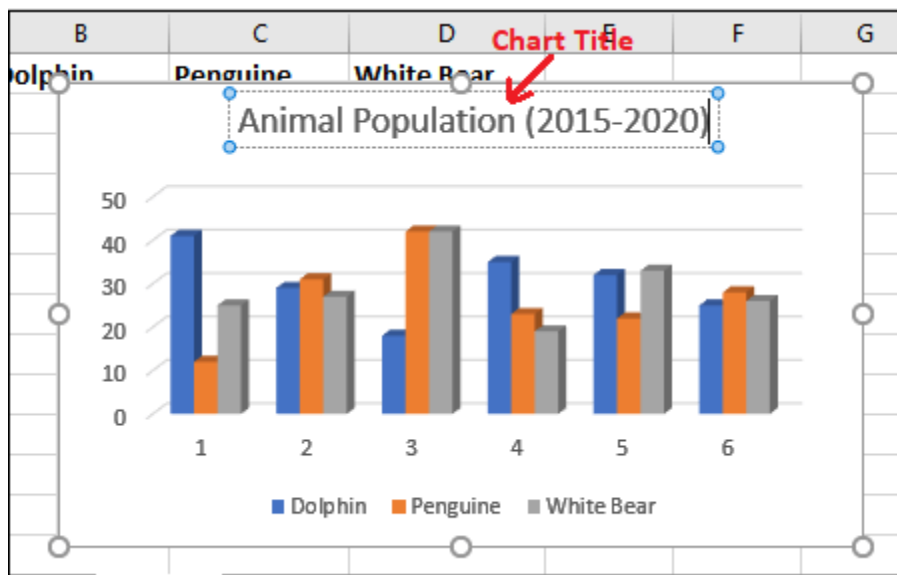


Step 5: The selected chart is inserted into your Excel worksheet. Initially, the chart looks like this for the data selected in step 2.



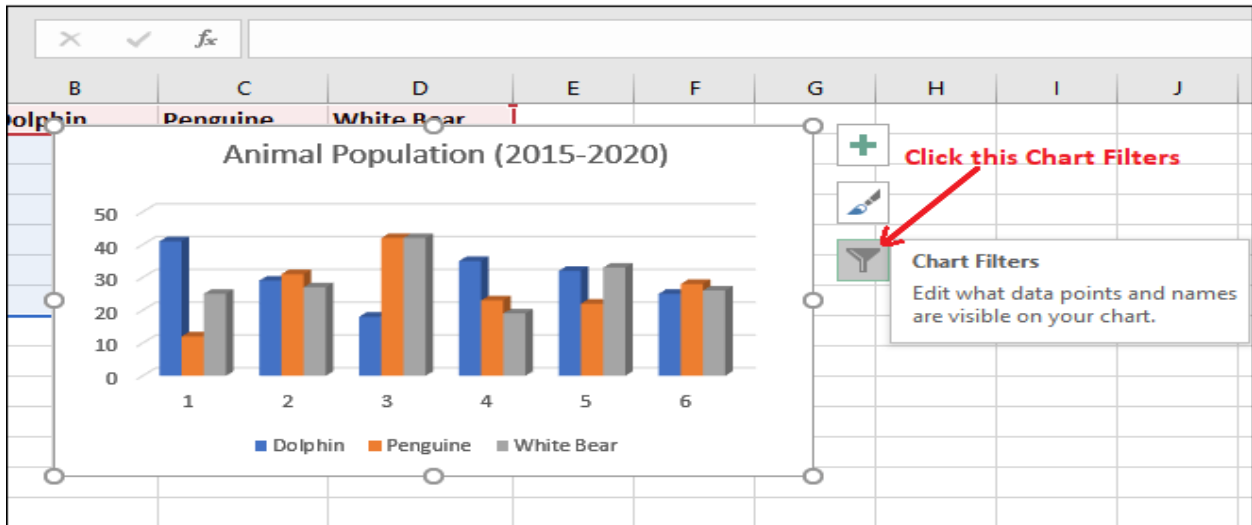
Currently, this chart does not have a valid title, clear values for analysis, and more. You can set up all these things in your chart by modifying it.

Step 6: Double-tap on the Chart Title to make it editable and then provide a new valid title that suites to it.

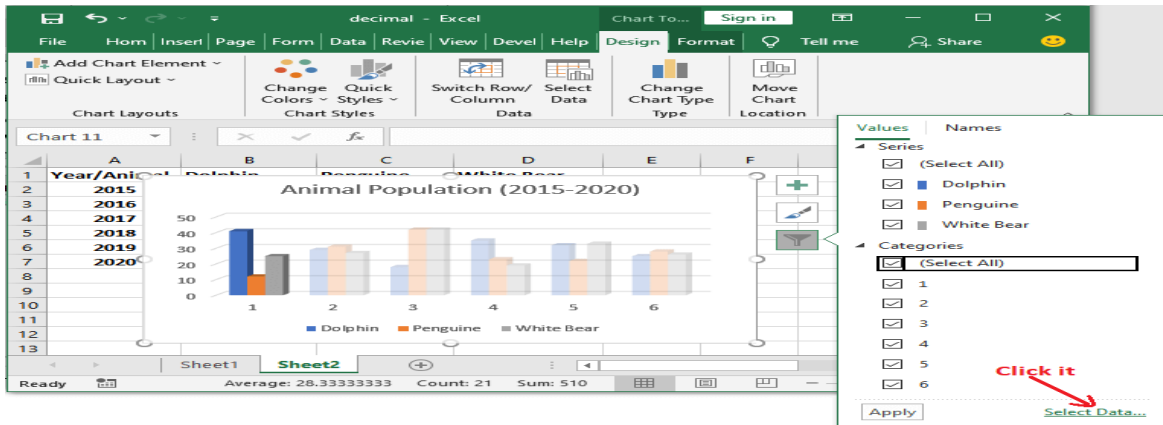


Here, Blue color vertical bar is representing to Dolphin, Orange vertical bar to Penguin, and Grey vertical bar to White Bear population count.

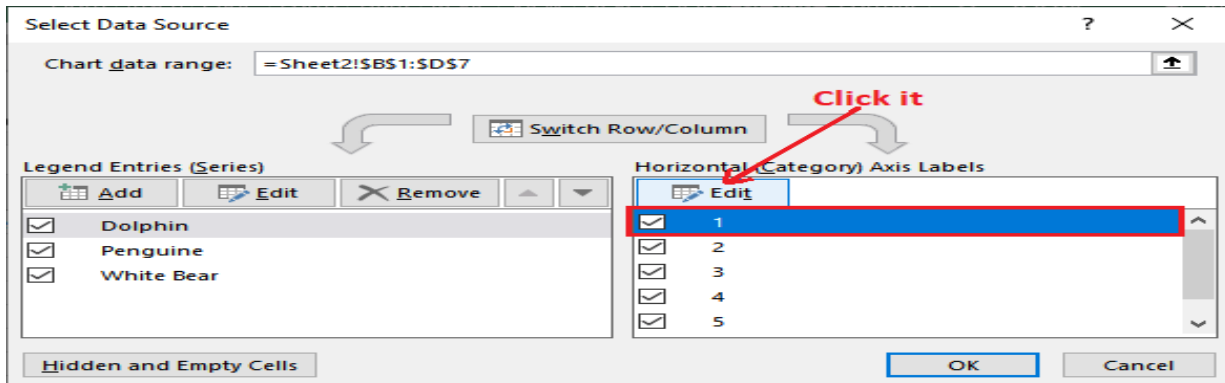
Step 7: You can also define each vertical bar for its year so that the user can easily analyze the values. Click on the Chart Filters icon here.



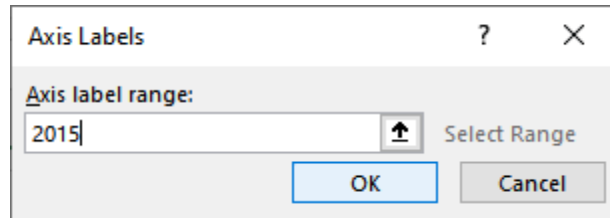
Step 8: Click on the Select data present at the bottom of the list to replace the years 2015 for each vertical bar.



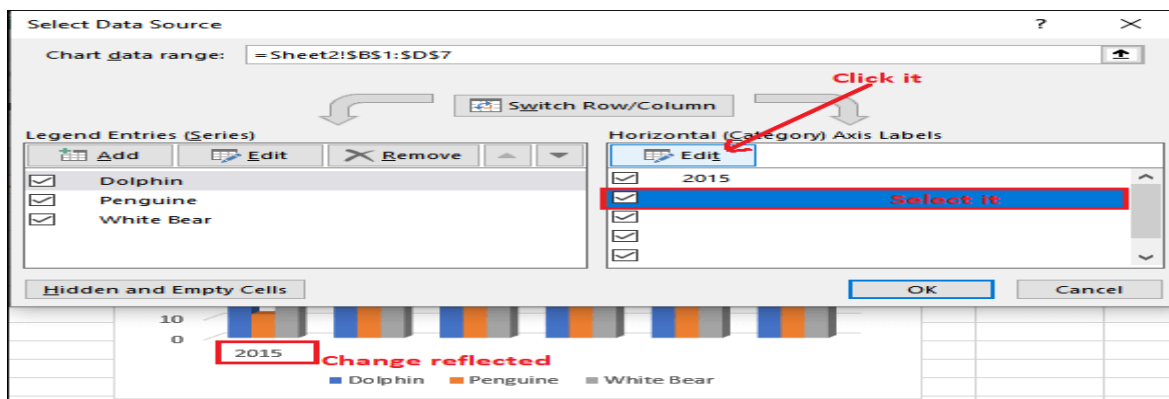
Step 9: Here, select the number 1 to replace it with year 2015 and click the Edit button.



Step 10: Enter the year and click OK.

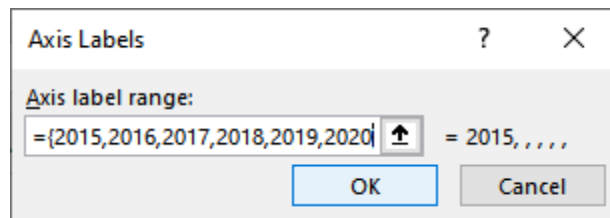


Step 11: The year 2015 will immediately reflect on the chart, and all other become blank. Now, to put all other years for each vertical bar, click one more time here.

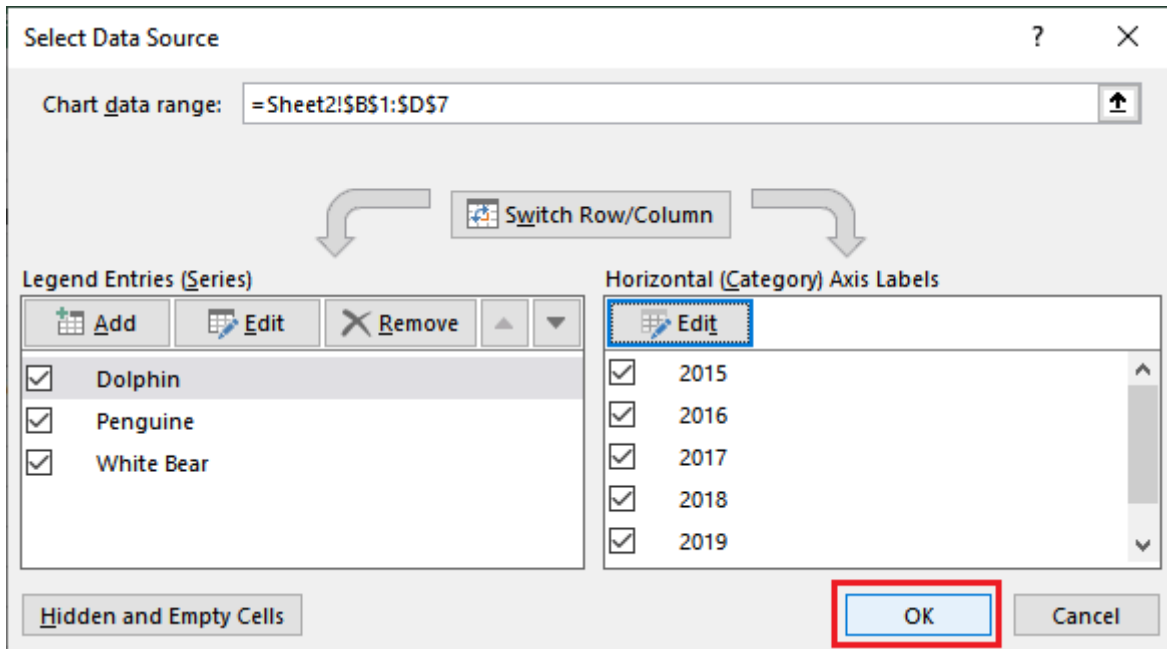


Step 12: Add more years from 2016 to 2020 inside curly braces separated by a comma and click OK.

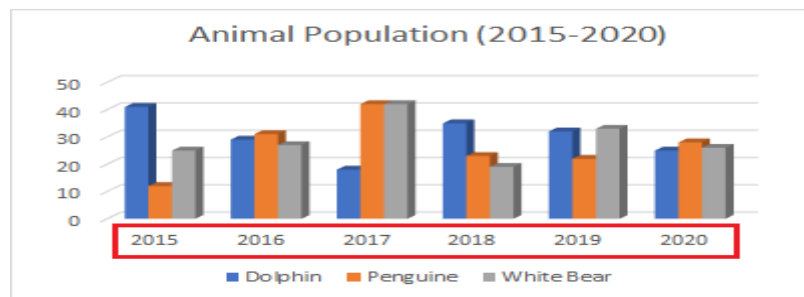
{2015,2016,2017,2018,2019,2020}



Step 13: All values are now added. So, click OK.



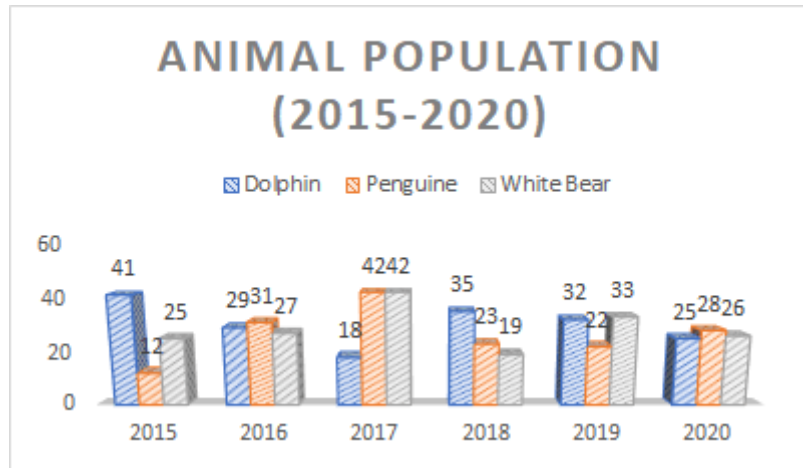
Step 14: See the charts that the years are reflected on the chart correspond to each vertical bar.



We can see that the exact value is not defined at the end of each bar. Only the graph is showing. Excel enables the users to choose a detailed bar.

Step 15: Choose another chart style for the Column chart for detailed description from the Chart Style in the ribbon.

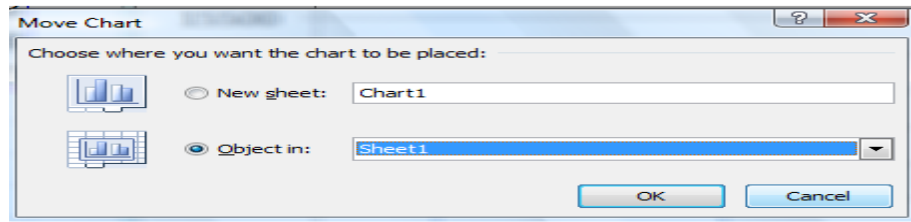
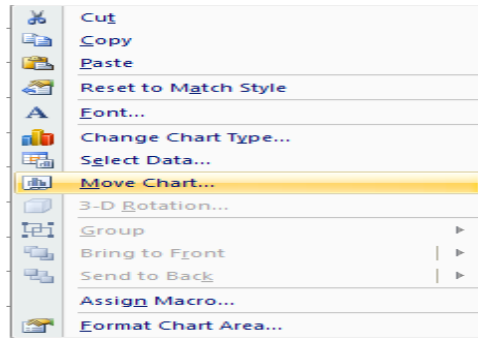
Step 16: You can now see the exact value is also showing for each bar in this chart.



Similarly, you can insert, create, and modify other charts in your Excel worksheet for your data.

4.3.9 Read a chart

You should learn how to read a chart to understand and analyze the data presented through it. It is most important to learn how to read a chart. A chart contains several components, such as elements and parts that help you to interpret the data. This is one of the most important points because if you are unable to read and understand the chart, you might misinterpret the value and analyze the data. The chart will be inserted within the current spreadsheet at the position of the cursor. You can quickly create a chart with the support of shortcut. First, select the data to be displayed in the chart then press Alt + F1 or F11 shortcut to draw the chart. MS Excel uses the default chart type when shortcuts are used. When you press Alt + F1 shortcut after selection of data, the chart is inserted in the current worksheet. When you press F11 after selection of data the chart is displayed on a separate worksheet. You can resize or move the chart to new place within the same worksheet by drag action of mouse. You have already learnt skills of moving or resizing pictures in earlier unit. The same skill can be applied here to move and resize the chart. Right click on the border of chart to bring up a pop-up menu having command to improve presentation of chart and value of data displayed there in. Select Move Chart option from the pop-up menu to move chart to new location. A dialogue box will appear to know whether chart will be moved in one of the existing worksheets or to be moved to new worksheet. The pop-up menu on right click over the chart and dialogue window display on the selection of the Move Chart option is shown on right side in the figure

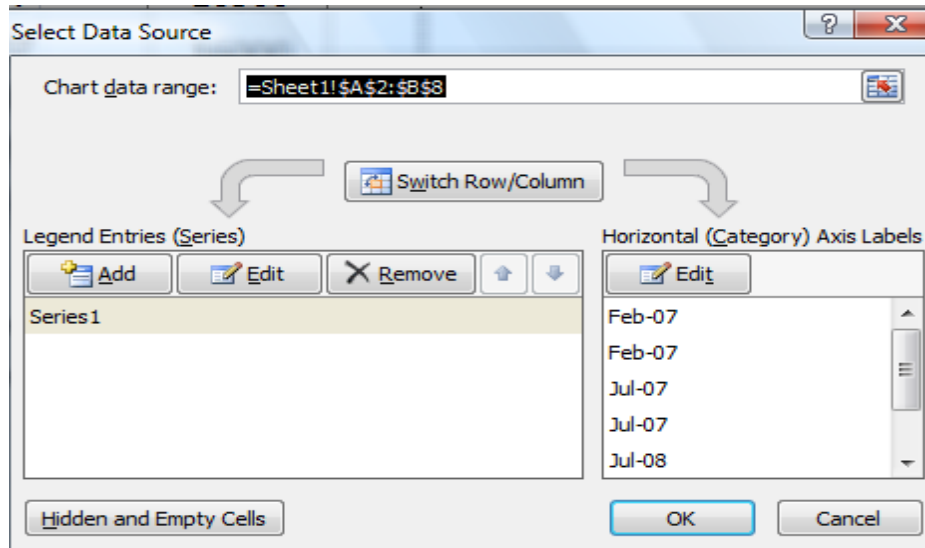


When you click over a chart, Chart Tools Design tab is displayed on the ribbon. There are various institutions features and styles in the Chart Tools Design tab to improve the presentation of chart and change values of data presented in chart.

4.3.10 Options under the Chart Tools Design

If you had made mistake in selection of data for chart or designing of chart earlier, you can bring desirable effect in the chart without redesign chart from the beginning with the help of various options available under the Chart tools.

Change Chart Type: This option is available under Type subgroup, which allows you to select one of the predefined chart types by clicking on its icon. Other details in the chart such as data values, title of charts and axis etc. remain unchanged after changing of the chart type. Select Data If you want to change the data to be depicted in the chart then click of Select data icon under Data sub group of the Design tab of the ribbon. You will get the following dialogue box to redefine data value for the chart



Horizontal (Category) Axis Labels list displays the label of the X-axis of the chart in sequential manner. The entries reflected on top of the list are displayed next to intersection of X and Y-axis, following entries in the list is displayed in sequential manner on X-axis of the chart. If you have not assigned label to horizontal axis than horizontal axis is assigned labels by default as 1, 2, 3, and so on. Legend Entries (Series) label in Select Data Source window (fig. 20) display a series of data reflected on the Chart. You can edit existing series or add new series to the chart. Select the series to be edited from the list and click on edit you will get the following window. Assign a name to the series in the textbox of Series name: label. Modify the series values either typing the range of series values manually, as you have learnt in feeding formulas to worksheet (Starting range followed by “:” then ending range) or click on rectangle icon at the right end of the textbox of Series values: label, select the range of cells through drag action. For example, value =sheet!\$B2\$2:\$B\$6 in the range box indicates that chart will display data from b2 to b16 range in sheet 1. You can simultaneously display two or more series on the same chart to make comparison of data. Number of series in a chart depends on type of chart. For example pie chart displays only one series and a line chart can display many series at a time. Click on Add button in the dialogue box shown the Figure 20, to define a new series assign a name and define range for the new series. Once you have selected ranges to be displayed and labels for horizontal axis, click on OK Button to apply change into the chart. Chart Layouts You can improve the presentation of chart by adding chart title, labels for axis, series legends, major and minor axis. These features can be easily added by selecting one of the predefined layouts from the Chart

Layouts subgroup of the Chart Tools Design tab. Different Layouts are displayed in miniature form under these options. Chart Styles This subgroup of functions improve the style of chart by changing colors to lines, areas, bars or background areas of the chart. Make sure that your printer should support colored printing otherwise design chart in mono-color.

4.4 Summary

This unit describes the essential features of Microsoft Excel. After studying this unit, we will be able to create Excel sheet documents, open the excel sheet documents that were earlier created save the documents as well as perform simple operations. The unit also describes operations such as including text, graphics, tables, clip artistic. It also describes the process of creation of different kinds of documents such as simple notes and memos as well as multi-column reports with tables, graphics data in different ways with click of button. In this Unit, we learned some great features like type of dashboards, benefits, getting your data ready, different features of excel which help to make our dashboard work like grouping, table, Charts, Slicers, etc. Converting data in to Table made our Pivot table dynamically linked to source data while appending. Slicers and timelines (Excel 2013 feature) are great features to make our dashboard dynamic. A statistical package is defined as the software used to collect, organize, interpret and present numerical information. The need of a statistical package arises due to the complexity of calculations involved therein for analysis and inference. It helps to bring accuracy in results.

4.5 Self-Assessment Questions

1. How can you sum up the Rows and Column number quickly in the Excel sheet?
2. How can you resize the column?
3. What are charts in MS-Excel?
4. What is ribbon?

4.6 References

- www.support.office.com
- www.tutorialspoint.com

- www.chandoo.org
- www.ignou.ac.in
- Greg Harvey, Excel 2010 For Dummies, Wiley Publishing, Inc., Indianapolis, Indiana, 2010.
- Debra Dalglish, Beginning Pivot Tables in Excel2007, A press, 2011.
- www.contextures.com

UNIT: 5**COMPUTATIONS WITH MS EXCEL - PART III**

Structure

- 5.1 Introduction
- 5.2 Objectives
- 5.3 Formula and Functions of Statistics by MS Excel
 - 5.3.1 SUM
 - 5.3.2 AVERAGE
 - 5.3.3 COUNT
 - 5.3.4 SUB TOTAL
 - 5.3.5 MODULES
 - 5.3.6 POWER
 - 5.3.7 CEILING
 - 5.3.8 FLOOR
 - 5.3.9 CONCATENATE
 - 5.3.10 LEN
 - 5.3.11 REPLACE
 - 5.3.12 SUBSTITUTE
 - 5.3.13 LEFT, RIGHT, MID
 - 5.3.14 UPPER, LOWER, PROPER
 - 5.3.15 NOW ()
 - 5.3.16 TODAY ()
 - 5.3.17 TIME ()
 - 5.3.18 HOUR, MINUTES, SECOND
 - 5.3.19 DATEDIF
 - 5.3.20 VLOOK UP
 - 5.3.21 HLOOK UP
- 5.4 Statistical Analysis with solver by MS Excel
- 5.5 Feature of Some Statistical Packages
- 5.6 Describe the Features of Statistical Package MS-Excel

- 5.7 Summary
- 5.8 Self-Assessment Questions
- 5.9 References

5.1 Introduction

In this unit will cover all topics from MS Excel to advance, A number of calculations can be programmed and carried out automatically in the MS Excel. There are specialized functions to perform financial calculations and statistical analysis. The terms worksheet, spreadsheet and workbook are used interchangeably by different books and authors for calculation worksheet prepared with MS-Excel. Here is some of main usage of the MS Excel.

1. Do statistical analysis and survey,
2. Organize large volume of data and records, and Automatic computation of complex calculations.

The MS Excel workbook consists of grids of cells, arrange in tabular form as crossing of rows and columns. Each cell is an independent unit for keeping records. Data store in cells may be numeric, string, date or time

5.2 Objectives

MS Excel is a component of the MS Office. It is used for preparation of calculation worksheets, conduct data analysis and present data in the form of charts. After reading this unit, we will be able:

- describe the features of statistical packages MS-Excel
- Statistical Analysis with solver by MS Excel
- Feature of Some Statistical Packages.

5.3 Formula and Functions of Statistics by MS Excel

There are plenty of Excel formulas and functions depending on what kind of operation you want to perform on the dataset. We will look into the formulas and functions on mathematical operations, character-text functions, data and time, sum if-count if, and few lookup

functions. Let's now look at the top 25 Excel formulas you must know. In this section we have categorized 25 Excel formulas based on their operations. Let's start with the first Excel formula on our list. In Microsoft Excel, a formula is an expression that operates on values in a range of cells. These formulas return a result, even when it is an error. Excel formulas enable you to perform calculations such as addition, subtraction, multiplication, and division. In addition to these, we can find out averages and calculate percentages in excel for a range of cells, manipulate date and time values, and do a lot more. Formulas are used to perform calculation automatically. When we apply a formula to a cell, its value mathematically depends upon values of other fields. For example formula $c3 = b3 + a3$, represents that value of c3 is sum of values placed in b3 and a3 fields. As you change value of b3 or/and a3 cells, the value of c3 is changed automatically. Every formula begins with an equals sign (=) or plus sign (+). These are recognized as reserved words for inputting formula in a worksheet. Click on the cell, where formula is to enter. Enter = or + sign before entering a formula in the formula bar and then type the formula. For example for performing $c3 = b3 + a3$ calculation take the follow steps. First, enter values in b3 and a3 field. Move focus to c3 cell by clicking over it. Now enter the following string in the formula text box "= b3 + a3" and press enter; sum of b3 and a3 is now available in c3. Instead of writing formula just enter = or + symbols in the formula bar; take mouse to b3 cell and click; b3 is display on formula bar. Now entre + symbol from the keyboard, it is displayed on formula bar next to b3, now click on a3 cell and your formula is ready (type „=“; click on b3; type +; click on a3). Thus, you can either directly write address or select it by clicking on specific cell to insert it in formula. If you have assigned names to cells as explained in sub section 8.10, then you can use names of cells instead of their address. If a formula is not proceeded by + or = sign then this formula will be treated as a string. Some time formulas use standard functions instead of simple mathematical calculation. The MS Excel defines the following procedure to do the calculation.

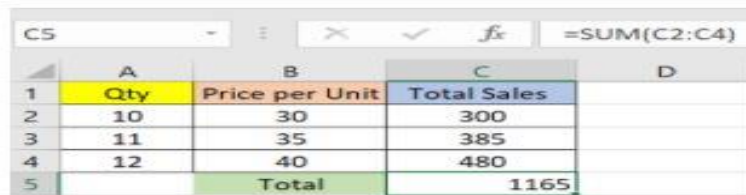
- Choose a cell.
- To enter an equal sign, click the cell and type =.
- Enter the address of a cell in the selected cell or select a cell from the list.
- You need to enter an operator.
- Enter the address of the next cell in the selected cell.

- Press enter

There is another term that is very familiar to Excel formulas, and that is "function". The two words, "formulas" and "functions" are sometimes interchangeable. They are closely related, but yet different. A formula begins with an equal sign. Meanwhile, functions are used to perform complex calculations that cannot be done manually. Functions in excel have names that reflect their intended use. There are plenty of Excel formulas and functions depending on what kind of operation you want to perform on the dataset. We will look into the formulas and functions on mathematical operations, character-text functions, data and time, sum if-count if, and few lookup functions. Let's now look at the top 25 Excel formulas you must know. In this article, we have categorized 25 Excel formulas based on their operations. Let's start with the first Excel formula on our list.

5.3.1 SUM

The SUM () function, as the name suggests, gives the total of the selected range of cell values. It performs the mathematical operation which is addition. Here's an example of it below:
Sum =SUM" (C2:C4)"



	A	B	C	D
1	Qty	Price per Unit	Total Sales	
2	10	30	300	
3	11	35	385	
4	12	40	480	
5		Total	1165	

Fig: Sum function in Excel

As you can see above, to find the total amount of sales for every unit, we had to simply type in the function “= SUM (C2:C4)”. This automatically adds up c2, c3, and c4. The result is stored in C5.

5.3.2 AVERAGE

The AVERAGE () function focuses on calculating the average of the selected range of cell values. As seen from the below example, to find the average of the total sales, we have to simply type in:

AVERAGE =AVERAGE(C2, C3, C4).

	A	B	C	D	E
1	Qty	Price per Unit	Total Sales		
2	10	30	300		
3	11	35	385		
4	12	40	480		
5		Total	1165		
6		Average	388.3333333		

Fig: Average function in Excel

It automatically calculates the average, and you can store the result in your desired location.

5.3.3 COUNT

The function COUNT() counts the total number of cells in a range that contains a number. It does not include the cell, which is blank, and the ones that hold data in any other format apart from numeric.

COUNT =COUNT(C1:C4)

	A	B	C	D
1	Qty	Price per Unit	Total Sales	
2	10	30	300	
3	11	35	385	
4	12	40	480	
5		Count	3	

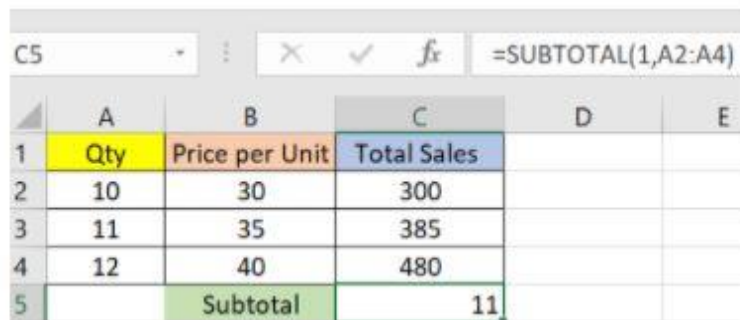
Fig: Microsoft Excel Function - Count

As seen above, here, we are counting from C1 to C4, ideally four cells. But since the COUNT function takes only the cells with numerical values into consideration, the answer is 3 as the cell containing “Total Sales” is omitted here. If you are required to count all the cells with numerical values, text, and any other data format, you must use the function ‘COUNTA()’.

However, COUNTA() does not count any blank cells. To count the number of blank cells present in a range of cells, COUNTBLANK() is used.

5.3.4 SUBTOTAL

Moving ahead, let's now understand how the subtotal function works. The SUBTOTAL () function returns the subtotal in a database. Depending on what you want, you can select average, count, sum, min, max, min, and others. Let's have a look at two such examples.



	A	B	C	D	E
1	Qty	Price per Unit	Total Sales		
2	10	30	300		
3	11	35	385		
4	12	40	480		
5		Subtotal	11		

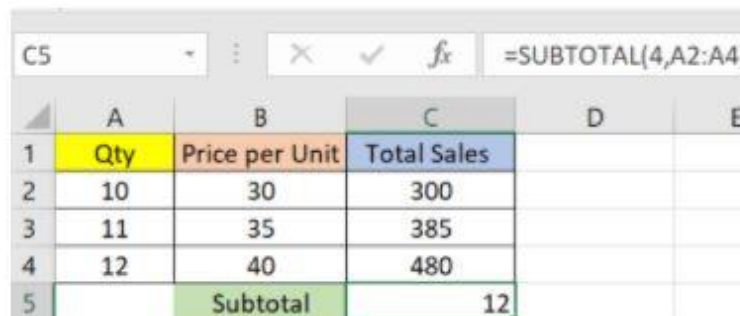
Fig: Subtotal function in Excel

In the example above, we have performed the subtotal calculation on cells ranging from A2 to A4. As you can see, the function used is

`SUBTOTAL =SUBTOTAL(1, A2: A4)`

In the subtotal list “1” refers to average. Hence, the above function will give the average of A2: A4 and the answer to it is 11, which is stored in C5. Similarly, “=SUBTOTAL(4, A2: A4)”

This selects the cell with the maximum value from A2 to A4, which is 12. Incorporating “4” in the function provides the maximum result.



	A	B	C	D	E
1	Qty	Price per Unit	Total Sales		
2	10	30	300		
3	11	35	385		
4	12	40	480		
5		Subtotal	12		

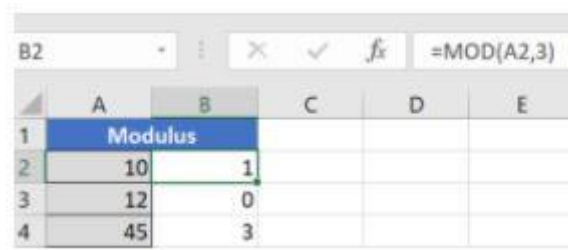
Fig: Count function in Excel

5.3.5 MODULUS

The MOD () function works on returning the remainder when a particular number is divided by a divisor. Let's now have a look at the examples below for better understanding. Example, we have divided 10 by 3. The remainder is calculated using the function

`MODULUS =MOD(A2,3)`

The result is stored in B2. We can also directly type “=MOD(10,3)” as it will give the same answer.



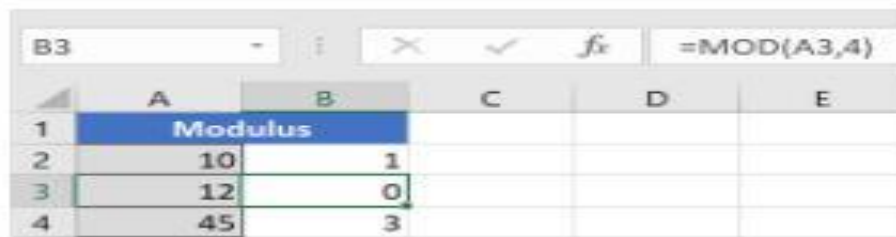
The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E
1	Modulus				
2	10	1			
3	12	0			
4	45	3			

The formula bar at the top shows the formula for cell B2: `=MOD(A2,3)`.

Fig: Modulus function in Excel

Similarly, here, we have divided 12 by 4. The remainder is 0, which is stored in B3.



The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E
1	Modulus				
2	10	1			
3	12	0			
4	45	3			

The formula bar at the top shows the formula for cell B3: `=MOD(A3,4)`.

Fig: Modulus function in Excel

5.3.6 POWER

The function “Power ()” returns the result of a number raised to a certain power. Let's have a look at the examples shown below:

	A	B	C	D	E
1	Power				
2	10	1000			
3	4	256			
4					

Fig: Power function in Excel

As you can see above, to find the power of 10 stored in A2 raised to 3, we have to type:

Power =POWER (A2,3)

This is how power function works in Excel.

5.3.7 CEILING

Next, we have the ceiling function. The CEILING() function rounds a number up to its nearest multiple of significance.

	A	B	C	D	E
1	Ceiling				
2	35.316	40			

Fig: Ceiling function in Excel

The nearest highest multiple of 5 for 35.316 is 40.

5.3.8 FLOOR

Contrary to the Ceiling function, the floor function rounds a number down to the nearest multiple of significance.

	A	B	C	D	E
1	Floor				
2	35.316	35			

Fig: Floor function in Excel

The nearest lowest multiple of 5 for 35.316 is 35.

5.3.9 CONCATENATE

This function merges or joins several text strings into one text string. Given below are the different ways to perform this function. example, we have operated with the syntax:

CONCATENATE =CONCATENATE(A25, " ", B25)

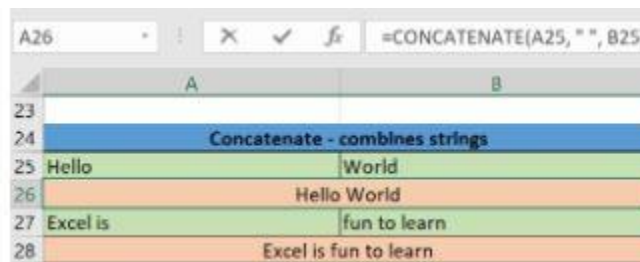


Fig: Concatenate function in Excel

In example, we have operated with the syntax:

"=CONCATENATE(A27&" "&B27)"

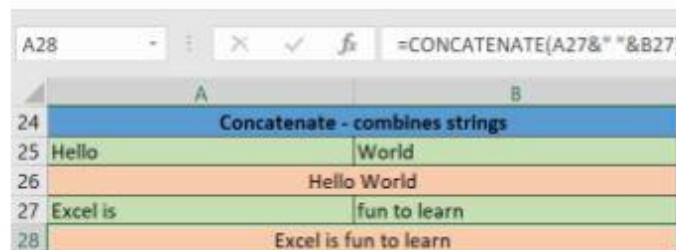


Fig: Concatenate function in Excel

Those were the two ways to implement the concatenation operation in Excel.

5.3.10 LEN

The function LEN() returns the total number of characters in a string. So, it will count the overall characters, including spaces and special characters. Given below is an example of the Len function.

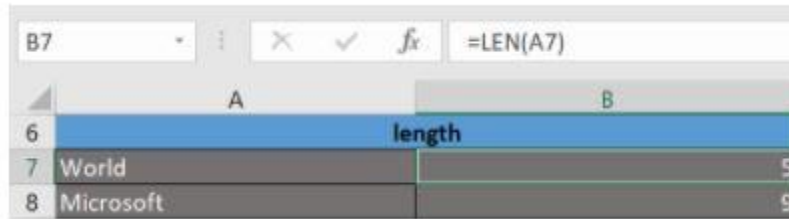


Fig: Len function in Excel

Let's now move onto the next Excel function on our list of this article.

5.3.11 REPLACE

As the name suggests, the REPLACE () function works on replacing the part of a text string with a different text string.

The syntax is “=REPLACE(old text, start num, num chars, new text)”. Here, start num refers to the index position you want to start replacing the characters with. Next, num chars indicate the number of characters you want to replace.

Let's have a look at the ways we can use this function. Here, we are replacing A101 with B101 by typing

REPLACE =REPLACE(A15,1,1,"B")

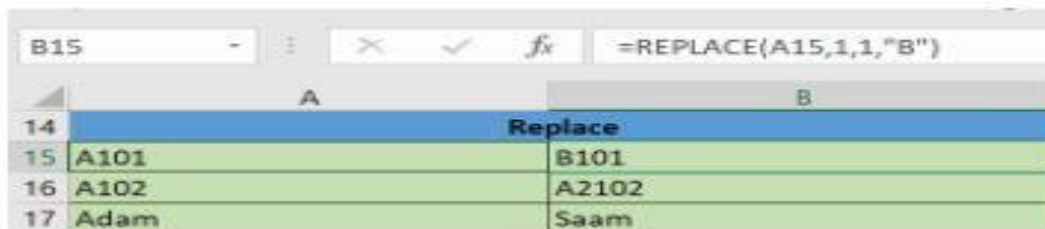


Fig: Replace function in Excel

Next, we are replacing A102 with A2102 by typing: “=REPLACE (A16,1,1, "A2")”

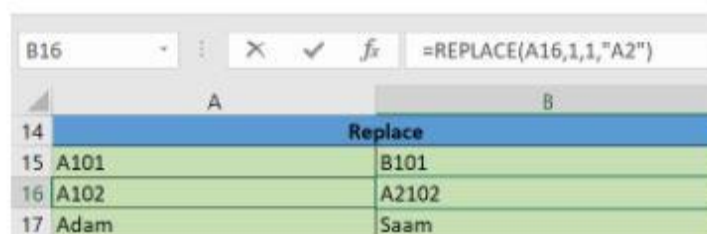


Fig: Replace function in Excel

Finally, we are replacing Adam with Saam by typing: “=REPLACE(A17,1,2, "Sa")”

	A	B
14	Replace	
15	A101	B101
16	A102	A2102
17	Adam	Saam

Fig: Replace function in Excel

Let's now move to our next function.

5.3.12 SUBSTITUTE

The SUBSTITUTE() function replaces the existing text with a new text in a text string. The syntax is “=SUBSTITUTE(text, old text, new text, [instance num])”.

Here, [instance num] refers to the index position of the present texts more than once. Given below are a few examples of this function: Here, we are substituting “I like” with “He likes” by typing: “=SUBSTITUTE(A20, "I like", "He likes")”

	A	B
19	Substitute	
20	I like Excel	He likes Excel
21	MS Excel 2010, MS Word 2010	MS Excel 2010, MS Word 2016
22	MS Excel 2010, MS Word 2010	MS Excel 2016, MS Word 2016

Fig: Substitute function in Excel

Next, we are substituting the second 2010 that occurs in the original text in cell A21 with 2016 by typing “=SUBSTITUTE(A21,2010, 2016,2)”.

Substitute function i

The screenshot shows the Excel formula bar with the formula `=SUBSTITUTE(A21,2010, 2016,2)`. Below the formula bar is a table with columns A and B. Row 19 is a header row with the title "Substitute". Row 20 shows "I like Excel" in column A and "He likes Excel" in column B. Row 21 shows "MS Excel 2010, MS Word 2010" in column A and "MS Excel 2010, MS Word 2016" in column B. Row 22 shows "MS Excel 2010, MS Word 2010" in column A and "MS Excel 2016, MS Word 2016" in column B.

	A	B
19	Substitute	
20	I like Excel	He likes Excel
21	MS Excel 2010, MS Word 2010	MS Excel 2010, MS Word 2016
22	MS Excel 2010, MS Word 2010	MS Excel 2016, MS Word 2016

Fig: n Excel

Now, we are replacing both the 2010s in the original text with 2016 by typing “=SUBSTITUTE(A22,2010,2016)”.

The screenshot shows the Excel formula bar with the formula `=SUBSTITUTE(A22,2010,2016)`. Below the formula bar is a table with columns A and B. Row 19 is a header row with the title "Substitute". Row 20 shows "I like Excel" in column A and "He likes Excel" in column B. Row 21 shows "MS Excel 2010, MS Word 2010" in column A and "MS Excel 2010, MS Word 2016" in column B. Row 22 shows "MS Excel 2010, MS Word 2010" in column A and "MS Excel 2016, MS Word 2016" in column B.

	A	B
19	Substitute	
20	I like Excel	He likes Excel
21	MS Excel 2010, MS Word 2010	MS Excel 2010, MS Word 2016
22	MS Excel 2010, MS Word 2010	MS Excel 2016, MS Word 2016

Fig: Substitute function in Excel

That was all about the substitute function; let’s now move on to our next function.

5.3.13 LEFT, RIGHT, MID

The LEFT() function gives the number of characters from the start of a text string. Meanwhile, the MID() function returns the characters from the middle of a text string, given a starting position and length. Finally, the right() function returns the number of characters from the end of a text string. Let’s understand these functions with a few examples. In example, we use the function left to obtain the leftmost word on the sentence in cell A5.

B5				
=LEFT(A5,5)				
	A	B	C	D
4				
5	Excel is fun to learn	Excel	is fun	to learn
6				

Fig: Left function in Excel

Shown below is an example using the mid function.

C5				
=MID(A5,7,6)				
	A	B	C	D
4				
5	Excel is fun to learn	Excel	is fun	to learn
6				

Fig: Mid function in Excel

Here, we have an example of the right function.

D5				
=RIGHT(A5,8)				
	A	B	C	D
4				
5	Excel is fun to learn	Excel	is fun	to learn
6				

Fig: Right function in Excel

5.3.14 UPPER, LOWER, PROPER

The UPPER() function converts any text string to uppercase. In contrast, the LOWER() function converts any text string to lowercase. The PROPER() function converts any text string to proper case, i.e., the first letter in each word will be in uppercase, and all the other will be in lowercase. Let's understand this better with the following examples: Here, we have converted the text in A6 to a full uppercase one in A7.

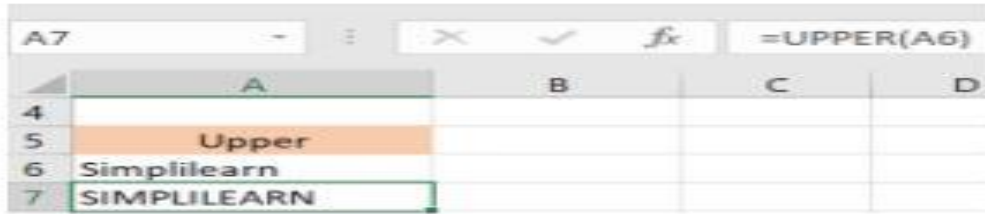


Fig: Upper function in Excel

Now, we have converted the text in A6 to a full lowercase one, as seen in A7.

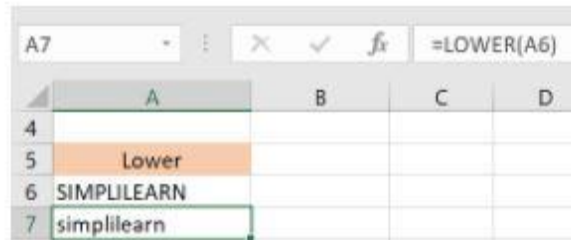


Fig: Lower function in Excel

Finally, we have converted the improper text in A6 to a clean and proper format in A7.

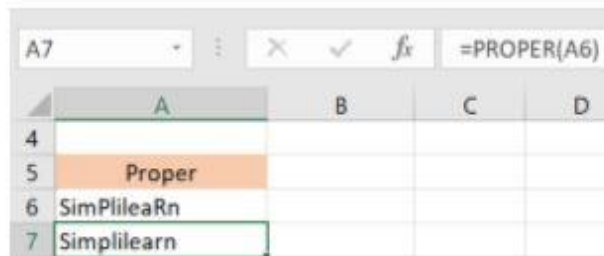


Fig: Proper function in Excel

Now, let us hop on to exploring some date and time functions in Excel.

5.3.15 NOW()

The NOW() function in Excel gives the current system date and time.



Fig: Now function in Excel

The result of the NOW() function will change based on your system date and time.

5.3.16 TODAY()

The TODAY() function in Excel provides the current system date.



Fig: Today function in Excel

The function DAY() is used to return the day of the month. It will be a number between 1 to 31. 1 is the first day of the month, 31 is the last day of the month.



Fig: Day function in Excel

The MONTH() function returns the month, a number from 1 to 12, where 1 is January and 12 is December.



Fig: Month function in Excel

The YEAR() function, as the name suggests, returns the year from a date value.



Fig: Year function in Excel

5.3.17 TIME()

The TIME() function converts hours, minutes, seconds given as numbers to an Excel serial number, formatted with a time format.



Fig: Time function in Excel

5.3.18 HOUR, MINUTE, SECOND

The HOUR() function generates the hour from a time value as a number from 0 to 23. Here, 0 means 12 AM and 23 is 11 PM.



Fig: Hour function in Excel

The function MINUTE(), returns the minute from a time value as a number from 0 to 59.



Fig: Minute function in Excel

The SECOND() function returns the second from a time value as a number from 0 to 59.



Fig: Second function in Excel

5.3.19 DATEDIF

The DATEDIF() function provides the difference between two dates in terms of years, months, or days. Below is an example of a DATEDIF function where we calculate the current age of a person based on two given dates, the date of birth and today's date.



Fig: Datedif function in Excel

Now, let's skin through a few critical advanced functions in Excel that are popularly used to analyze data and create reports.

5.3.20 VLOOKUP

Next up in this article is the VLOOKUP() function. This stands for the vertical lookup that is responsible for looking for a particular value in the leftmost column of a table. It then returns a value in the same row from a column you specify. Below are the arguments for the VLOOKUP function:

Lookup value - This is the value that you have to look for in the first column of a table.

Table - This indicates the table from which the value is retrieved.

Col index - The column in the table from the value is to be retrieved.

Range lookup - [optional] TRUE = approximate match (default). FALSE = exact match.

We will use the below table to learn how the VLOOKUP function works.

If you wanted to find the department to which Stuart belongs, you could use the VLOOKUP function as shown below:

	A	B	C	D	E
1	First Name	Last Name	Department	City	Date Hired
2	Ben	Zampa	HR	Chicago	10-11-2001
3	Stuart	Carry	Marketing	Kansas	20-06-2002
4	Jenson	Button	Operations	New York	01-12-2004
5	Lucy	Davis	Sales	Los Angeles	25-02-2011
6	Trent	Patinson	IT	Boston	17-08-2015
7	Jhonny	Evans	Sales	Houston	10-01-2018

Fig: Vlookup function in Excel

Here, A11 cell has the lookup value, A2: E7 is the table array, 3 is the column index number with information about departments, and 0 is the range lookup.

9	Vlookup				
10	First Name	Last Name	Department	City	Date Hired
11	Stuart		=VLOOKUP(A11,A2:E7,3,0)		

If you hit enter, it will return “Marketing”, indicating that Stuart is from the marketing department.

9	Vlookup				
10	First Name	Last Name	Department	City	Date Hired
11	Stuart		Marketing		

5.3.21 HLOOKUP

Similar to VLOOKUP, we have another function called HLOOKUP() or horizontal lookup. The function HLOOKUP looks for a value in the top row of a table or array of benefits. It gives the value in the same column from a row you specify.

Below are the arguments for the HLOOKUP function:

lookup_value - This indicates the value to lookup.

table - This is the table from which you have to retrieve data.

row_index - This is the row number from which to retrieve data.

range_lookup - [optional] This is a boolean to indicate an exact match or approximate match.

The default value is TRUE, meaning an approximate match.

Given the below table, let's see how you can find the city of Jenson using HLOOKUP.

G	H	I	J	K	L	M
First Name	Ben	Stuart	Jenson	Lucy	Trent	Jhonny
Last Name	Zampa	Carry	Button	Davis	Patinson	Evans
Department	HR	Marketing	Operations	Sales	IT	Sales
City	Chicago	Kansas	New York	Los Angeles	Boston	Houston
Date Hired	10-11-2001	20-06-2002	01-12-2004	25-02-2011	17-08-2015	10-01-2018

Hlookup	
First Name	Jenson
City	=HLOOKUP(H23,G1:M5,4,0)

Fig: Hlookup function in Excel

Here, H23 has the lookup value, i.e., Jenson, G1:M5 is the table array, 4 is the row index number, 0 is for an approximate match. Once you hit enter, it will return "New York".

Hlookup	
First Name	Jenson
City	New York

Our Data Analyst Master's Program will help you learn analytics tools and techniques to become a Data Analyst expert! It's the perfect course for you to jumpstart your career. Enroll now!

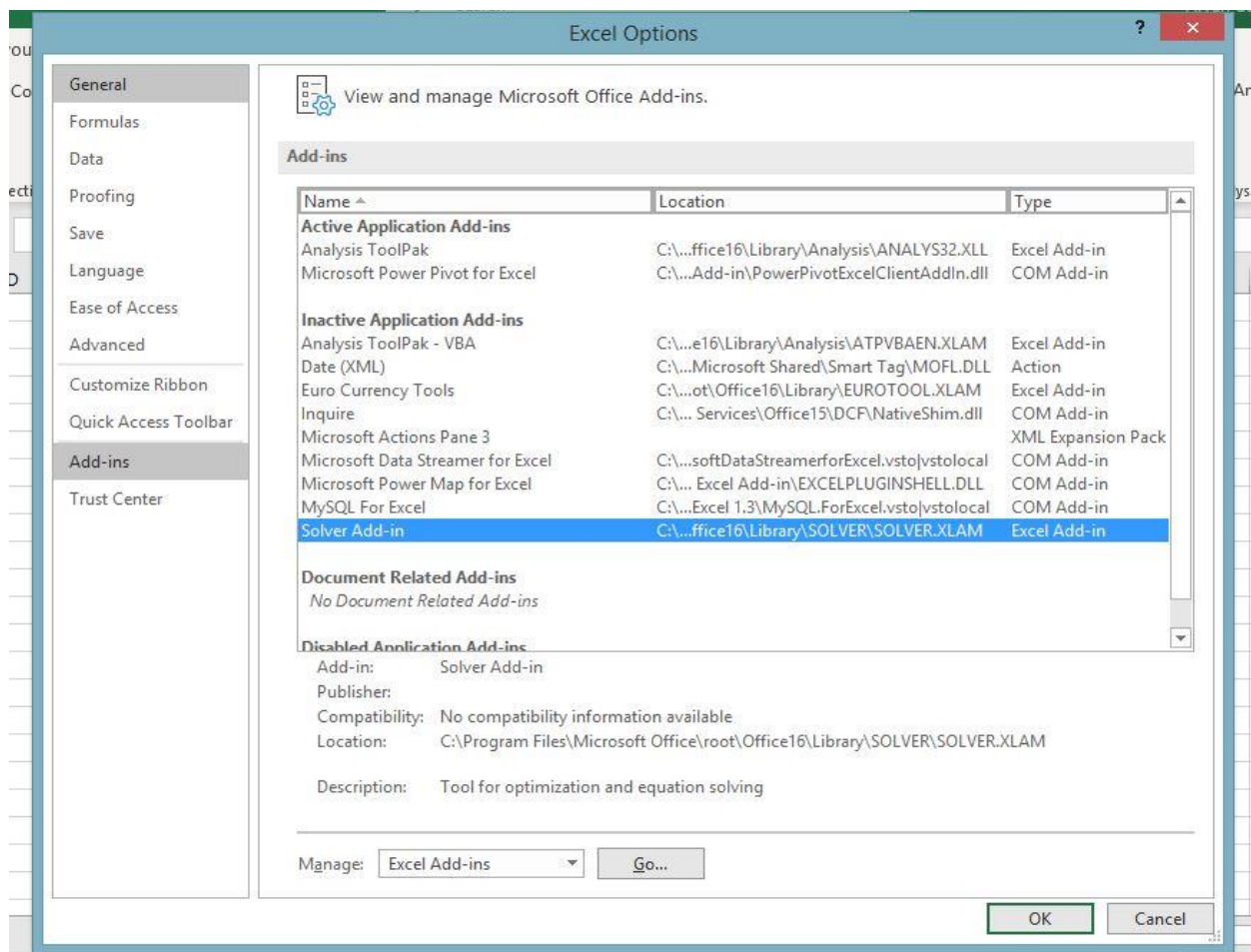
Loan Amount	\$400,000.00
Terms In Month	120
Rate of Interest	9%
Payment	(\$5,000.00)

5.4 Statistical Analysis with solver by MS Excel

What-If Analysis is the method of changing the values to try out different scenarios for formulas in Advanced excel. Several different sets of values can be used in one or multiple of these

Advanced excel formulas to explore the different results. A solver is ideal for what-if analysis. It is an add-in program in Microsoft Excel and is helpful on many levels. The feature can be used to identify an optimal value for a formula in the cell known as the objective cell. Some constraints or limits are however applicable on other formula cell values on a worksheet. Solver works with decision variables which are a group of cells used in computing the formulas in the objective and constraint cells. The solver adjusts the value of decision variable cells to work on the limits on constraint cells. This process aids in determining the desired result for the objective cell. Activating Solver Add-in

- On the File tab, click Options.
- Go to Add-ins, select Solver Add-in, and click on the Go button.



- Check Solver Add-in and click OK.

Many people intend to use statistical techniques in their research. It is definitely a good practice to substantiate your claims with the help of data. Statistics has various functions, which can be broadly categorized as follows:

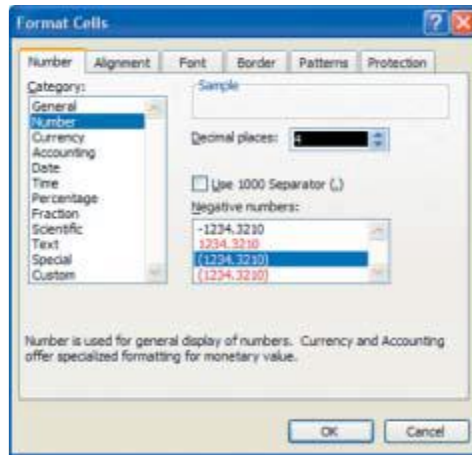
- 1) Summaries and Describe data: One summarizes and describes the data in order to view data at a glance. If it is nominal or ordinal data, one makes cross-tabulations and graphs; if it is interval or ratio data then z-scores are calculated.
- 2) Variance and distribution of the data: In order to measure the spread of the data and knowing its distributions one makes tables and charts and graphs for nominal/ordinal data and histograms with normal curve or box plots with interquartile range for interval/ratio data.
- 3) Compare groups: When one has to compare two or more populations then one makes cross-tabulations for nominal/ordinal data and employ testing of hypothesis for continuous/numeric data divided into groups.
- 4) Identify relationships: In order to identify relationships in the data, one uses cross- tabulations for nominal/ordinal data; calculate correlation coefficient and scatter plot for Interval/ratio data or go for linear regression/ ANOVA for data with one dependent and 2 or more predictor variables.
- 5) Identify groups of similar cases: Carrying out hierarchical cluster analysis solves the problem of identifying groups of similar cases or k-means cluster analysis. One uses Discriminant analysis for identify characteristics of known groups.
- 6) Identify groups of similar variables: Factor analysis is carried out to identify groups of similar variables.

5.5 FEATURES OF SOME STATISTICAL PACKAGES

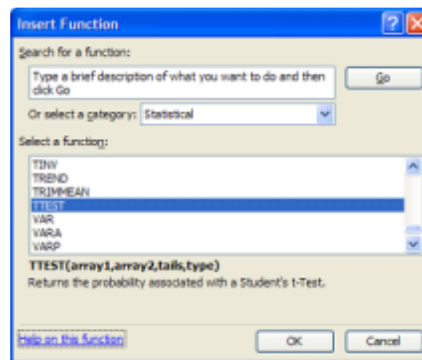
Advances in computing especially the advent of the personal computer (PC) have made computing a game of the commoners. Today one has the computing power as one can easily load software of his choice or need into his PC. There is a plethora of read-made computer packages available today. No one can find different statistical packages for applications to different disciplines. We will describe two such packages that are readily available and are popular and user friendly. We will also give a glimpse of some other packages in the subsequent section Microsoft Excel, is a big worksheet (it can take data rows in thousands across 256 columns).

This worksheet can be used for data entry and for performing calculations by click of buttons. It has a “paste function where you can paste any formula from a big list of inbuilt functions. MS Excel can be used to create tables, and graphs and perform statistical calculations. The work done in MS Excel can be easily copied and pasted to many window-based programs for further analysis.

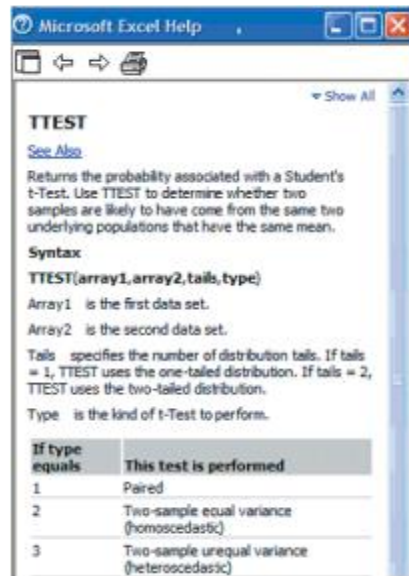
According to Pottel, “Spreadsheets are a useful and popular tool for processing and presenting data. In fact, Microsoft Excel spreadsheets have become somewhat of a standard for data storage, at least for smaller data sets. The fact that the program is often being packaged with new computers, which increases its easy availability, naturally encourages its use for statistical analysis. However, many statisticians find this unfortunate, since Excel is clearly not a statistical package. There is no doubt about that, and Excel has never claimed to be one. But one should face the facts that due to its easy availability many people, including professional statisticians, use Excel, even on a daily basis, for quick and easy statistical calculations. Therefore, it is important to know the flaws in Excel, which, unfortunately, still exist today! “Excel is clearly not an adequate statistics package because many statistical methods are simply not available. This lack of functionality makes it difficult to use it for more than computing summary statistics and simple linear regression and hypothesis testing”. However in MS Excel 2003 aspects of the some statistical functions, including rounding results, and precision have been enhanced. The MS Excel worksheet is a collection of cells. As we have earlier said, there are 65,000 (rows) X 256 (columns) cells in an MS Excel worksheet. Each row or column can be used to enter data belonging to one category. Data entry in MS Excel is as simple as writing on a piece of paper. MS Excel assigns each column a field depending upon the type of data. It supports various data formats; one can choose a data format by formatting the cells.



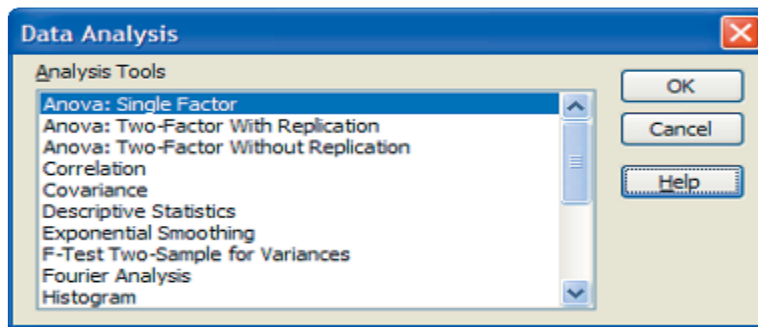
Once the type of cells is defined it is easy to enter the data without taking care of the format. MS Excel can perform usual calculations on the data so entered. It has an insert function (fx) icon that contains many inbuilt functions like sum, count, max/min, standard deviation etc. In fact it has a plethora of built-in functions that performs special calculations without even typing the formula. To perform a calculation, one has to select a function and specify the range of values on which it has to be applied. These functions are known as paste functions.



We will concentrate on the statistical functions and see some of the major statistical Packages functions of MS Excel. As you can see in figure 9.2, once you go to the function menu and choose “statistical” category, you will be asked to select a function. Suppose you have chosen t-test. You will be told on the same screen that t-test returns the probability associated with a student’s t-test. Now if you are still not comfortable with the description, you may select help on this function, which is at the bottom left of the screen. More help is offered in the following form.



MS Excel has a built-in statistical package for taking you in further details of data analysis. It provides a set of data analysis tools called the Analysis Tool Pack, which you can use to save steps when you develop complex statistical analyses. You provide the data and parameters for each analysis; the tool uses the appropriate statistical macro functions and then displays the results in an output table. Some tools generate charts in addition to output tables. To access these tools, click Data Analysis on the Tools menu



Let us have a brief description of these tools. The table given below highlights their functions and uses. Data Analysis Tools in MS-Excel

SNO	Tool	Function	Use
1	ANOVA	The ANOVA tools provide Different types of variance analysis.	Test of the hypothesis that each sample is drawn from the same underlying probability distribution
2	Correlation Calculate	Calculate the correlation	Examine each pair of

		Coefficient between two variables when measurements on each variable are observed for each of N subjects	measurement variables to determine whether the two measurement variables tend to move together
3	Descriptive Statistics	Generates a report of univariate statistics for data, Providing information about the central tendency and variability of your data.	Describes the data in an interpretable format and show summary statistics like mean, mode, median, std. deviation, skewness, kurtosis and range etc.
4	Exponential Smoothing	Predicts a value based on the forecast for the prior period, adjusted for the error in that prior forecast	Forecast on the basis of a smoothing constant.
5	Regression analysis	Performs linear regression analysis by using the “least squares” method to fit a line through a set of observations	Analyze how a single dependent variable is affected by the values of one or more independent variables
6	Sampling analysis	Creates a sample from a population by treating the input range as a population	Infer about a population on the basis of a sample

You have seen that MS Excel can do virtually most of common statistical calculations.

There are two more features that are worth mentioning when one talks about the statistical functions of MS Excel. These two are cross tabulations, pivot tables and the graphical features. MS Excel can be used to create cross tabulations or two-way frequency tables across categorical variables. In MS Excel there is a pivot table wizard which helps in creating tables in multi-dimensions. Let us explain these concepts with the help of an example. The data given below is the percentage contribution of a country to world research in a particular subject area.

5.6 Describe the Features of Statistical Package MS-Excel

One of the features of the development of modern world is the development of the capacity to convert observations in numbers. The science, which deals with numbers, is statistics.

It crunches the numbers and organizes them in a meaningful way so that information is generated. This information builds up knowledge and thus the development goes on. Advances in computing have come handy in this as they help in doing this part of job accurately, timely effectively and convincingly.

Computer can help immensely in the statistical analysis. There exist numerous statistical tools and the need is to identify their actual usage. Even with the use of a statistical package many statistical procedures require a lot of prior knowledge and insight. In this Unit, we have tried to build-up a case for the usage of statistics by first defining statistics and what it can do to your data. You will further find the definition of the data types. An explanation of the common tasks that are performed in a preliminary analysis is also given. This Unit also presents a description of two popular packages (MS Excel 148 and SPSS) and gives a glimpse of some other statistical packages.

5.7 Summary

In this unit, we learned some great features like type of dashboards, benefits, getting your data ready, different features of excel which help to make our dashboard work like grouping, table, Charts, Slicers, etc. Converting data in to Table made our Pivot table dynamically linked to source data while appending. Slicers and timelines (Excel 2013 feature) are great features to make our dashboard dynamic

- 1) A statistical package is defined as the software used to collect, organise, interpret and present numerical information. The need of a statistical package arises due to the complexity of calculations involved therein for analysis and inference. It helps to bring accuracy in results.
- 2) The data analysis tools in MS-Excel are ANOVA, correlation covariance, Descriptive Statistics, Exponential Smoothing, F-Test, Moving Average, Regression Analysis, sampling Analysis, t-Test and z-Test.
- 3) The statistical analysis tools in SPSS are Report, Descriptive Statistics, Compare Means, General Linear Model, Correlate, Regression, Classify, Data Reduction, Scale, Non-Parametric Tests and Multiple Response.

5.8 Self-Assessment Questions

1. Create a custom number format to display the cell values in thousands
2. Create a worksheet in which column A has the label Roll-no, column B has the label Names, Column C has the label Marks and Column D has the label percentage
3. Create a range of 15 values and name it as Test. Find out the maximum value from the text range, by using the range name in the formula.
4. Create a worksheet with data in four pages and send only the first two pages for printing

5.9 References

- www.support.office.com
- www.tutorialspoint.com
- www.chandoo.org
- www.ignou.ac.in
- www.contextures.com



U.P. Rajarshi Tandon Open
University, Prayagraj

DCESTAT – 106

Basic Knowledge of Statistical Softwares

Block: 2 Statistical Computation with R

Unit – 6 : Basics of R

Unit – 7 : Statistical Analysis with R

Unit – 8 : Testing of Hypothesis with R

Course Design Committee

Dr. Ashutosh Gupta

Chairman

Director, School of Sciences
U. P. Rajarshi Tandon Open University, Prayagraj

Prof. Anup Chaturvedi

Member

Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. S. Lalitha

Member

Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. Himanshu Pandey

Member

Department of Statistics
D. D. U. Gorakhpur University, Gorakhpur.

Prof. Shruti

Member-Secretary

Professor, School of Sciences
U.P. Rajarshi Tandon Open University, Prayagraj

Course Preparation Committee

Dr. Anuj Kumar Singh

Writer

School of Sciences,
U. P. Rajarshi Tandon Open University, Prayagraj

Dr. P. S. Pundir

Editor

Department of Statistics
University of Allahabad, Prayagraj

Prof. Shruti

Course Coordinator

School of Sciences,
U. P. Rajarshi Tandon Open University, Prayagraj

DCESTAT – 106/ DCESTAT – 106 BASIC KNOWLEDGE OF STATISTICAL SOFTWARES

©UPRTOU

First Edition: July 2023

ISBN :

©All Rights are reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from the Uttar Pradesh Rajarshi Tandon Open University, Prayagraj. Printed and Published by Col.. Vinay Kumar, Registrar, Uttar Pradesh Rajarshi Tandon Open University, 2023.

Printed By:

Block & Units Introduction

The **Block - 2 – Statistical Computations with R**, is the second block with three units.

- In **Unit – 6 – Basics of R**, is being introduced the Terminology and basic Principles of R software, R Studio and R-Commander, creation of data files. Import Export of Data files, Transformation of Data
- In **Unit – 7 – Statistical Analysis with R** is discussed Statistical Analysis using R – Descriptive Statistics, Curve fitting, correlation and regression analysis, graphs
- In **Unit – 8 -Testing of Hypothesis with R** has been introduced Studying of Statistical Analysis using R. Studying of general procedure of testing a hypothesis.

At the end of unit, the summary, self-assessment questions are given.

UNIT:6 BASICS OF R

Structure

- 6.1 Introduction
- 6.2 Objectives
- 6.3 Basics of R Software
- 6.4 R Studio and R-Commander
- 6.5 Creation of Data Files
- 6.6 Command Line, Data Editor and R Studio
- 6.7 Basic and R as a Calculator
- 6.8 Import & Export of Data Files
 - 6.8.1 Import of Excel Data File
 - 6.8.2 Import of Minitab Data File
 - 6.8.3 Import of SPSS Data File
 - 6.8.5 Import of Table Data File
 - 6.8.5 Import of CSV Data File
 - 6.8.6 Working Directory
- 6.9 Export of Data Files (Importing Data Files of Other Software and Redirecting Output)
- 6.10 Transformation of Data
 - 6.10.1 Using Arrange () Method
 - 6.10.2 Numeric
 - 6.10.3 Integer
 - 6.10.4 Complex
 - 6.10.5 Logical
 - 6.10.6 Character
 - 6.10.7 Factor
 - 6.10.8 Vector
 - 6.10.9 Combining Vectors
 - 6.10.10 Value Coercion

6.10.11	Vector Arithmetic's
6.10.12	Logical Index Vector
6.10.13	Matrix
6.10.14	Matrix Elements
6.10.15	Data Frame
6.11	Application of Transformation of Data
6.11.1	Matrix Construction
6.11.2	Transpose
6.11.3	Combining Matrices
6.11.4	Matrix Arithmetic
6.11.5	Addition and Subtraction
6.11.6	Matrix Multiplication
6.12	Summary
6.13	Self-Assessment Questions
6.14	References

6.1 Introduction

How we analyze data has changed dramatically in recent years. With the advent of personal computers and the internet, the sheer volume of data we have available has grown enormously. The science of data analysis (statistics, psychometrics, econometrics, and machine learning) has kept pace with this explosion of data. Before personal computers and the internet, new statistical methods were developed by academic researchers who published their results as theoretical papers in professional journals. Today new methodologies appear daily. The advent of personal computers had another effect on the way we analyze data. When data analysis was carried out on mainframe computers, computer time was precious and difficult to come by. Analysts would carefully set up a computer run with all the parameters and options thought to be needed. Today's data analysts need to access data from a wide range of sources (database management systems, text files, statistical packages, and spreadsheets), merge the pieces of data together, clean and annotate them, analyze them with the latest methods, present the findings in meaningful and graphically appealing ways, and incorporate the results into attractive reports

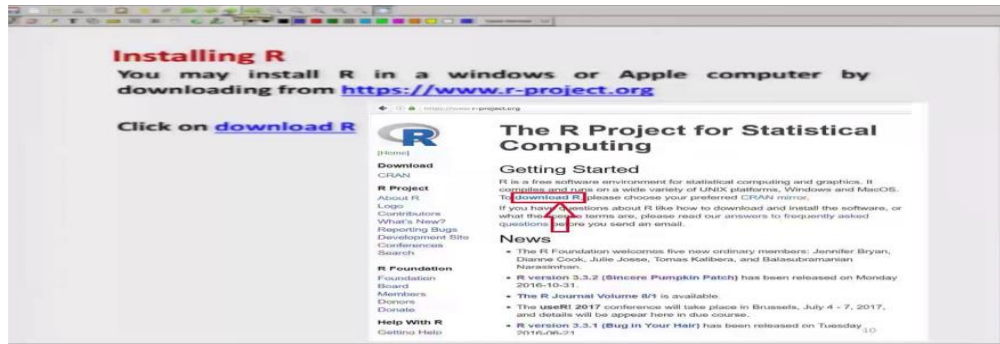
that can be distributed to stakeholders and the public. As we will see in the following pages, R is a comprehensive software package that's ideally suited to accomplish these goals. R is essentially an environment for the data Manipulation, statistical computing as well as graphical display and data analysis, right. R is just like any other software. There are different types of software which helps us in mathematical calculation and statistical data analysis. R is software, and R has an advantage that R can do data manipulation; R can do statistical computing as well as simulations, R is capable of graphical display and R can also help us in doing different type of data analysis. R has an effective way of data handling. So, it can handle the data easily and it can store the input and output variables. This can store the outcome in the form of a scalar as well as form of a vectors or a matrix, and in R software, simple calculations are possible as well as complicated calculations are also possible and it is not difficult, the mathematical calculation like addition, subtraction and these vectors and matrices, everything is possible, just like any other software.

6.2 Objectives

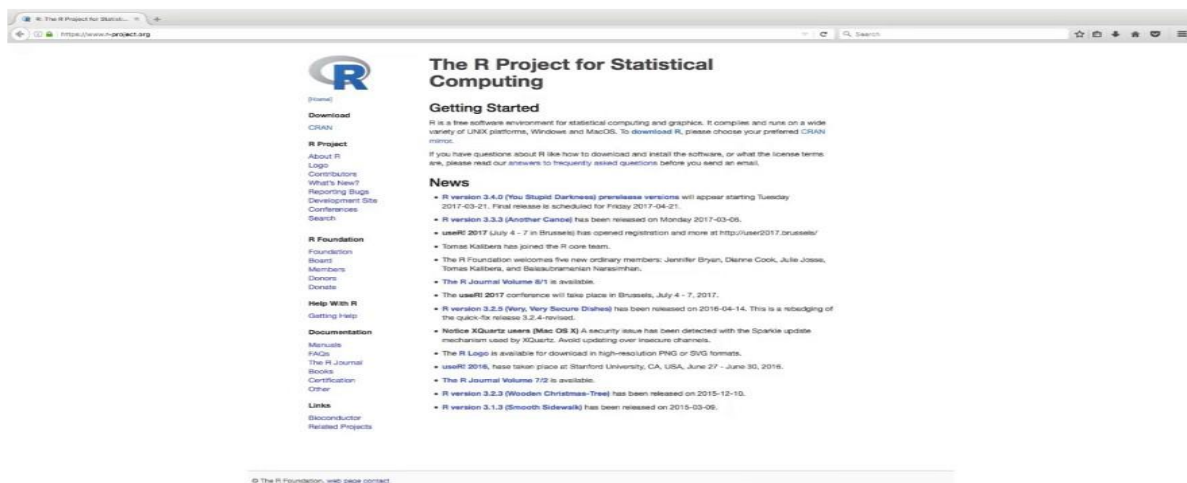
After studying this unit, you will be able to understand the following objectives:

- Studying of Basics of R.
- Studying of R Studio and R-Commander.
- Study of Creation of data files.
- Study Command line, Data Editor and R Studio
- Study of Import Export of Data files.
- Study of Transformation of Data.

6.3 Basics of R Software



How should we get R and how to install it on the computer? Let us try to understand this thing. This is a very simple thing means if you have a little bit idea about using the internet, anybody can do it. So, in order to install the R, what we try to do that there is a website. This website is www R hyphen project dot org and what we try to do here that we try to execute this command. And then I will try to show you what are we going to get and the same outcome I am trying to write down here. If you try to see, I have simply copied and pasted but in order to make you more confident; what I will try to do here that that I will try to show you this online. www R project dot org, if we try to see, we have got this website.



So, you can see here. Now what I have done just for the sake of convenience and in order to illustrate it, I have taken a screenshot of this webpage and I have copied it here. So, I am showed that this will not create any confusion for you, but it will help us in understanding the

things. So, once you come to this homepage, then what you have to do here that you need to go here, there is a command here download R.

Example, you can see here, there is a download R here and here you double click it and once you double click it, it will give you this home page. That is the same thing which I am trying to show you here also. We can click over here at the download icon and then in the next site, you will get here this type of site what you have obtained.



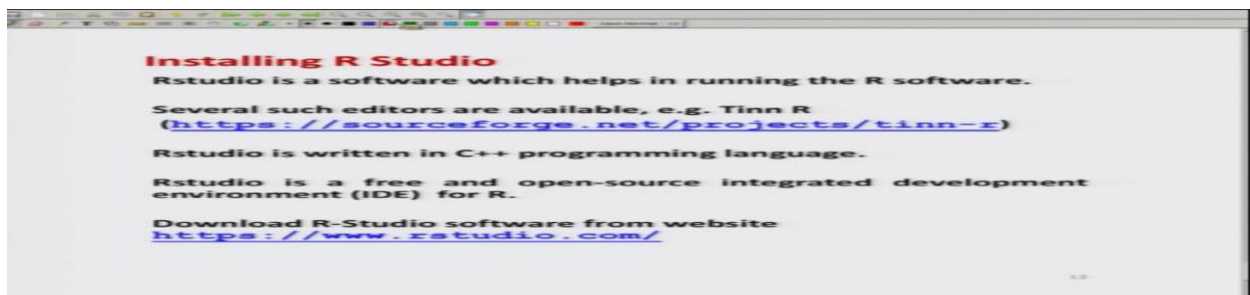
On the left-hand side, you can see here that there are different types of addresses which are given. So, actually different people in different countries, they have uploaded the software. So, you can click on any of the link and it will open the page for downloading the R software. So, you can just click over here or say here, whatever you want, and then once you do it, then you will get the software over here. For example, I can show you here that once you try to do it here.

6.4 R Studio and R-Commander

R Studio software as we had discussed earlier, that is free software that can be downloaded from the website and this is actually a sort of interface between R software and us. Whatever information is contained in R, whatever execution is being done in R, they can be seen through R Studio also. And usually, it is more helpful to work in R Studio rather than working directly into the R software. And particularly if you are beginning to learn the R software, this is more helpful because you can see each and every thing just before your eyes in a single shot. And whenever you are trying to write down the program, you are trying to write down the code of a program, and then it is easier actually, at every step. For example, you can highlight you can

run and you can check whether your commands are working fine or not. In case if you find any mistake, at the same step you can correct it.

So, as we have seen earlier that whenever we start the R Studio, we have 4 windows. Now, our objective is that we want to learn what are these four windows indicating, what type of information is being contained and provided by these four windows shows, so you have seen here we have four windows. So, I am calling this as window 1, this as window 2, this was here say window 3 and this here as say window 4. And now, let us try to understand the information provided by each of the windows one by one first let us try to come to the window 1. This is a place where we try to write down the script or in simple words, this is the place where we type our all the commands. They can be a single line command or they can be multi level commands or that can be entire file containing the one program So, now, you try to look at the minor details of this slide. So, you can see here that this is the place where we try to click to add a new script file and this is a place where we try to click to save the file and this is the place where we try to open an already existing file, this is a place where we click to run the program and if we want to rerun the program, then we need to click over here. These lines can be single level, single line or they can be multi lines which have to be run. So, for example, you can see here I can highlight it in the R Studio software also. For example, you can see here, if you try to click over here, you get here R script and then you can say open here a new R script and then you can see here first script and second script, both are here. This is the place where you can save the details and here is the place where here by clicking here, you can run the program; this is the place here way where you can rerun the program.



Now, after this, we come to another aspect. Whenever we are working with R software, we have two options- either we can work directly with the R software, we can execute the command inside the R software and second option is that I can take help of supplementary

software which works inside the R from outside. So, so there are different types of software which are which are available and, in this course, we are going to use software, what is called as R Studio. So, R Studio is essentially software which helps in the execution of R software and beside R Studio, there are other types of software which are available. For example, one of another software is the Tinn R and this can be downloaded from this site and, but I can use any one actually. I am not trying to say at all that either Tinn R is better than R Studio or vice versa. I have chosen R Studio to work. So, this R Studio is actually written in the C++ Programming language.

R Studio is also free software because in open-source software and this can be freely downloaded from this website. So, what we have to do here that we simply need to copy this link and then we have to open it in the internet browser. So, for example, I can show you here that if I try to say here type here R, R Studio dot com then I get here this thing. So, I have simply taken here a screenshot of this thing. So, what we need to do, this software will be opened and now we have to simply come over here and then I have to click over here at the download part and this software will be downloaded and once this software is downloaded, then you can install it on the computer. So, essentially you will see that once you have installed the R software and R Studio software, you will have a link like this here R and here R Studio and now, we are ready to move into the learning of R software.

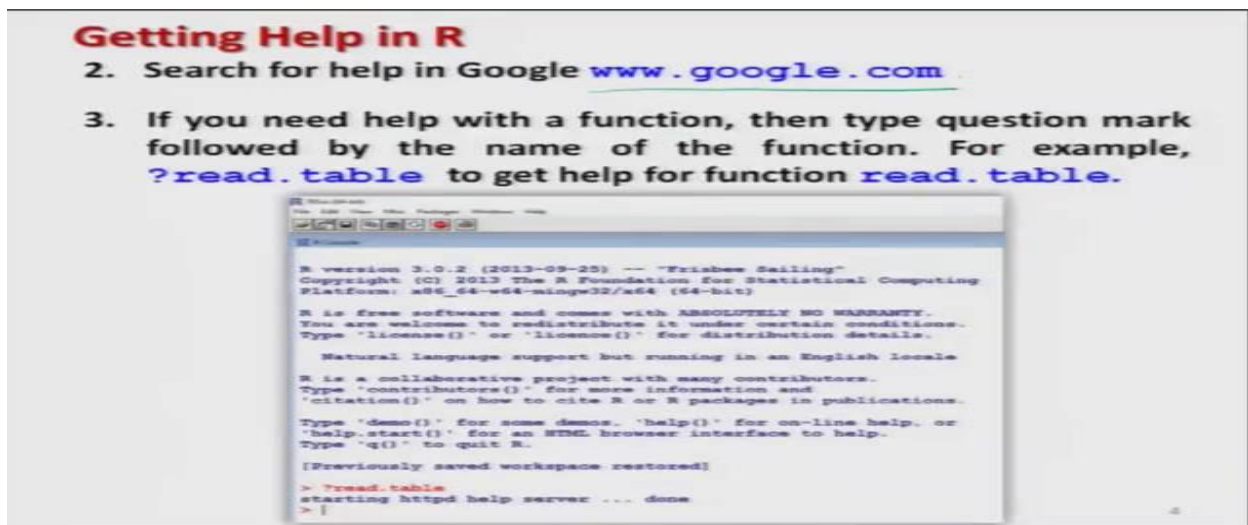
6.5 Creation of Data Files



So now, you can see here on the desktop of your computer that there is going to be a icon of here R like this. So you need to double click on this icon, right. So, let us try to do it here. So now, I have taken a screenshot of the R GUI and this is pasted over here and now let us tries to see what is there in this window. If you try to read it here for the first 2 paragraph, it is trying to

give you the details about the R software and in the last 2 paragraphs; it is trying to give you different options. For example, how to get a citation here and how to obtain the Demonstration of a particular function, how to get here a help and how to quit the R program and so on.

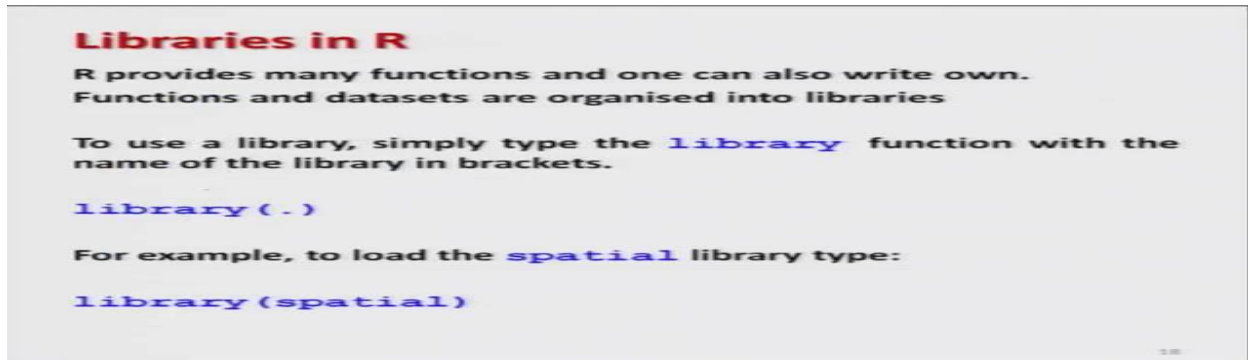
So, if you try to see, if you try to read it, you will get information about this, but now after this, once we have started the program over here, now our objective is that how to obtain the help in the R. So first of all, what we can do, you can see in this window. So, what you can do, you can just double click over here and if you try see it here, we will get here, something like this help.



The next option is that, we can go to Google.com and there are different resources which are available at different sites including the support from the website of the R software. So, we can just look into different types of examples, different types of syntax and they will also try to extend their help to us. If you try to see here that suppose I want to take a help on some function, say called as `read.table`. Well, I am using here some new names like as function or `read.table`. This things should not to bother you because later on, you are going to learn all this things, but my objective here is simply to show you that suppose you know that there is command `read.table` and so obtain the help on that function.

So, suppose want to obtain some help on the `read.table` function, what I have to do here, I simply have to write say question mark and after this I have to type here the function name. For example, here I have done here question mark followed by `read.table` and we try to type this

read.table on the command line. For example, I will simply try to copy and paste over here so that we save some time and I tried to paste it over here.



Libraries in R
R provides many functions and one can also write own.
Functions and datasets are organised into libraries

To use a library, simply type the `library` function with the name of the library in brackets.

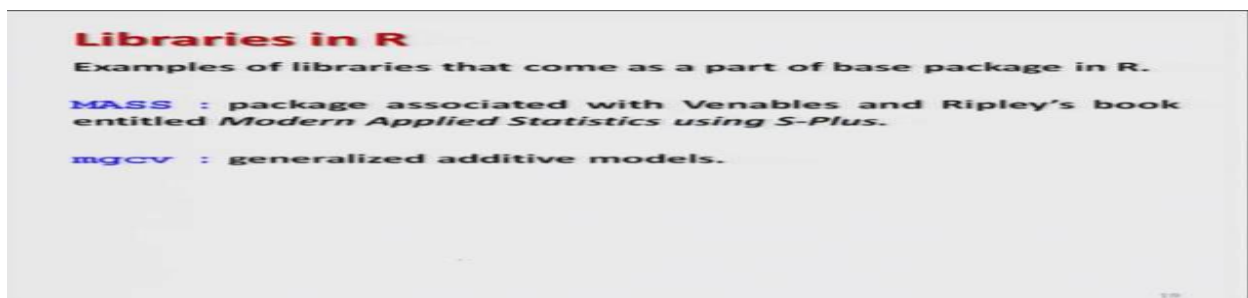
```
library(.)
```

For example, to load the `spatial` library type:

```
library(spatial)
```

18

The next aspect on which I would like to concentrate is the libraries in R. The first question comes, what is a library? So, if you try to see, in simple words, the literal meaning of the library is that a place where there is a collection of many books. So similarly, these libraries are also the collection of several types of commands to execute different types of tasks and R provides many many functions and out of which there are two types of functions one function which are built in inside the R software and say another type of software which you can create. Beside those things, when you try to install the R software, you will see that there are several options given. One option is when you are starting the R for the first time, and then you are actually downloading the base software. So in the base software they have given most of the say these common commands which are useful for a common user and beside that if you want to use any specialized function, then they have given the commands for that specialized task inside a library. For example, in case if I want to fit the times series model, so, there will be a library for the time series. Suppose if I want to fit a fit a special data model, so there will be another library that is dealing with the spatial data.



Libraries in R
Examples of libraries that come as a part of base package in R.

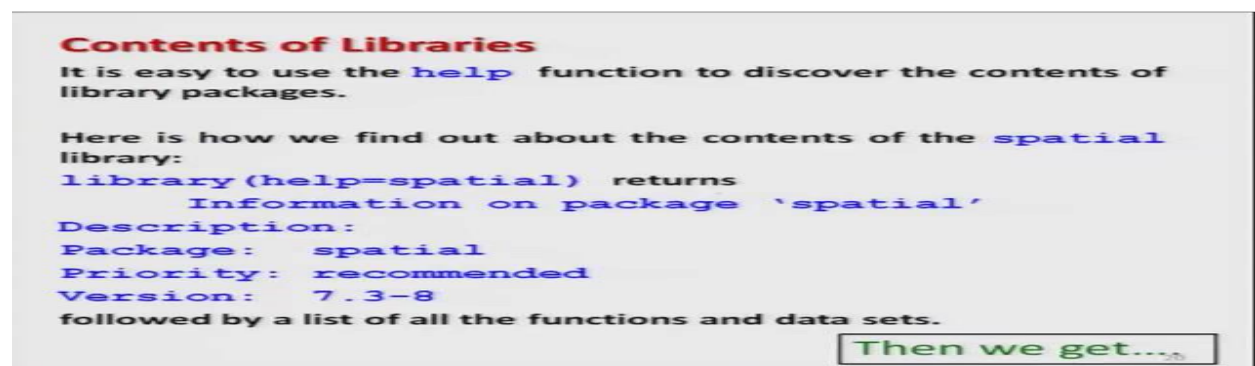
MASS : package associated with Venables and Ripley's book entitled *Modern Applied Statistics using S-Plus*.

mgcv : generalized additive models.

19

Some built in libraries which comes as a part of the base package in R, they are something like here, one is MASS and say another is mgcv and please remember one thing, this

is capital M capital A and capital S S. This MASS package actually, this contains various type of data sets and the tools which are related to a book Modern Applied Statistics using S-plus which was written by Venables and Ripley. Actually, that was the book about the software Splus in which they have use different types of data sets, different types of command and this library contains all those commands over here. Similarly, there is another library m g c v. So, this library contains about the details about the generalized additive model. So if you want to use some generalized additive models, then you have to first load this library and only after that you can use it and in order to load a library, simply write library and inside the brackets the name of the library.



```
Contents of Libraries
It is easy to use the help function to discover the contents of
library packages.

Here is how we find out about the contents of the spatial
library:
library(help=spatial) returns
  Information on package 'spatial'
Description:
Package:  spatial
Priority:  recommended
Version:  7.3-8
followed by a list of all the functions and data sets.
Then we get...
```

So, once you load a library means, obviously, you would like to know what the contents of that library are because as such you have no idea. So, I can use the help function to discover the contents of a library package. For example, earlier, I had just installed the package say here spatial.

Now, we want to know that what is there in this say spatial package and I need to know the help. So, I will try to write down here the library help equal to the package name. So this is help equal to say spatial and once I try to execute it, I will get here this type of screenshot, means the package name is spatial, priority this is good, this is the version and after this, this will give you all the details about this package.

Installing Packages and Libraries

The base R package contains programs for basic operations.

It does not contain some of the libraries necessary for advanced statistical work.

Specific requirements are met by special packages.

They are downloaded and their downloading is very simple.

22

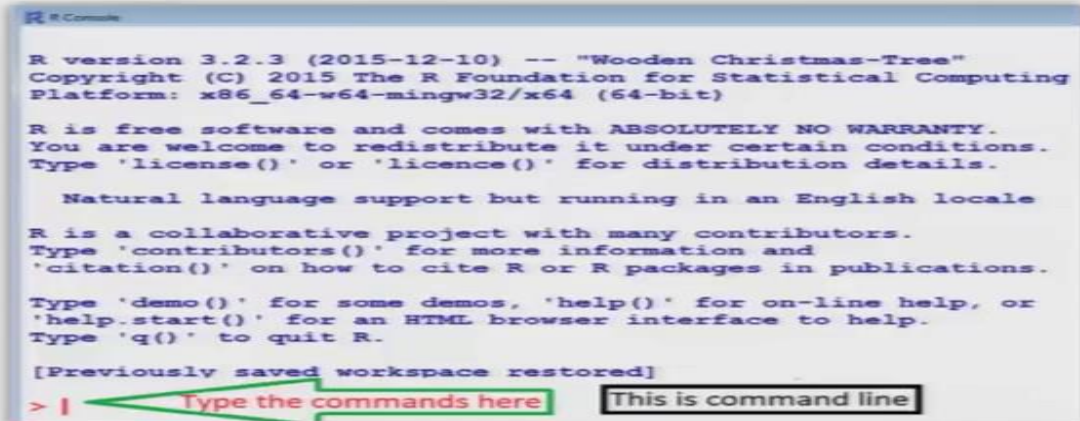
Now, before you try to use this library or say packages, we need to install them. There are some packages which comes as a part of the base package of R and there are certain packages which need to be installed externally and these packages have different types of qualities, they are used for different types of tasks and the R package does not contain all these packages. We have to download it from the website of the R software. So here, we try to learn here how we are going to do. So, first thing is this that we have to do downloading of the software, but believe me, this downloading is very simple. So, now, if you want to install any package, first step is that run the R program and then use the command install. Packages and as soon as you say install. Packages, then the package will be downloaded and it will be installed.

6.6 Command Line, Data Editor and R Studio

Welcome to the lecture on introduction to R software. In this lecture, we will continue with some introductory topics and we will talk about command line, data editors and say software, R Studio. So, let us try to start with one by one.

Command Line versus Scripts

What is command line?



The screenshot shows the R console output. At the bottom, there is a red arrow pointing to the prompt '> |' with the text 'Type the commands here'. To the right of the prompt, there is a black box with the text 'This is command line'.

```
R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

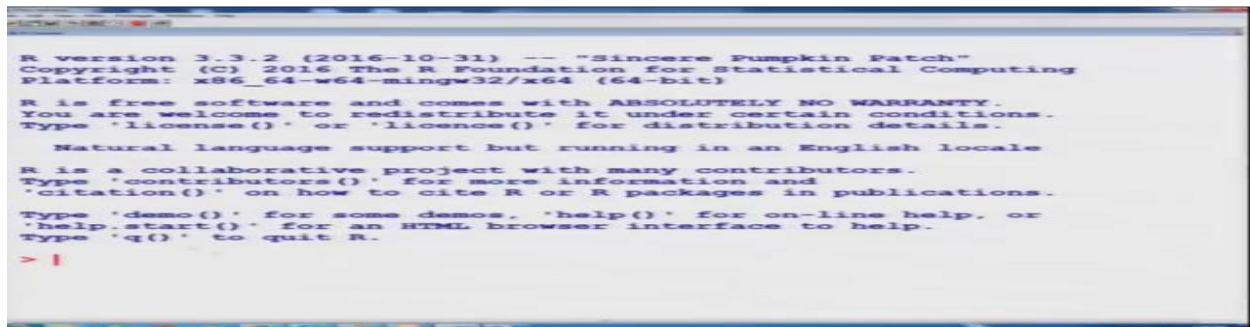
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]
> |
```

The first question comes what is a command line? Why command line? Because in order to write a program in R, there are two options that I can write the program on the command line or I can write it inside a script file.

So, you have seen that when you start the R, you get this type of screen over here. And here you will see that there is a sign something like greater than. This sign, greater than sign is actually the command prompt. For example, if you try to start here R, this is here this thing you can see.



The screenshot shows the R console output. At the bottom, there is a red arrow pointing to the prompt '> |'.

```
R version 3.3.2 (2016-10-31) -- "Sincere Pumpkin Patch"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

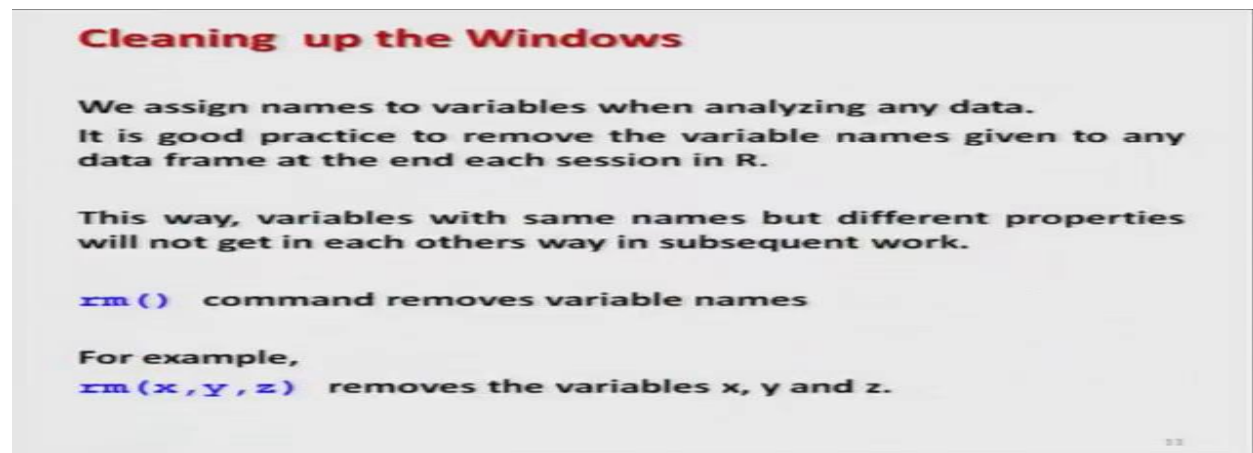
> |
```

So, this is actually command line, and this is the place where we try to write down the syntax or command. For example, if I want to find out the mean, I have to write down here and so on. So, the same thing is being demonstrated here. So, this is here the command line and here we try to type our commands. The R software is not a menu driven software, but here you have to type the commands.

Now, whenever we are trying to write down the commands, there are two options, that the command can be written in a or it can be written in more than one lines. So, single line commands and multi-line commands- both are possible to write in R programming.

Whenever we are writing a single line program; that means, one command is executed on the same line and when we are trying to write down the multi line commands; that means, we will try to write down or type one command at a time in the sequential order. So, whenever we are trying to write such multi line programs, it is always better to write all the commands in a single file and then try to execute the entire file in a single shot.

So that means, whenever we want to write a program, the program is essentially a combination of several syntax, several commands which are written in a logical order depending on the objective of the task. Whatever we want to do it, we have to write down the program. For example, if you simply want to find out the arithmetic mean, in arithmetic mean the first step is to sum all the observations and the second step is to divide the sum by the total number of observations. So, these are two steps. So, if I want to write down a program for finding out the arithmetic mean, first of all I have to write some lines to compute the sum and in the second step, I have to write some lines to divide the sum by the number of observations. And this will complete the entire program for finding of the arithmetic mean.



Cleaning up the Windows

We assign names to variables when analyzing any data. It is good practice to remove the variable names given to any data frame at the end each session in R.

This way, variables with same names but different properties will not get in each others way in subsequent work.

`rm()` command removes variable names

For example,
`rm(x, y, z)` removes the variables x, y and z.

22

Now obviously, once you have done a program, you will get ready for the next program. So, it is always better that you first clean up all the windows, whatever variable names you have defined, whatever information you have defined, that should be cleaned up. For example, suppose I am writing a program and in which I have used a variable name say age and by age I


am trying to denote the ages of some older person. Now, I am trying to use another program in which I want to find the age of some children. So obviously, I will try to define as my natural instinct the variable name to be age and I will try to enter my data, but then there can be some contradiction that when I am trying to run the program, possibly it might be using the information contained in the earlier defined variable age that was containing the ages of some elderly persons.

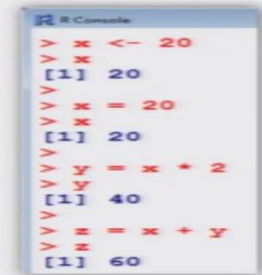
6.7 Basic and R as a Calculator

Earlier, we had discussed about the basic fundamentals mainly related to how to start and how to work with R. From this lecture onwards, in the next couple of lectures, we will be talking about how to do calculations in R, and again I will say I will be concentrating on the basic fundamentals and my objective is that I should help you so that you can learn the course yourself. So, here again, what I am going to do that I have taken some simple calculations, I will try to show you it online and my request is that you also try to do the same thing yourself on your computer, and not only the examples which I am taking, but try to take more example from your area, from your subject and try to solve them, the more you practice, better you we will be. One thing I can also accept that here I am trying to copy and paste the commands from my slides and, but I would request you to at least type those command yourself. The advantage of typing the commands yourself is that, you will remember where to put comma and where to put inverted comma, where to put say this say full stop and where to put colon. These things can come only when you type the command yourself. So, let us now start with our lecture.

So, in this lecture I am going to talk about the basics and how I can use R as a simple calculator. A simple calculator is one where you can do addition, subtraction, multiplication, division and some bracket rules also. So, let us try to start it, but before that let us try to understand the terminology and the symbols and notation used in R.

Basics

- `>` is the prompt sign in R. 
- The assignment operators are the left arrow with dash `<-` and equal sign `=`.
 - `> x <- 20` assigns the value 20 to `x`.
 - `> x = 20` assigns the value 20 to `x`.
 - Initially only `<-` was available in R.
- `> x = 20` assigns the value 20 to `x`.
 - `> y = x * 2` assigns the value $2 \times x$ to `y`.
 - `> z = x + y` assigns the value $x + y$ to `z`.



First thing what you have to keep in mind that as soon as we start our R, there is a prompt sign and the prompt sign in R is denoted by greater than sign. So, this will always be the sign when you start your R and that will be the first line on the R GUI window, that is the R graphic user interface window. Now after this, whenever we want to do anything, I have to assign a value to a variable.

For example, in mathematics, you have seen that usually we write x is equal to 2. So, the question is this that how to do this thing and what is the meaning of this thing. The meaning of writing x is equal to 2 is this that I am trying to consider here a variable, and I am assigning it a value 2.

So, in case if I want to do this thing in R, I have 2 options. This equality sign can be used as just as equality sign and another option is this instead of equality sign, this symbol less than and hyphen (`<-`), this can be used. So, if we try to see here, I am trying to write down here $x <- 20$ or I can also write $x=20$. Now, the question comes out why there are two symbols for the same job. Actually, when R started, that was developed on the lines of S-Plus and in S-Plus, the assignment operator was this one, less than hyphen.

6.8 Import Export of Data Files

Welcome to the next lecture on the course introduction to R software. Now, in this lecture, we are going to understand some concepts related to the import of data. What do we really mean by import of data? You see whenever you want to manipulate any data set, the data

set has to be present on your computer and this data set may come from different sources; first option is that you can create that data set yourself. For example, you open a spreadsheet and try to enter the data and even when you are trying to save the spreadsheet, there can be different forms to save that spreadsheet. For example, command separated values, txt format and says some other thing.

Beside this thing, it is also possible that data is available from some other source; these sources can be somebody already has entered the data in some other computer and you want to import that data on your computer where you want to do the analysis using the R software.

Another option can be that data is uploaded somewhere on some internet site and you want to download it or you want to work on that data set directly. So, these are different possible ways in which the data set can be available to us and in this lecture, we are going to concentrate that how to read this data set into the R software. Once you can read the data set into the R software, then you can do all other manipulations which you have learnt in the earlier lectures and those which we are going to learn in the further lectures.

So, let us try to start our lecture first thing comes that whenever you want to work with a data file, that data file has to be located in some directory on your computer and you have to instruct the R software to read the data from that directory only. There is a default directory in the R software and it is not always possible to put the data in that default directory, but we would like to put the data at a place which is convenient to us. So, for example, if you want to read the path of the directory in the R, then how to get it done; means first of all once you start the R, you would like to see where is the working directory, from where is or say from which location this R software is fetching the data.

So, in order to do that thing, we have say one option here which is called here `getwd`, this means get working directory working directory and then we put an argument. By this command, you can get the working directory in which the R is presently working. Now, suppose I want to change this directory. For that we have another command which is called as `setwd`; that means set working directory and then inside the argument, you have to specify the location of the data set inside the double quotes, location has to be in terms of the path of the computer. For example, in this course, what I have done, I can show you here that I have created a directory on the C

drive, you can see here where I am highlighting here; this I have created on the say here C drive, that you can see over here.

6.8.1 Import of Excel Data File

The sample data is in Excel format, and needs to be imported into R prior to use. For this, we can use the function `read.xls` from the `gdata` package. It reads from an Excel spreadsheet and returns a data frame. The following shows how to load an Excel spreadsheet named "mydata.xls". This method requires Perl runtime to be present in the system.

```
> library(gdata)
> help(read.xls)
> mydata = read.xls("mydata.xls")
```

Alternatively, we can use the function `loadWorkbook` from the `XLConnect` Package to read the entire workbook, and then load the worksheets with `readWorksheet`. The `XLConnect` package requires Java to be pre-installed.

```
> library(XLConnect)
> wk = loadWorkbook("mydata.xls")
> df = readWorksheet(wk,sheet="Sheet1")
```

6.8.2 Import of Minitab Data file

If the data file is in Minitab Portable Worksheet format, it can be opened with the function `read.mtp` from the `foreign` package. It returns a list of components in the Minitab worksheet

```
> library(foreign)
> help(read.mtp)
> mydata = read.mtp("mydata.mtp")
```

6.8.3 Import of SPSS Data File

For the data files in **SPSS** format, it can be opened with the function `read.spss` also from the foreign package. There is a `to.data.frame` option for choosing whether a data frame is to be returned. By default, it returns a list of components.

instead.

```
> library(foreign)
> help(read.spss)
> mydata = read.spss("myfile",to.data.frame=TRUE)
```

6.8.4 Import of Table Data File

A data table can reside in a text file. The cells inside the table are separated by blank characters.

```
> mydata = read.table("mydata.txt")
> mydata
```

For further detail of the function `read.table`, please consult the R documentation.

```
> help(read.table)
```

6.8.5 Import of CSV Data File

The sample data can also be in comma separated values (CSV) format. Each cell inside such data file is separated by a special character, which usually is a comma, although other characters can be used as well. The first row of the data file should contain the column names instead of the actual data. Here is a sample of the expected format. After we copy and paste the data above in a file named "mydata.csv" with a text editor, we can read the data with the function `read.csv`.

```
> mydata = read.csv("mydata.csv")
> mydata
```

In various European locales, as the comma character serves as the decimal point, the function `read.csv2` should be used instead. For further detail of the `read.csv` and `read.csv2` functions, please consult the R documentation.

```
> help(read.csv)
```

6.8.6 Working Directory

Finally, the code samples above assume the data files are located in the R **working directory**, which can be found with the function `getwd`.

```
> getwd() # working directory
```

We can select a different working directory with the function `setwd`, and thus avoid entering the full path of the data files.

```
> setwd("<new path>")
```

Note that the forward slash should be used as the path separator even on Windows platform.

```
> setwd("C:/MyDoc")
```

6.9 Export of Data Files (Importing Data Files of Other Software and Redirecting Output)

We had discussed the aspect how to import data sets from some external sources and we had discussed about different types of files structure like as dot csv dot txt and we discussed how to import them in your R software. Now, we are going to continue our discussion and we are going to learn that how one can import a data set that was created in some other software.

first source which I am going to take is say how to import the data from a spreadsheet and one of the important and popular packages to create a spreadsheet is say Microsoft Excel software and, in that case, the extension of the file is dot xlsx or this can also be dot xls in the earlier versions of Microsoft Excel software. We are going to address is how to import a data which is created in Excel software and has got an extension dot xlsx; actually, dot xls was the extension in the earlier version of the Microsoft Excel software. So, in order to read a file which is created in xlsx package, we have a command `read dot xlsx` and then inside the arguments, we have to write down the file name, but when I try to use this command, this is going to read only the first sheet of the Excel spreadsheet. When we are trying to create an Excel sheet in the Excel software, then it is possible to create different sheets inside the same file.

In order to use the older Excel files in dot xls format, we have to use another package `gdata` and then I have to use the command `read dot xls` and rest everything remains the same, but in case if you try to use this `read xls`, this will again only read the first sheet of the file.

So, in order to read a file of the format dot xls; so, the first step is that we have to install the package called as `gdata`, after the `gdata` package has been installed, we need to upload it. So, I use the command `library gdata` and this can be done exactly in the same way as we did in the earlier case and after this, you want to read the file. So, for that use the command `read dot xls`, give the name of the file and here you have to specify the `sheetIndex` or `sheetName` exactly in the same way as we did in the case of dot xlsx format. We take another source from which the data can come.

Suppose we are going to read a SPSS data file. SPSS is a very popular statistical software. In order to read a data file that is created in the SPSS package, we need to install a special package which is called as `foreign`; all in small letters. So, first we install this foreign package and then we use the function `read dot SPSS` and we give the name inside this argument. So in case if you want to read a data file from SPSS package, first step is to install the package `foreign` using this command, then load this package using `library` command and then simply try to use the `read dot SPSS` command and inside the argument enclosed by the double quotes, try to write down the name of the file.

Similarly, if we want to read a data file from say software SAS; SAS is another a statistical software, Statistical Analysis System and in that case, we use the command `read dot XPORT`;

`X P O R T` and then the same format inside the argument you have to specify the name of the file along with its paths inside the double quotes and similarly, there is another statistical software what is called as a STATA. So, if you want to read a data file that was created in the software STATA, you just use the command `read dot dta` and inside the arguments, try to give the name of the file and its path inside double quotes.

In case if you want to have some more description on the data import and export that can be also found in the R manual which is located at this data set on the website of the R project. So means if you want to have a specific thing, you can always read it from here. So, after learning

that how we can read the data from different sources or in different formats, the next objective is that whenever we are trying to run the program, how that program can be saved?

how one can save the outcome of a program inside a file, but in order to understand it, first we also need to know that how to see the contents of the working directory because whenever you are trying to save the file in a directory, you would also like to check whether that file is there or not or what are their contents So, first of all, I try to explain you here that how we can see the contents of the working directory

6.10 Transformation of Data

The data transformation in R is mostly handled by the external packages tidyverse and dplyr . These packages provide many methods to carry out the data simulations. There are a large number of ways to simulate data transformation in R. These methods are widely available using these packages, which can be downloaded and installed using the following command

```
>install.packages("tidyverse")
```

6.10.1 Using ARRANGE () Method

For data transformation in R, we will use the arrange () method, to create an order for the sequence of the observations given. It takes a single column or a set of columns as the input to the method and creates an order for these. The arrange () method in the tidyverse package inputs a list of column names to rearrange them in a specified order. By default, the arrange() method arranges the data in ascending order. It has the following syntax.

There are several basic R data types that are of frequent occurrence in routine R calculations. Though seemingly innocent, they can still deliver surprises. Instead of chewing through the language specification, we will try to understand them better by direct experimentation with R code.

For simplicity, we defer the concept of vector for later discussion. Here, we use only vectors of unit length for demonstration

6.10.2

NUMERIC

Decimal values are called numeric in R. It is the default computational data type.

If we assign a decimal value to a variable `x` as follows, `x` will be of numeric type.

```
> x = 10.5 # assign a decimal
```

```
> x # print x
```

```
[1] 10.5
```

```
> class(x) # print class name
```

```
[1] "numeric"
```

Furthermore, even if we assign an integer to a variable `k`, it is still being saved as a numeric value.

```
> k = 1
```

```
> k # print k
```

```
[1] 1
```

```
> class(k) # print class name
```

```
[1] "numeric"
```

The fact that `k` is not an integer can be confirmed with the `is.integer` function.

We will discuss how to create an integer in our next tutorial on the integer type.

```
> is.integer(k) # is k an integer?
```

```
[1] FALSE
```

6.10.3

INTEGER

In order to create an integer variable in R, we invoke the `integer` function. We can be assured that `y` is indeed an integer by applying the `is.integer` function.

```
> y = as.integer(3)
```

```
> y # print y
```

```
[1] 3
```

```
> class(y) # print class name
```

```
[1] "integer"
```

```
> is.integer(y) # is y an integer?
```

```
[1] TRUE
```

Incidentally, we can coerce a numeric value into an integer with the `as.integer` function.

```
> as.integer(3.14)           # integer cast
```

```
[1] 3
```

And we can parse a string for decimal values in much the same way.

```
> as.integer("5.27")       # parse string
```

```
[1] 5
```

On the other hand, it is erroneous trying to parse a non-decimal string.

```
> as.integer("Joe")
```

```
[1] NA
```

Warning message:

NAs introduced by coercion

Often, it is useful to perform arithmetic on logical values. Just like the C language,

TRUE has the value 1, while FALSE has value 0.

```
> as.integer(TRUE)
```

```
[1] 1
```

```
> as.integer(FALSE)
```

```
[1] 0
```

6.10.4 COMPLEX

A complex value in R is defined via the pure imaginary value i .

```
> z = 1 + 2i # a complex number
```

```
> z # print z
```

```
[1] 1+2i
```

```
> class(z) # print class name
```

```
[1] "complex"
```

The following gives an error since -1 is not a complex value.

```
> sqrt(-1) # square root of -1
```

```
[1] NaN
```

Warning message:

In `sqrt(-1)` : NaNs produced

Instead, we have to use the complex value `-1+0i`.

```
> sqrt(-1+0i)
```

```
[1] 0+1i
```

An alternative is to coerce `-1` into a complex value.

```
> sqrt(as.complex(-1))
```

```
[1] 0+1i
```

6.10.5 LOGICAL

A logical value is often created via comparison between variables.

```
> x = 1; y = 2 # sample values
```

```
> z = x > y # is x larger?
```

```
> z # print result
```

```
[1] FALSE
```

```
> class(z) # print class
```

```
[1] "logical"
```

Standard logical operations are `&` (and), `|` (or), and `!` (negation).

```
> u = TRUE; v = FALSE
```

```
> u & v # u AND v
```

```
[1] FALSE
```

```
> u | v # u OR v
```

```
[1] TRUE
```

```
> !u # negation of u
```

```
[1] FALSE
```

Further details and related logical operations can be found in the R documentation.

```
> help("&")
```

6.10.6 CHARACTER

A character object represents string values in R. For example, the following is a character string made from a Shakespeare quote:

```
> s = "Brevity is the soul of wit."
```

We can find out its length with the function `nchar`.

```
> nchar(s)
```

```
[1] 27
```

We can also convert simple data values into character strings with the function `as.character`.

```
> x = as.character(3.14)
```

```
> x # print x
```

```
[1] "3.14"
```

```
> class(x) # print class name
```

```
[1] "character"
```

And we can merge two character strings into one with the function `paste`.

```
> fname = "Joe"; lname = "Smith"
```

```
> paste(fname, lname)
```

```
[1] "Joe Smith"
```

However, it is often more convenient to create a readable string with the `sprintf` function, which has a C language syntax.

```
> sprintf("%s has %d dollars",
```

```
+ "Sam", 100)
```

```
[1] "Sam has 100 dollars"
```

To extract a substring, we apply the `substr` function. Here is an example showing how to extract the substring between the third and twelfth positions in a character string.

```
> substr("Mary has a little lamb.",start=3, stop=12)
```

```
[1] "ry has a l"
```

And to replace the first occurrence of the word "little" by another word "big" in the character string, we apply the `sub` function.

```
> sub("little", "big",
```

```
+ "Mary has a little lamb.")
```

```
[1] "Mary has a big lamb."
```


More functions for character string manipulation can be found in the R documentation.

```
> help("sub")
```

6.10.7 FACTOR

A factor object represents a categorical data type in R. Its sole purpose is to represent qualitative data, such as colors or shoe sizes. We can create factor values from simple data types using the function `factor`.

```
> a = factor("A")
```

The following confirms that the object is indeed a factor.

```
> class(a)
```

```
[1] "factor"
```

In particular, we can create factors from numbers:

```
> x = factor(1)
```

```
> y = factor(2)
```

Since `x` and `y` are factor values, there is no arithmetic operation allowed. Try adding them up, and we get errors instead.

```
> x + y
```

```
[1] NA
```

Warning message:

```
In Ops.factor(x, y) : + not meaningful for factors
```

More functions for handling factors can be found in the R documentation.

```
> help("factor")
```

6.10.8 VECTOR

A vector is a sequence of data elements of the same basic type. Members in a vector are officially called components. Nevertheless, we will just call them members here. Here is a vector containing three numeric members 2, 3 and 5.

```
> c(2, 3, 5)
```

```
[1] 2 3 5
```

And here is a vector of logical values.

```
> c(TRUE, FALSE, TRUE, FALSE, FALSE)
```

```
[1] TRUE FALSE TRUE FALSE FALSE
```

A vector can contain character strings.

```
> c("aa", "bb", "cc", "dd")
```

```
[1] "aa" "bb" "cc" "dd"
```

And we can find the number of members in a vector with the length function.

```
> length(c("aa", "bb", "cc", "dd"))
```

```
[1] 4
```

6.10.9 COMBINING VECTORS

Vectors can be combined via the function `c`. For examples, the following two vectors `n` and `s` are combined into a new vector containing members from both vectors.

```
> n = c(2, 3, 5)
```

```
> s = c("aa", "bb", "cc", "dd")
```

```
> c(n, s)
```

```
[1] "2" "3" "5" "aa" "bb" "cc" "dd"
```

6.10.10 VALUE COERCION

In the code snippet above, notice how the numeric values are being coerced into character strings when the two vectors are combined. This is necessary so as to maintain the same primitive data type for members in a single vector

6.10.11 VECTOR ARITHMETIC'S

Arithmetic operations on vectors are performed in a member-by-member fashion, i.e., member-wise. For example, suppose we have two vectors `a` and `b` as follows.

```
> a = c(1, 3, 5, 7)
```

```
> b = c(1, 2, 4, 8)
```

Then, if we multiply `a` by 5, we would get a vector with each of its members

multiplied by 5.

```
> 5 * a
```

```
[1] 5 15 25 35
```

And if we add a and b together, the sum would be a vector whose members are the sum of the corresponding members from a and b.

```
> a + b
```

```
[1] 2 5 9 15
```

Similarly for subtraction, multiplication and division, we get new vectors via member-wise operations.

```
> a - b
```

```
[1] 0 1 1 -1
```

```
> a * b
```

```
[1] 1 6 20 56
```

```
> a / b
```

```
[1] 1.000 1.500 1.250 0.875
```

6.10.12 LOGICAL INDEX VECTOR

A new vector can be sliced from a given vector with a logical index vector, which has the same length as the original vector. Its members are TRUE if the corresponding members in the original vector are to be included in the slice, and FALSE if otherwise.

For example, consider the following vector s of length 5.

```
> s = c("aa", "bb", "cc", "dd")
```

To retrieve the second and fourth members of s, we define a logical vector L of the same length, and have its second and fourth members set as TRUE.

```
> L = c(FALSE, TRUE, FALSE, TRUE)
```

```
> s[L]
```

```
[1] "bb" "dd"
```

The code can be abbreviated into a single line.

```
> s[c(FALSE, TRUE, FALSE, TRUE)]
```

```
[1] "bb" "dd"
```

6.10.13 MATRIX

Matrix is a collection of data elements arranged in a two-dimensional rectangular layout. The following is an example of a matrix with 2 rows and 3 columns.

$$A = \begin{bmatrix} 2 & 4 & 3 \\ 1 & 5 & 7 \end{bmatrix}$$

6.10.14 MATRIX ELEMENTS

We create a matrix in R with the namesake function from a vector. The elements in a matrix must be all of the same basic type. By default, matrix elements are arranged along the *column* direction.

```
> A = matrix( c(2, 1, 4, 5, 3, 7), nrow=2)
> A
```

```
      [,1] [,2] [,3]
[1,]  2   4   3
[2,]  1   5   7
```

We can also input matrix elements along the *row* direction by enabling the `by row` option.

```
> B = matrix( c(2, 1, 4, 5, 3, 7), nrow=2, byrow=TRUE)
> B
```

```
      [,1] [,2] [,3]
[1,]  2   1   4
[2,]  5   3   7
```

In general, a matrix of M rows and N columns is called a $M \times N$ matrix. An element at the m th row and n th column of A can be accessed via the expression $A[m, n]$.

```
> A[2, 3]           # 2nd row, 3rd column
[1] 7
```

The entire m th row of A can be extracted as $A[m,]$.

```
> A[2, ]           # the 2nd row of A
[1] 1 5 7
```

Similarly, the entire n th column of A can be extracted as $A[, n]$.

```
> A[,3]           # the 3rd column of A
```

```
[1] 3 7
```

Note that the code above produces a vector, instead of a 2x1 matrix. In order to produce the latter outcome, an extra argument, `drop`, must be explicitly set as `FALSE`.

```
> A[,3, drop=FALSE]
```

```
      [,1]
```

```
[1,]  3
```

```
[2,]  7
```

We can also extract more than one rows or columns at a time.

```
> A[,c(1,3)]
```

```
      [,1] [,2]
```

```
[1,]  2    3
```

```
[2,]  1    7
```

If we assign names to the rows and columns, then we can access matrix elements by names instead of coordinates.

```
> dimnames(A) = list(c("row1", "row2"),c("col1", "col2", "col3"))
```

```
> A
```

```
      col1 col2 col3
row1     2   4   3
row2     1   5   7
```

```
> A["row2", "col3"]
```

```
[1] 7
```

6.10.15 DATA FRAME

A data frame is used for storing data tables. It is a list of vectors of equal length. For example, in the following, `f` is a data frame containing a numeric vector `n`, a character vector `s`, and a logical vector `b`.

```
> n = c(2, 3, 5)
```

```
> s = c("aa", "bb", "cc")
```

```
> b = c(TRUE, FALSE, TRUE)
```

```
> f = data.frame(n, s, b)> help(mtcars)
```

6.11 Application of Transformation of Data

There are following applications:

6.11.1 Matrix Construction

There are various ways to construct a matrix.

6.11.2 Transpose

Consider the following 3x2 matrix B. It has 3 rows and 2 columns.

```
> B = matrix(c(2, 4, 3, 1, 5, 7), nrow=3)
```

```
> B
```

```
      [,1] [,2]
[1,]    2    1
[2,]    4    5
[3,]    3    7
```

We construct the **transpose** of a matrix by interchanging its columns and rows using the function `t`. Thus the transpose of B is a 2x3 matrix, and has 2 rows and 3 columns.

```
> t(B) # transpose
```

```
      [,1] [,2] [,3]
[1,]    2    4    3
[2,]    1    5    7
```

6.11.3 Combining Matrices

We can combine two matrices having same number of rows into a larger matrix.

For example, suppose we have another matrix C also of 3 rows.

```
> C = matrix(c(7, 4, 2),nrow=3)
```

```
> C
```

```
      [,1]
[1,]  7
[2,]  4
[3,]  2
```

Then we can combine B and C with the function `cbind`.

```
> cbind(B, C)
      [,1] [,2] [,3]
[1,]    2    1    7
[2,]    4    5    4
[3,]    3    7    2
```

Similarly, we can combine two matrices having same number of columns with the function `rbind`.

```
> D = matrix(c(6, 2), nrow=1, ncol=2)
> D # D has 2 columns
```

```
      [,1] [,2]
[1,]    6    2
```

```
> rbind(B, D)
```

```
      [,1] [,2]
[1,]    2    1
[2,]    4    5
[3,]    3    7
[4,]    6    2
```

6.11.4 Matrix Arithmetic

When the dimensions of two matrices are compatible, it is possible to perform various arithmetic operations with them.

6.11.5 Addition and Subtraction

We can add or subtract two matrices when they have the same dimensions. For example, suppose A and B are both 3x2 matrices.

```
> A = matrix(1:6, nrow=3); A
```

```
      [,1] [,2]
[1,]   1   4
[2,]   2   5
[3,]   3   6
```

```
> B = matrix(5:10, nrow=3); B
```

```
      [,1] [,2]
[1,]   5   8
[2,]   6   9
[3,]   7  10
```

Then we can compute their **sum**:

```
> A + B
```

```
      [,1] [,2]
[1,]   6  12
[2,]   8  14
[3,]  10  16
```

Similarly, we can compute their **difference**:

```
> A - B
```

```
      [,1] [,2]
[1,]  -4  -4
[2,]  -4  -4
[3,]  -4  -4
```

6.11.6 Matrix Multiplication

We can multiply two matrices together if the column dimension of the first matrix is the same as the row dimension of the second matrix.

More specifically, if the dimension of the first matrix is $M \times N$, and the dimension of the second matrix is $N \times K$, then their matrix product will be of dimension $M \times K$.

Furthermore, the element at the m th row and n th column of the product is the *dot product* of the m th row of the first matrix with the n th column of the second matrix.

For example, consider the following two matrices C and D. As C is a 3x4 matrix, and D is a 4x5 matrix, the column dimension of C matches the row dimension of D.

Hence we can define the matrix product of C and D.

```
> C = matrix(1:12, nrow=3); C
```

	[,1]	[,2]	[,3]	[,4]
[1,]	1	4	7	10
[2,]	2	5	8	11
[3,]	3	6	9	12

```
> D = matrix(-4:15, nrow=4); D
```

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	-4	0	4	8	12
[2,]	-3	1	5	9	13
[3,]	-2	2	6	10	14
[4,]	-1	3	7	11	15

We can multiply C and D together using the operator `%*%` in R. The product is a 3x5 matrix as expected.

```
> C %*% D
```

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	-40	48	136	224	312
[2,]	-50	54	158	262	366
[3,]	-60	60	180	300	420

6.12 Summary

This unit describes the essential features of R software. After studying this unit, we will be able to different types of software which helps us in mathematical calculation and statistical data analysis. R is software, and R has an advantage that R can do data manipulation; R can do statistical computing as well as simulations, R is capable of graphical display and R can also help us in doing different type of data analysis. R has an effective way of data handling. So, it can handle the data easily and it can store the input and output variables. This can store the outcome in the form of a scalar as well as form of a vectors or a matrix, and in R software, simple

calculations are possible as well as complicated calculations are also possible and it is not difficult, the mathematical calculation like addition, subtraction and these vectors and matrix, everything is possible, just like any other software.

6.13 Self-Assessment Question

The following data are given value

```
> library(MASS)
```

```
> head(painters)
```

	Composition	Drawing
Da Udine	10	8
Da Vinci	15	16
Del Piombo	8	13
Del Sarto	12	16
Fr. Penni	0	15
Guilio Romano	15	16

	Colour	Expression
Da Udine	16	3
Da Vinci	4	14
Del Piombo	16	7
Del Sarto	9	8
Fr. Penni	8	0
Guilio Romano	4	14

	School
Da Udine	A
Da Vinci	A
Del Piombo	A
Del Sarto	A
Fr. Penni	A
Guilio Romano	A

The last column in the data set `painters` contains the school information of each painter. The schools are named as A, B, ..., *etc.* The `School` column is therefore qualitative, and consists of factor values.

```
> painters$School
```

```
[1] A A A A A A A A A A B B B ...
```

```
Levels: A B C D E F G H
```

Note: For further details of the painter's data set, please consult the R documentation.

```
> help(painters)
```

1. Find the bar graph of the painter schools in the data set `painters`.
2. Find the bar graph of the composition scores in `painters`.

6.14 References

- Introduction to Statistics and Data Analysis with Exercises, Solutions and Applications in R. Christian Heumann, Michael Schomaker and Shalabh, Springer, (2022).
- Auguie, B. (2012). `gridExtra: functions in Grid graphics`. <http://CRAN.R-project.org/package=gridExtra>, R package.
- Baddeley, A.J. and Turner, R. (2005). "spatstat: An R package for analysing spatial point patterns." *Journal of Statistical Software*, 12(6), 1–42.
- Becker, R.A., Chambers, J.M., and Wilks, A.R. (1988). *The New S Language*. Chapman & Hall, UK.
- Bollen, K.A. and Jackman, R.W. (1990). "Regression diagnostics: An expository treatment of outliers and influential cases." In J. Fox and J.S. Long (eds.), "Modern Methods of Data Analysis," Sage, USA.

UNIT: 7**STATISTICAL ANALYSIS WITH R**

Structure

- 7.1 Introduction
- 7.2 Objectives
- 7.3 Descriptive Statistics
 - 7.3.1 Organization of Data
 - 7.3.1.1 Classification
 - 7.3.1.1.1 Frequency Distribution can be with Ungrouped Data and Grouped Data
 - 7.3.1.1.2 Construction of Frequency Distribution
 - 7.3.1.1.3 Types of Frequency Distribution
 - 7.3.1.2 Tabulation
 - 7.3.2 Graphical Presentation of Data
 - 7.3.2.1 Histogram
 - 7.3.2.2 Frequency Polygon
 - 7.3.2.3 Frequency Curve
 - 7.3.2.4 Cumulative Frequency Curve or Ogive
 - 7.3.3 Diagrammatic Presentation of Data
 - 7.3.3.1 Bar Diagram
 - 7.3.3.2 Sub- Divided Bar Diagram
 - 7.3.3.3 Multiple Bar Diagram
 - 7.3.3.4 Pie diagram
- 7.4 Summarisation of Statistical Data
 - 7.4.1 Measures of Central Tendency
 - 7.4.1.1 Arithmetic Mean
 - 7.4.1.2 Median
 - 7.4.1.3 Median
 - 7.4.2 Measures of Dispersion

- 7.4.2.1 Range
 - 7.4.2.2 Mean deviation
 - 7.4.2.3 Standard deviation
 - 7.4.3 Skewness and Kurtosis Statistical Data
 - 7.4.3.1 Skewness
 - 7.4.3.2 Kurtosis
- 7.5 Advantages of Descriptive Statistics
- 7.6 Disadvantages of Descriptive Statistics
- 7.7 Scattered Diagram
- 7.8 Curve Fitting
- 7.9 Correlation and Regression Analysis
 - 7.9.1 Types of Correlation
 - 7.9.2 Coefficient of Correlation
- 7.10 Regression using R
- 7.11 Graphical Presentation of some Functions using R
 - 7.11.1 Bar Plots
 - 7.11.2 Simple Bar Plots
 - 7.11.3 Stacked and Grouped Bar Plots
 - 7.11.4 Mean Bar Plots
 - 7.11.5 Spinograms
 - 7.11.6 Pie Charts
 - 7.11.7 Kernel Density Plots
 - 7.11.8 Box Plots
 - 7.11.9 Violin Plots
- 7.12 Summary
- 7.13 Self-Assessment Questions
- 7.14 References

7.1 Introduction

Welcome to the Introduction towards R Software. Well, in this course, we are going to learn about different aspects of software, what is called as R. So, in this section, we are going to understand that why should we learn R and how should we install the R software and related software on our computer. The basic idea is that we are am going to start this section at a very elementary level so that anyone who has no idea about even about how to install the software or how to operate the computer even he should be able to understand it.

7.2 Objectives

After studying this unit, you will be able to understand the following objectives:

- Studying of Statistical Analysis using R.
- Studying of Descriptive Statistics.
- Studying of Curve fitting.
- Studying of Correlation and regression analysis
- Studying of Graphs

7.3 Descriptive Statistics

Descriptive statistics is a branch of statistics, which deals with descriptions of obtained data. On the basis of these descriptions a particular group of population is defined for corresponding characteristics. Most of the observations in this universe are subject to variability, especially observations related to human behavior. It is a well-known fact that attitude, intelligence and personality differ from individual to individual. In order to make a sensible definition of the group or to identify the group with reference to their observations/ scores, it is necessary to express them in a precise manner. For this purpose, observations need to be expressed as a single estimate which summary, the observations The descriptive statistics include classification, tabulation, diagrammatic and graphical presentation of data, measures of central tendency and variability. These measures enable the researchers to know about the tendency of data or the scores, which further enhance the ease in description of the phenomena. Such single estimate of the series of data which summaries the distribution known as parameters of the distribution. These parameters define the distribution completely.

Descriptive statistics involves two operations:

- (1) Organisation of Data.
- (2) Summarization of Data

7.3.1 Organization of Data

There are four major statistical techniques for organizing the data. These are:

- (i) Classification
- (ii) Tabulation
- (iii) Graphical Presentation, and
- (iv) Diagrammatical Presentation

7.3.1.1 Classification

The arrangement of data in groups according to similarities is known as classification.

A classification is a summary of the frequency of individual scores or ranges of scores for a variable. In the simplest form of a distribution, we will have such value of variable as well as the number of persons who have had each value.

Once data are collected, it should be arranged in a format from which they would be able to draw some conclusions. Thus, by classifying data, the investigators move a step ahead in regard to making a decision.

A much clear picture of the information of score emerges when the raw data are organized as a frequency distribution. Frequency distribution shows the number of cases following within a given class interval or range of scores. A frequency distribution is a table that shows each score as obtained by a group of individuals and how frequently each score occurred

7.3.1.1.1 Frequency Distribution can be with Ungrouped Data and Grouped Data

- (i) An ungrouped frequency distribution may be constructed by listing all score values either from highest to lowest or lowest to highest and placing a tally mark (/) besides each scores every times it occurs. The frequency of occurrence of each score is denoted by 'f'.

(ii) Grouped frequency distribution: If there is a wide range of score value in the data, then it is difficult to get a clear picture of such series of data. In this case grouped frequency distribution should be constructed to have a clear picture of the data. A group frequency distribution is a table that organizes data into classes.

It shows the number of observations from the data set that fall into each of the class.

7.3.1.1.2 Construction of Frequency Distribution

To prepare a frequency distribution it is essential to determine the following:

- (1) The range of the given data =, the difference between the highest and lowest scores.
- (2) The number of class intervals = There is no hard and fast rules regarding the number of classes into which data should be grouped. If there are very few scores it is useless to have a large number of class-intervals. Ordinarily, the number of classes should be between 5 to 30.
- (3) Limits of each class interval = another factor used in determining the number of classes is the size/ width or range of the class which is known as 'class interval' and is denoted by 'i'. Class interval should be of uniform width resulting in the same-size classes of frequency distribution. The width of the class should be a whole number and conveniently divisible by 2, 3, 5, 10, or 20.

There are three methods for describing the class limits for distribution:

- (i) Exclusive method,
- (ii) Inclusive method
- (iii) True or actual class method.

(i) Exclusive Method: In this method of class formation, the classes are so formed that the upper limit of one class become the lower limit of the next class. In this classification, it is presumed that score equal to the upper limit of the class is exclusive, i.e., a score of 40 will be included in the class of 40 to 50 and not in a class of 30 to 40 (30-40, 40-50, 50-60)

(ii) Inclusive Method: In this method the classes are so formed that the upper limit of one class does not become the lower limit of the next class. This classification includes scores, which are equal to the upper limit of the class. Inclusive method is preferred when measurements are given in whole numbers. (30-39, 40-49, 50-59)

(iii) True or Actual Class Method: Mathematically, a score is an internal when it extends from 0.5 units below to 0.5 units above the face value of the score on a continuum. These class limits are known as true or actual class limits. (29.5 to 39.5, 39.5 to 49.5) etc.

7.3.1.1.3 Types of Frequency Distribution

There are various ways to arrange frequencies of a data array based on the requirement of the statistical analysis or the study. A couple of them are discussed below:

(i) Relative Frequency Distribution: A relative frequency distribution is a distribution that indicates the proportion of the total number of cases observed at each score value or internal of score values.

(ii) Cumulative Frequency Distribution: Sometimes investigator may be interested to know the number of observations less than a particular value. This is possible by computing the cumulative frequency. A cumulative frequency corresponding to a class-interval is the sum of frequencies for that class and of all classes prior to that class.

(iii) Cumulative Relative Frequency Distribution: A cumulative relative frequency distribution is one in which the entry of any score of class interval expresses that score's cumulative frequency as a proportion of the total number of cases.

7.3.1.2 Tabulation

Frequency distribution can be either in the form of a table or it can be in the form of graph. Tabulation is the process of presenting the classified data in the form of a table. A tabular presentation of data becomes more intelligible and fit for further statistical analysis. A table is a systematic arrangement of classified data in row and columns with appropriate headings and sub-headings. The main components of a table are:

(i) Table Number: When there is more than one table in a particular analysis a table should be marked with a number for their reference and identification. The number should be written in the center at the top of the table.

(ii) Title of the Table: Every table should have an appropriate title, which describes the content of the table. The title should be clear, brief, and self-explanatory.

Title of the table should be placed either centrally on the top of the table or just below or after the table number.

(iii) Caption: Captions are brief and self-explanatory headings for columns. Captions may involve headings and sub-headings. The captions should be placed in the middle of the columns. For example, we can divide students of a class into males and females, rural and urban, high SES and Low SES etc.

(iv) Stub: Stubs stand for brief and self-explanatory headings for rows.

(v) Body of the Table: This is the real table and contains numerical information or data in different cells. This arrangement of data remains according to the description of captions and stubs.

(vi) Head Note: This is written at the extreme right hand below the title and explains the unit of the measurements used in the body of the tables.

(vii) Footnote: This is a qualifying statement which is to be written below the table explaining certain points related to the data which have not been covered in title, caption, and stubs.

(viii) Source of Data: The source from which data have been taken is to be mentioned at the end of the table.

7.3.2 Graphical Presentation of Data

The purpose of preparing a frequency distribution is to provide a systematic way of “looking at” and understanding data. To extend this understanding, the information contained in a frequency distribution often is displayed in graphic and/or diagrammatic forms. In graphical presentation of frequency distribution, frequencies are plotted on a pictorial platform formed of horizontal and vertical lines known as graph.

A graph is created on two mutually perpendicular lines called the X and Y–axes on which appropriate scales are indicated. The horizontal line is called the abscissa and vertical the ordinate. Like different kinds of frequency distributions there are many kinds of graph too, which enhance the scientific understanding of the reader. The commonly used graphs are Histogram, Frequency polygon, Frequency curve, Cumulative frequency curve. Here we will discuss some of the important types of graphical patterns used in statistics. We shall illustrate the method of creating line plots, with the help of some suitable problems. We illustrate the methods of creation

of graphs of gamma and beta functions. Different functions such as plot(), abline(), curve(), text(), outer(), persp() and

7.3.2.1 Histogram

It is one of the most popular methods for presenting continuous frequency distribution in a form of graph. In this type of distribution, the upper limit of a class is the lower limit of the following class. The histogram consists of series of rectangles, with its width equal to the class interval of the variable on horizontal axis and the corresponding frequency on the vertical axis as its heights.

7.3.2.2 Frequency Polygon

Prepare an abscissa originating from 'O' and ending to 'X'. Again, construct the ordinate starting from 'O' and ending at 'Y'. Now label the class-intervals on abscissa stating the exact limits or midpoints of the class intervals. You can also add one extra limit keeping zero frequency on both side of the class-interval range.

The size of measurement of small squares on graph paper depends upon the number of classes to be plotted. Next step is to plot the frequencies on ordinate using the most comfortable measurement of small squares depending on the range of whole distribution.

To plot a frequency polygon, you have to mark each frequency against its concerned class on the height of its respective ordinate. After putting all frequency marks a draw a line joining the points. This is the polygon.

7.3.2.3 Frequency Curve

A frequency curve is a smooth free hand curve drawn through frequency polygon. The objective of smoothing of the frequency polygon is to eliminate as far as possible the random or erratic fluctuations that are present in the data.

7.3.2.4 Cumulative Frequency Curve or Ogive

The graph of a cumulative frequency distribution is known as cumulative frequency curve or ogive. Since there are two types of cumulative frequency distribution e.g., "less than" and "more than" cumulative frequencies. We can have two types of ogives.

(i) **‘Less than’ Ogive:** In ‘less than’ ogive, the less than cumulative frequencies are plotted against the upper-class boundaries of the respective classes. It is an increasing curve having slopes upwards from left to right.

(ii) **‘More than’ Ogive:** In more than ogive, the more than cumulative frequencies are plotted against the lower-class boundaries of the respective classes. It is decreasing curve and slopes downwards from left to right.

7.3.3 Diagrammatic Presentation of Data

A diagram is a visual form for the presentation of statistical data. They present the data in simple, readily comprehensible form. Diagrammatic presentation is used only for presentation of the data in visual form, whereas graphic presentation of the data can be used for further analysis. There are different forms of diagram e.g., Bar diagram, Sub-divided bar diagram, Multiple bar diagram, Pie diagram and Pictogram.

7.3.3.1 Bar Diagram

Bar diagram is most useful for categorical data. A bar is defined as a thick line. Bar diagram is drawn from the frequency distribution table representing the variable on the horizontal axis and the frequency on the vertical axis. The height of each bar will be corresponding to the frequency or value of the variable.

7.3.3.2 Sub- Divided Bar Diagram

Study of sub classification of a phenomenon can be done by using sub-divided bar diagram. Corresponding to each sub-category of the data the bar is divided and shaded. There will be as many shades as there will sub portion in a group of data. The portion of the bar occupied by each sub-class reflects its proportion in the total.

7.3.3.3 Multiple Bar Diagram

This diagram is used when comparisons are to be shown between two or more sets of interrelated phenomena or variables. A set of bars for person, place or related phenomena are

drawn side by side without any gap. To distinguish between the different bars in a set, different colours , shades are used.

7.3.3.4 Pie Diagram

It is also known as angular diagram. A pie chart or diagram is a circle divided into component sectors corresponding to the frequencies of the variables in the distribution. Each sector will be proportional to the frequency of the variable in the group. A circle represents 360°. So 360° angles is divided in proportion to percentages. The degrees represented by the various component parts of given magnitude can be obtained by using this formula. After the calculation of the angles for each component, segments are drawn in the circle in succession, corresponding to the angles at the center for each segment. Different segments are shaded with different colours, shades or numbers.

7.4 Summarisation of Statistical Data

In the previous section we have discussed about tabulation of the data and its representation in the form of graphical presentation. In research, comparison between two or more series of the same type is needed to find out the trends of variables.

For such comparison, tabulation of data is not sufficient and it is further required to investigate the characteristics of data. The frequency distribution of obtained data may differ in two ways, first in measures of central tendency and second, in the extent to which scores are spread over the central value. Both types of differences are the components of summary statistics

7.4.1 Measures of Central Tendency

It is the middle point of a distribution. Tabulated data provides the data in a systematic order and enhances their understanding. Generally, in any distribution values of the variables tend to cluster around a central value of the distribution. This tendency of the distribution is known as central tendency and measures devised to consider this tendency is known as measures of central tendency. A measure of central tendency is useful if it represents accurately the

distribution of scores on which it is based. A good measure of central tendency must possess the following characteristics:

- It should be clearly defined- The definition of a measure of central tendency should be clear and unambiguous so that it leads to one and only one information.
- It should be readily comprehensible and easy to compute.
- It should be based on all observations- A good measure of central tendency should be based on all the values of the distribution of scores.
- It should be amenable for further mathematical treatment.
- It should be least affected by the fluctuation of sampling.

In Statistics there are three most commonly used measures of central tendency.

These are:

- 1) Arithmetic Mean
- 2) Median
- 3) Mode

7.4.1.1 Arithmetic Mean

The arithmetic mean is most popular and widely used measure of central tendency. Whenever we refer to the average of data, it means we are talking about its arithmetic mean. This is obtained by dividing the sum of the values of the variable by the number of values. It is also a useful measure for further statistics and comparisons among different data sets. One of the major limitations of arithmetic mean is that it cannot be computed for open-ended class-intervals.

Example: Find the mean eruption duration in the data set faithful.

We apply the mean function to compute the mean value of eruptions.

```
> duration = faithful$eruptions
```

```
> mean(duration)
```

```
[1] 3.4878
```

The mean eruption duration is about 3.5 minutes

7.4.1.2 Median

Median is the middle most value in a data distribution. It divides the distribution into two equal parts so that exactly one half of the observations is below and one half is above that point. Since median clearly denotes the position of an observation in an array, it is also called a position average. Thus more technically, median of an array of numbers arranged in order of their magnitude is either the middle value or the arithmetic mean of the two middle values. It is not affected by extreme values in the distribution.

Example: Find the median of eruption duration in the data set faithful.

We apply the median function to compute the median value of eruptions.

```
> duration = faithful$eruptions
```

```
> median(duration)
```

```
[1] 4
```

7.4.1.3 Mode

Mode is the value in a distribution that corresponds to the maximum concentration of frequencies. It may be regarded as the most typical of a series value. In more simple words, mode is the point in the distribution comprising maximum frequencies therein.

7.4.2 Measures of Dispersion

In the previous section we have discussed about measures of central tendency. By knowing only, the mean, median or mode, it is not possible to have a complete picture of a set of data. Average does not tell us about how the score or measurements are arranged in relation to the center. It is possible that two sets of data with equal mean or median may differ in terms of their variability. Therefore, it is essential to know how far these observations are scattered from each other or from the mean. Measures of these variations are known as the 'measures of dispersion'. The most commonly used measures of dispersion are range, average deviation, quartile deviation, variance and standard deviation.

7.4.2.1 Range

Range is one of the simplest measures of dispersion. It is designated by 'R'. The range is defined as the difference between the largest score and the smallest score in the distribution. It gives the two extreme values of the variable. A large value of range indicates greater dispersion while a smaller value indicates lesser dispersion among the scores. Range can be a good measure if the distribution is not much skewed.

Example: Find the range of eruption duration in the data set faithful.

We apply the max and min functions to compute the largest and smallest values of eruptions, and then we take their difference.

```
> duration = faithful$eruptions
> max(duration) - min(duration)
[1] 3.5
```

Find the interquartile range of eruption duration in the data set faithful

We apply the IQR function to compute the interquartile range of eruptions.

```
> duration = faithful$eruptions
> IQR(duration)
[1] 2.2915
```

7.4.2.2 Mean Deviation

Average deviation refers to the arithmetic mean of the differences between each score and the mean. It is always better to find the deviation of the individual observations with reference to a certain value in the series of observation and then take an average of these deviations. This deviation is usually measured from mean or median. Mean, however, is more commonly used for this measurement.

Remark: It is less affected by extreme values as compared to standard deviation. It provides better measure for comparison about the formation of different distributions.

7.4.2.3 Standard Deviation

Standard deviation is the most stable index of variability. In standard deviation, instead of the actual values of the deviations we consider the squares of deviations and the outcome is known as variance. Further, the square root of this variance is known as standard deviation and

designated as SD. Thus, standard deviation is the square root of the mean of the squared deviations of the individual observations from the mean. If all the score has an identical value in a sample, the SD will be 0 (zero).

Example: (i) Find the standard deviation of eruption duration in the data set faithful.

We apply the SD function to compute the standard deviation of eruptions.

```
> duration = faithful$eruptions
```

```
> sd(duration)
```

```
[1] 1.1414
```

(ii) Find the variance of eruption duration in the data set faithful

We apply the var function to compute the variance of eruptions.

```
> duration = faithful$eruptions
```

```
> var(duration)
```

```
[1] 1.3027
```

Remark: It is based on all observations. It is amenable to further mathematical treatments.

Of all measures of dispersion, standard deviation is least affected by fluctuation of sampling.

7.4.3 Skewness and Kurtosis of Statistical Data

There are two other important characteristics of frequency distribution that provide useful information about its nature. They are known as skewness and kurtosis.

7.4.3.1 Skewness

Skewness is the degree of asymmetry of the distribution. In some frequency distributions scores are more concentrated at one end of the scale. Such a distribution variability are usually related, the more the skewness the greater the variability.

Example: Find the skewness of eruption duration in the data set faithful

We apply the function skewness from the package to compute the skewness coefficient of eruptions. As the package is not in the core R library, it has to be installed and loaded into the R workspace.

```
> library(e1017)
```

```
> duration = faithful$eruptions
```

```
> skewness(duration)
```

```
[1] -0.41355
```

The skewness of eruption duration is -0.41355. It indicates that the eruption duration distribution is skewed towards the left.

7.4.3.2 Kurtosis

The term 'kurtosis' refers to the 'peakedness' or flatness of a frequency distribution curve when compared with normal distribution curve. The kurtosis of a distribution is the curvedness or peakedness of the graph. If a distribution is more peaked than normal it is said to be leptokurtic. This kind of peakedness implies a thin distribution. If a distribution is more flat than the normal distribution it is known as Platykurtic distribution. A normal curve is known as Mesokurtic.

Example: Find the kurtosis of eruption duration in the data set faithful.

We apply the function kurtosis from the e1071 package to compute the kurtosis of eruptions. As the package is not in the core R library, it has to be installed and loaded into the R workspace.

```
> library(e1071)
```

```
> duration = faithful$eruptions
```

```
> kurtosis(duration)
```

```
[1] -1.5116
```

The kurtosis of eruption duration is -1.5116, which indicates that eruption duration distribution is platykurtic. This is consistent with the fact that its histogram is not bell-shaped.

7.5 Advantages of Descriptive Statistics

The Advantages of Descriptive statistics are given below:

- It is essential for arranging and displaying data.
- It forms the basis of rigorous data analysis.
- It is easier to work with, interpret, and discuss than raw data.
- It helps in examining the tendencies, variability, and normality of a data set.
- It can be rendered both graphically and numerically.

- It forms the basis for more advanced statistical methods.

7.6 Disadvantages of Descriptive Statistics

The disadvantages of descriptive statistics can be listed as given below:

- It can be misused, misinterpreted, and incomplete.
- It can be of limited use when samples and populations are small.
- It offers little information about causes and effects.
- It can be dangerous if not analysed completely. There is a risk of distorting the original data or losing important detail.

Descriptive statistics is a branch of statistics, which deals with descriptions of obtained data. On the basis of these descriptions a particular group of population is defined for corresponding characteristics. The descriptive statistics include classification, tabulation, diagrammatic and graphical presentation of data, measures of central tendency and variability. These measures enable the researchers to know about the tendency of data or the scores, which further enhance the ease in description of the phenomena. Such single estimate of the series of data which summarizes the distribution are known as parameters of the distribution. These parameters define the distribution completely. Statistics is a branch of knowledge that is used to summarize and present data. The word 'Statistic' has been derived from the Latin word 'status' meaning simply the collection of numerical data on different aspects of the life of the people useful to the state. However, the scope broadened to include collection of numerical data in tabular and graphical form. The statistical data must be arranged and compiled in a systematic and organized way for comprehensive understanding of the compiled information.

Therefore, statistical analyses are used for the following purposes to:

- (i) Summarize the data for analysis and interpretation
- (ii) Describe and summarize researcher's observation and measurements of variables under study. (i.e. Descriptive statistics)
- (iii) Test hypotheses and to estimate the population parameters (i.e. Inferential statistics).

In this unit you will learn about measures of variability as range and standard deviation. The meaning, use and computation SD standard deviation are the most commonly used to measure variability for research data. This unit also describes the meaning and computation of

correlation coefficient to find the relationship between two variables. There are different methods of computing correlation coefficient but this unit describes about rank difference method and two major statistical tests which have been presented to compute correlation. Measures of variability refer to the "spread or "dispersion of the scores around the central tendency.

7.7 Scatter Diagram

Scatter diagram is a statistical tool for determining the potentiality of correlation between dependent variable and independent variable. Scatter diagram does not tell about exact relationship between two variables but it indicates whether they are correlated or not. Let (X_i, Y_i) ($i= 1,2,\dots,n$) be the bivariate distribution. If the values of the dependent variable Y are plotted against corresponding values of the independent variable X in the XY plane, such diagram of dots is called scatter diagram or dot diagram. It is to be noted that scatter diagram is not suitable for large number of observations.

7.8 Curve Fitting

Let Y and X be the dependent and independent variables respectively and we have a set of values $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$, i.e. observations are taken from n individuals on X and Y. We are interested in studying the function $Y = f(X)$. If \hat{Y}_i is the estimated value of Y obtained by the function and Y_i is the observed value of Y at x_i then we can define residual. The difference between \hat{y}_i and Y_i i.e. the difference between observed value and estimated value is called error of the estimate or residual for Y_i .

Principle of least squares consists in minimizing the sum of squares of the residuals, i.e. according to principle of least squares

$$W = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \text{ Should be minimum}$$

Let us consider a curve (function) of the type $Y = a + bX + cX^2 + \dots \dots \dots tX^k \dots \dots \dots (A)$

Where, Y is dependent variable, X is independent variable and a, b, c, ..., t are unknown constants. Suppose we have $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ values of two variables (X, Y)

i.e. data for variables X and Y. These variables may be height and weight, sales and profit, rainfall and production of any crop, etc. In all these examples, first variables, i.e. height, sales and rainfall seem to be independent variables, while second variables, i.e. weight, profit and production of crop seem to be dependent variables. With the given values have $(X_1, Y_1) (X_2, Y_2), \dots, (X_n, Y_n)$ curve (function) given in equation (A) produces set of n equations.

$$Y_1 = a + bX_1 + cX_1^2 + \dots \dots tX_1^k$$

$$Y_2 = a + bX_2 + cX_2^2 + \dots \dots tX_2^k$$

.
.

.....

..... (B)

$$Y_n = a + bX_n + cX_n^2 + \dots \dots tX_n^k$$

Our problem is to determine the constants a, b, c, ..., t such that it represents the curve of best fit given by equation (A) of degree k.

If $n = k+1$, i.e. number of equations and number of unknown constants are equal, there is no problem in determining the unknown constants and error can be made absolutely zero. But more often $n > k+1$ i.e. number of equations is greater than the number of unknown constants and it is impossible to do away with all errors i.e. these equations cannot be solved exactly which satisfy set of equations. Therefore, we try to determine the values of a, b, c, ..., t which satisfy set of equations (B) as nearly as possible.

Substituting X_1, X_2, \dots, X_n for X in equation (A) we have

$$Y_1 = a + bX_1 + cX_1^2 + \dots \dots tX_1^k$$

$$Y_2 = a + bX_2 + cX_2^2 + \dots \dots tX_2^k$$

.
.
.
.

.....

..... (C)

$$Y_n = a + bX_n + cX_n^2 + \dots \dots tX_n^k$$

The Quantities Y_1, Y_2, \dots, Y_n are called expected or estimated values of Y_1, Y_2, \dots, Y_n (given values of Y) for the given values of X_1, X_2, \dots, X_n . Here Y_1, Y_2, \dots, Y_n are the observed values of Y.

Let us define a quantity W, the sum of squares of errors i.e.

$$W = \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

$$W = \sum_{i=1}^n (Y_i - a - bX_i - cX_i^2 - \dots - tX_i^k)^2 \dots \dots \dots (D)$$

According to the principle of least squares the constant a, b, ..., t are chosen in such a way that the sum of squares of residuals is minimum. According to principle of maxima and minima (theorem of differential calculus), the extreme value (maximum or minimum) of the function W are obtained by

$$\frac{dW}{da} = 0 = \frac{dW}{db} = 0 = \frac{dW}{dc} = 0 \quad \dots = \frac{dW}{dt} = 0$$

(Provided that the partial derivatives exist)

Let us take

$$\frac{dW}{da} = 0$$

$$\frac{d}{da} \sum_{i=1}^n (Y_i - \bar{Y})^2 = 0$$

$$\frac{d}{da} \sum_{i=1}^n (Y_i - a - bX_i - cX_i^2 - \dots - tX_i^k)^2 = 0$$

$$2 \sum_{i=1}^n (Y_i - a - bX_i - cX_i^2 - \dots - tX_i^k) - (1) = 0$$

$$\sum_{i=1}^n Y_i = na + b\sum X_i + c\sum X_i^2 + \dots + t\sum X_i^k \dots \dots \dots (E)$$

$$\frac{d}{db} \sum_{i=1}^n (Y_i - \bar{Y})^2 = 0$$

$$\frac{d}{db} \sum_{i=1}^n (Y_i - a - bX_i - cX_i^2 - \dots - tX_i^k)^2 = 0$$

$$2 \sum_{i=1}^n (Y_i - a - bX_i - cX_i^2 - \dots - tX_i^k) - (X_i) = 0$$

$$\sum_{i=1}^n Y_i X_i = a\sum X_i + b\sum X_i^2 + c\sum X_i^3 + \dots + t\sum X_i^{k+1} \dots \dots \dots (F)$$

Therefore, by the conditions

$$\frac{dW}{da} = 0 = \frac{dW}{db} = 0 = \frac{dW}{dc} = 0 \quad \dots = \frac{dW}{dt} = 0$$

ultimately we get the following (k+1) equations

$$\sum_{i=1}^n Y_i = na + b\sum X_i + c\sum X_i^2 + \dots + t\sum X_i^k.$$

$$\sum_{i=1}^n Y_i X_i = a\sum X_i + b\sum X_i^2 + c\sum X_i^3 + \dots + t\sum X_i^{k+1}$$

.....(G)

$$\sum_{i=1}^n Y_i X_i^k = a\sum X_i^k + b\sum X_i^{k+1} + c\sum X_i^{k+2} + \dots + t\sum X_i^{2k}$$

These equations are solved as simultaneous equations and give the value of (k+1) constants a, b, c, ..., t. Substitution of these values in second order partial derivatives gives positive value of the function. Positive value of the function indicates that the values of a, b, c, ..., t obtained by solving the set of equations (G), minimize U which is sum of squares of residuals . With these values of a, b, c, ..., t , curve in equation (A) is the curve of best fit.

7.9 Correlation Analysis

We often encounter situations where data appears as pairs of figures relating to two variables. A correlation problem considers the joint variation of two measurements neither of which is restricted by the experimenter. Considers the frequency distributions of one variable (called the dependent variable) when another (independent variable) is held fixed at each of several levels. **Examples:** we are study of the relationship between IQ and aggregate percentage marks obtained by a person in UPPSC examination, blood pressure and metabolism or the relation between height and weight of individuals. In these examples both variables are observed as they naturally occur, since neither variable is fixed at predetermined levels. In practical applications, we might come across the situation where observations are available on two or more variables. The following examples will illustrate the situations clearly:

(1). Heights and weights of persons of a certain group;
(2). Sales revenue and advertising expenditure in business; and
(3). Time spent on study and marks obtained by students in exam. If data are available for two variables, say X and Y, it is called bivariate distribution. Let us consider the example of sales revenue and expenditure on advertising in business. A natural question arises in mind that is there any connection between sales revenue and expenditure on advertising? Does sales revenue increase or decrease as expenditure on advertising increases or decreases? If we see the example of time spent on study and marks obtained by students, a natural question appears whether marks increase or decrease as time spent on study increase or decrease. In all these situations, we try to find out relation between two variables and correlation answers the question, if there is any relationship between one variable and another. When two variables are related in such a way that change in the value of one variable affects the value of another variable, then variables are said to be correlated or there is correlation between these two variables.

7.9.1 Types of Correlation

There are two types of correlation.

(1). Positive Correlation: Correlation between two variables is said to be positive if the values of the variables deviate in the same direction i.e. if the values of one variable increase (or decrease) then the values of other variable also increase (or decrease). Some examples of positive correlation are correlation between

- (1). Heights and weights of group of persons;
- (2). House hold income and expenditure;
- (3). Amount of rainfall and yield of crops;
- (4). Expenditure on advertising and sales revenue.

Thus, the change is in the same direction. Hence the correlation is positive.

(2). Negative Correlation: Correlation between two variables is said to be negative if the values of variables deviate in opposite direction i.e. if the values of one variable increase (or decrease) then the values of other variable decrease (or increase). Some examples of negative correlations are correlation between

- (1). Volume and pressure of perfect gas.

- (2). Price and demand of goods.
- (3). Literacy and poverty in a country.
- (4). Time spent on watching TV and marks obtained by students in examination. Thus the change is in opposite direction. Therefore, the correlation between volume and pressure is negative. In remaining three examples also, values of the second variable change in the opposite direction of the change in the values of first variable.

7.9.2 Coefficient of Correlation

Scatter diagram tells us whether variables are correlated or not. But it does not indicate the extent of which they are correlated. Coefficient of correlation gives the exact idea of the extent of which they are correlated. Coefficient of correlation measures the intensity or degree of linear relationship between two variables.

Example: Find the correlation coefficient of eruption duration and waiting time in the data set faithful. Observe if there is any linear relationship between the variables.

We apply the cor function to compute the correlation coefficient of eruptions and waiting.

```
> duration = faithful$eruptions
> waiting = faithful$waiting
> cor(duration, waiting)
[1] 0.90081
```

The correlation coefficient of eruption duration and waiting time is 0.90081. Since it is close to one, we conclude that the variables are positively linearly correlated.

Remark: Linear relationships can be expressed in such a way that the independent variable is multiplied by the slope coefficient, added by a constant, which determines the dependent variable. If Y is a dependent variable, X is an independent variable, b is a slope coefficient and a is constant then linear relationship is expressed as $Y = a + b * X$. In fact linear relationship is the relationship between dependent and independent variables of direct proportionality. When these variables plotted on a graph give a straight line.

If X and Y are two random variables then correlation coefficient between X and Y is denoted by r and defined as

$$r = Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{V(X) * V(Y)}} \dots\dots\dots(A)$$

Correlation (x,y) is indication of correlation coefficient between two variables X and Y.

7.10 Regression Analysis using R

Regression analysis is a very widely used statistical tool to establish a relationship model between two variables. One of these variable is called predictor variable whose value is gathered through experiments. The other variable is called response variable whose value is derived from the predictor variable.

In Linear Regression these two variables are related through an equation, where exponent (power) of both these variables is one. Mathematically a linear relationship represents a straight line when plotted as a graph. A non-linear relationship where the exponent of any variable is not equal to one creates a curve. Simple example of regression is predicting weight of a person when his height is known. To do this we need to have the relationship between height and weight of a person. The general mathematical equation for a linear regression is $y = ax + b$

Carry out the experiment of gathering a sample of observed values of height and corresponding weight. Create a relationship model using the **lm()** functions in R. Find the coefficients from the model created and create the mathematical equation using these to predict the weight of new persons, use the **predict ()** function in R.

Regression analysis enables us in estimation and forecasting of the value of dependent variable for given value(Y) of the independent variable(X) and is extensively used in various disciplines. Clearly, while in simple linear regression there is a single independent variable; in multiple regression there are more than one independent variable. In case of two variables only; the latter is concerned with the study of more than two variables. Further, in case of simple regression if the relationship between the dependent and independent variables follows a straight line pattern, it is called linear regression. On the other hand, if the relation is expressed in the form of a curve, it is called curvilinear regression.

Linear Regression Here we are going to study the case of two variables X and Y. Thus, there are two lines of regression viz., X on Y and Y on X .The regression line on gives the most probable values of for a given value of and the regression line on gives the most probable values of for a given value of X for given value of Y . However, when there is a perfect correlation between X & Y i.e. +1 and -1, the two regression lines Y on X and X on Y coincides i.e. we will have one

regression line. Otherwise, there are two lines of regression which coincide at (mean(X), mean(Y)).

The regression line Y on X is given by

$$(Y - \bar{Y}) = b_{YX}(X - \bar{X})$$

Where $b_{YX} = r \frac{\sigma_y}{\sigma_x}$ is regression coefficient of Y on X. and \bar{X} is mean of X, \bar{Y} is mean of Y.

The regression line X on Y is given by

$$(X - \bar{X}) = b_{XY}(Y - \bar{Y})$$

Where $b_{XY} = r \frac{\sigma_x}{\sigma_y}$ is regression coefficient of X on Y. and \bar{X} is Mean of X, \bar{Y} is Mean of Y.

Example: Assuming the simple linear regression model for the data set faithful, estimate the next eruption duration if the waiting time since the last eruption has been 80 minutes.

We apply the function lm to a formula that describes the dependent variable eruptions by the independent variable waiting, and save the new linear regression model in a variable eruption.lm.

```
> eruption.lm = lm( eruptions ~ waiting,data=faithful)
```

Then we extract the parameters of the estimated regression equation with a function named coefficients.

```
> coeffs = coefficients(eruption.lm)
```

```
> coeffs (Intercept) waiting
```

```
-1.874016 0.075628
```

We now fit the eruption duration using the estimated regression equation.

```
> waiting = 80
```

```
> duration = coeffs[1] coeffs[2]*waiting
```

```
> duration (Intercept)
```

```
4.1762
```

Conclusion: Based on the simple linear regression model, if the waiting time since the last eruption has been 80 minutes, we expect the eruption to last about 4 minutes.

7.11 Graphical Presentation of some Functions using R

We have learnt that the visual presentation of data eliminates the dullness in the presentation of quantitative data and makes it more interesting. Visual presentation also helps in

comparison of data or determining the trends of the past performance. We have already studied about one of the techniques of visual presentation of data i.e., diagrammatic presentation. Another important technique of visual presentation of data is the presentation in the form of graphs. In this unit you will learn about the principles of preparing graphs, different types of graphs for time series and frequency distributions and the methods of preparing them. The term chart is used to refer to

- (i) A detailed map of a sea area. and
- (ii) Information presented in the form of a picture or a graph to make it easily understood.

We shall use it here in the latter sense. This presentation through diagrams and graphs can take various forms, including what is known as a flow diagram or a flowchart. These charts simplify the detailed information that is presented and help in its interpretation. Trends, movements, and distributions can be presented in a more comprehensive manner in graphs than in tables

In this unit, we shall mainly discuss different methods of creation of creating plots and graphs of different types of functions Using R software. We shall also discuss the method of computation of the cardinality of a given function.

The graphic presentation of data had the following advantages:

- (i) Graphic presentation of data renders comparison of data easier. The direction of curves or straight lines on the graphs makes it very simple and to draw comparisons.
- (ii) Graphic presentation of data helps in establishing trends of the past the presentation of time series data on a graph makes it possible to interpolate or extrapolate the values. Thus, it helps in forecasting.
- (iii) Graphic presentation of data makes it possible to determine the values of the positional averages such as median, quartiles, mode, etc. The graphs of frequency distribution help us to locate these values.
- (iv). Through Graphical presentation it is also possible to establish correlation between two variables. Scatter diagram is a graphic presentation technique to determine the degree of correlation

7.11.1 Bar Plots

A bar plot displays the distribution (frequency) of a categorical variable through vertical or horizontal bars. In its simplest form, the format of the `barplot()` function is `barplot(height)` where `height` is a vector or matrix. In the following examples we will plot the outcome of a study investigating a new treatment for rheumatoid arthritis. The data are contained in the `Arthritis` data frame distributed with the `vcd` package. This package isn't included in the default R installation, so install it before first use (`install.packages("vcd")`).

7.11.2 Simple Bar Plots

If `height` is a vector, the values determine the heights of the bars in the plot, and a vertical bar plot is produced. Including the option `horiz=TRUE` produces a horizontal bar chart instead. You can also add annotating options. The `main` option adds a plot title, whereas the `xlab` and `ylab` options add x-axis and y-axis labels, respectively.

In the `Arthritis` study, the variable `Improved` records the patient outcomes for individuals receiving a placebo or drug:

```
> library(vcd)
> counts <- table(Arthritis$Improved)
> counts
None Some Marked
42 14 28
```

Here, we see that 28 patients showed marked improvement, 14 showed some improvement, and 42 showed no improvement.

```
barplot(counts, main="Simple Bar Plot", xlab="Improvement", ylab="Frequency")
barplot(counts, main="Horizontal Bar Plot", xlab="Frequency", ylab="Improvement",
horiz=TRUE)
```

7.11.3 Stacked and Grouped Bar Plots

If `height` is a matrix rather than a vector, the resulting graph will be a stacked or grouped bar plot. If `beside=FALSE` (the default), then each column of the matrix produces a bar in the

plot, with the values in the column giving the heights of stacked “sub-bars.” If `beside=TRUE`, each column of the matrix represents a group, and the values in each column are juxtaposed rather than stacked. Consider the cross-tabulation of treatment type and improvement status

```
> library(vcd)
> counts <- table(Arthritis$Improved, Arthritis$Treatment)
> counts
Treatment
Improved Placebo Treated
None 29 13
Some 7 7
Marked 7 21
barplot(counts,main="StackedBarPlot",xlab="Treatment",ylab="Frequency",col=c("red","yellow",
"green"),
legend=rownames(counts))
barplot(counts,main="GroupedBarPlot",xlab="Treatment",ylab="Frequency",col=c("red","yellow",
"green"),
legend=rownames(counts), beside=TRUE)
```

The first `barplot()` function produces a stacked bar plot, whereas the second produces a grouped bar plot. We will also added the `col` option to add color to the bars plotted. The `legend.text` parameter provides bar labels for the legend (which are only useful when height is a matrix).

7.11.4 Mean Bar Plots

Bar plots needn't be based on counts or frequencies. You can create bar plots that represent means, medians, standard deviations, and so forth by using the aggregate function and passing the results to the `barplot()` function.

```
states <- data.frame(state.region, state.x77)
> means <- aggregate(states$Illiteracy, by=list(state.region), FUN=mean)
> means
```

	Group.	1 x
1	Northeast	1.00

```

2           South      1.74
3           North Central 0.70
4           West       1.02
> means <- means[order(means$x),]
> means
      Group.      1 x
3      North Central  0.70
1      Northeast     1.00
4      West          1.02
2      South         1.74
> barplot(means$x, names.arg=means$Group.1)
> title("Mean Illiteracy Rate")

```

7.11.5 Spinograms

Before finishing our discussion of bar plots, let's take a look at a specialized version called a spinogram. In a spinogram, a stacked bar plot is rescaled so that the height of each bar is 1 and the segment heights represent proportions. Spinograms are created through the `spine()` function of the `vcd` package. The following code produces a simple spinogram:

```

library(vcd)
attach(Arthritis)
counts <- table(Treatment, Improved)
spine(counts, main="Spinogram Example")
detach(Arthritis)

```

7.11.6 Pie Charts

Whereas pie charts are ubiquitous in the business world, they're denigrated by most statisticians, including the authors of the R documentation. They recommend bar or dot plots over pie charts because people are able to judge length more accurately than volume. Perhaps for this reason, the pie chart options in R are limited when compared with other statistical software. Pie charts are created with the function

```
pie(x, labels)
```

Where x is a non-negative numeric vector indicating the area of each slice and **labels** provides a character vector of slice labels.

```
par(mfrow=c(2, 2))
```

```
slices <- c(10, 12, 4, 16, 8)
```

```
lbls <- c("US", "UK", "Australia", "Germany", "France")
```

```
pie(slices, labels = lbls, main="Simple Pie Chart")
```

```
pct <- round(slices/sum(slices)*100)
```

```
lbls2 <- paste(lbls, " ", pct, "%", sep="")
```

```
pie(slices, labels=lbls2, col=rainbow(length(lbls2)),
```

```
main="Pie Chart with Percentages")
```

```
library(plotrix)
```

```
pie3D(slices, labels=lbls,explode=0.1, main="3D Pie Chart ")
```

```
mytable <- table(state.region)
```

```
lbls3 <- paste(names(mytable), "\n", mytable, sep="")
```

```
pie(mytable, labels = lbls3, main="Pie Chart from a Table\n (with sample sizes)")
```

Histograms display the distribution of a continuous variable by dividing the range of scores into a specified number of bins on the x-axis and displaying the frequency of scores in each bin on the y-axis. You can create histograms with the function

```
hist(x)
```

Where x is a numeric vector of values. The option `freq=FALSE` creates a plot based on probability densities rather than frequencies. The `breaks` option controls the number of bins. The default produces equally spaced breaks when defining the cells of the histogram.

```
par(mfrow=c(2,2))
```

```
hist(mtcars$mpg)
```

```
hist(mtcars$mpg,
```

```
breaks=12,
```

```
col="red",
```

```
xlab="Miles Per Gallon",
```

```
main="Colored histogram with 12 bins")
```



```

hist(mtcars$mpg,
freq=FALSE,
breaks=12,
col="red",
xlab="Miles Per Gallon", main="Histogram, rug plot, density curve")
rug(jitter(mtcars$mpg))
lines(density(mtcars$mpg), col="blue", lwd=2)
x <- mtcars$mpg
h<-hist(x,
breaks=12,
col="red",
xlab="Miles Per Gallon",
main="Histogram with normal curve and box")
xfit<-seq(min(x), max(x), length=40)
yfit<-dnorm(xfit, mean=mean(x), sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=2)
bar()

```

7.11.7 Kernel Density Plots

.kernel density estimation is a nonparametric method for estimating the probability density function of a random variable. Although the mathematics are beyond the scope of this text, in general, kernel density plots can be an effective way to view the distribution of a continuous variable. The format for a density plot (that's not being superimposed on another graph) is `plot(density(x))`

Where x is a numeric vector. Because the `plot()` function begins a new graph, use the `lines()` function (when superimposing a density curve on an existing graph).

```

par(mfrow=c(2,1))
d <- density(mtcars$mpg)
plot(d)

```

```
d <- density(mtcars$mpg)
plot(d, main="Kernel Density of Miles Per Gallon")
polygon(d, col="red", border="blue")
rug(mtcars$mpg, col="brown")
```

The `polygon()` function draws a polygon whose vertices are given by `x` and `y`.

These values are provided by the `density()` function in this case.

Kernel density plots can be used to compare groups.

7.11.8 Box Plots

A box-and-whiskers plot describes the distribution of a continuous variable by plotting its five-number summary: the minimum, lower quartile (25th percentile), median (50th percentile), upper quartile (75th percentile), and maximum. It can also display observations that may be outliers (values outside the range of $\pm 1.5 \cdot \text{IQR}$, where IQR is the inter quartile range defined as the upper quartile minus the lower quartile).

7.11.9 Violin Plots

Before we end our discussion of box plots, it's worth examining a variation called a *violin plot*. A violin plot is a combination of a box plot and a kernel density plot. You can create one using the `vioplot()` function from the `vioplot` package. Be sure to install the `vioplot` package before first use.

The format for the `vioplot()` function is `vioplot(x1, x2, ... , names=, col=)` Where `x1`, `x2`, ... represent one or more numeric vectors to be plotted (one violin plot is produced for each vector). The `names` parameter provides a character vector of labels for the violin plots, and `col` is a vector specifying the colors for each violin plot.

An example is given in the following listing.

```
library(vioplot)
x1 <- mtcars$mpg[mtcars$cyl==4]
x2 <- mtcars$mpg[mtcars$cyl==6]
x3 <- mtcars$mpg[mtcars$cyl==8]
vioplot(x1, x2, x3,
```

```
names= c("4 cyl", "6 cyl", "8 cyl"), col="gold")
```

```
title("Violin Plots of Miles Per Gallon", ylab="Miles Per Gallon", xlab="Number of Cylinders")
```

Dot plots provide a method of plotting a large number of labeled values on a simple horizontal scale. You create them with the `dotchart()` function, using the format

```
dotchart(x, labels=)
```

where x is a numeric vector and $labels$ specifies a vector that labels each point. You can add a `groups` option to designate a factor specifying how the elements of x are grouped. If so, the option `gcolor` controls the color of the groups label, and `cex` controls the size of the labels. Here's an example with the `mtcars` dataset: `dotchart(mtcars$mpg, labels=row.names(mtcars), cex=.7, main="Gas Mileage for Car Models", xlab="Miles Per Gallon")`

Remark: We discuss only graphs of time series and frequency distribution. Different functions such as `plot()`, `abline()`, `curve()`, `text()`, `outer()`, `persp()` and many others are discussed in this session to create different plots/graphs.

7.12 Summary

This unit reviewed the `ggplot2` package, which provides advanced graphical methods based on a comprehensive grammar of graphics. The package is designed to provide you with a complete and comprehensive alternative to the native graphics provided with R. It offers methods for creating attractive and meaningful visualizations of data that are difficult to generate in other ways. The `ggplot2` package can be difficult to learn, but a wealth of material is available to help you on your journey (I promised myself that I would never use that word, but learning `ggplot2` can certainly feel like one). A list of all `ggplot2` functions, along with examples, can be found at <http://docs.ggplot2.org>. To learn about the theory underlying `ggplot2`, see the original book by Wickham (2009). Finally, Chang (2013) has written a very practical book, chock full of useful examples. Chang's book is definitely where I would start.

You should now have a firm grasp of the many ways that R allows you to create visual representations of data. If a picture is worth a thousand words, and R provides a thousand ways to create a picture, then R must be worth a million words (or something to that effect).

7.13 Self-Assessment Question

Problem 1: Find the quartiles of eruption durations in the data set faithful

We apply the quantile function to compute the quartiles of eruptions.

```
> duration = faithful$eruptions
```

```
> quantile(duration)
```

0%	25%	50%	75%	100%
1.6000	2.1627	4.0000	4.4543	5.1000

Problem 2: Find the 32nd, 57th and 98th percentiles of eruption durations in the data set faithful.

We apply the quantile function to compute the percentiles of eruptions with the desired percentage ratios.

```
> duration = faithful$eruptions
```

```
> quantile(duration, c(.32, .57, .98))
```

32%	57%	98%
2.3952	4.1330	4.9330

7.14 References

- <http://docs.ggplot2.org>
- nptl.ac.in
- swayam.gov.in
- r-tutor.com.

UNIT: 8 TESTING OF HYPOTHESIS WITH R

Structure

3.0 Introduction

3.1 Objectives

3.2 Basic of Hypothesis

 3.2.1 Simple and Composite Hypotheses

 3.2.2 Null and Alternative Hypotheses

 3.2.3 Critical Region

 3.2.4 Type-I and Type-II Errors

 3.2.5 Level of Significance

 3.2.6 One-Tailed and Two-Tailed Test

3.3 General Procedure of Testing a Hypothesis

3.4 Statistical Analysis using R

3.5 Descriptive Statistics

3.6 Fitting and interpreting ANOVA type models

 3.6.1 One-way ANOVA

 3.6.2 Multiple comparisons

 3.6.3 Two-way factorial ANOVA

3.7 Using R to model basic experimental designs

 3.7.1 Completely Randomized Design

 3.7.2 Randomized Block Design

3.7 Population Mean between Two Matched Samples

3.9 Population Mean between Two Independent

3.10 Summary

3.11 Self-Assessment Questions

 3.12References

8.1 Introduction

In this unit, we have discussed one part of statistical inference, that is, estimation and we have learnt how we estimate the unknown population parameter(s) by using point estimation and interval estimation. We will focus on the second part of statistical inference which is known as testing of hypothesis. Whenever you are trying to estimate the value of a parameter, we try to take a sample and we observe that the different samples give different values of the estimates and the value of the Estimates may or may not be the same as of the unknown value of the parameter in the Population. And the difference in the estimated values which are obtained from different samples May be due to some random variation. In simple words, we just want to take the closeness of estimated value with some hypothetical value and then what are we trying to do? We are trying to find the difference between the estimated value and the hypothetical value and if this difference is small, this is less then we can expect to accept it and would say that there is not much significant difference between the two values. Example, In our day-to-day life, we see different commercials advertisements in television, newspapers, magazines, etc. such as

- (i) The television of certain brand saves up to 20% electric bill.
- (ii) The car of certain brand gives 60 km/litter mileage.
- (iii) A detergent of certain brand produces the cleanest wash, etc.

The Procedure of testing such type of claims or statements or assumptions is known as testing of hypothesis. The truth or falsity of a claim or statement is never known unless we examine the entire population. But practically it is not possible in mostly situations so we take a random sample from the population under study and use the information contained in this sample to take the decision whether a claim is true or false.

A customer of car wants to test whether the claim of car of certain brand gives the average mileage 60 km/liter is true or false. The decision maker is interested in making inference about the population parameter(s). However, he/she is not interested in estimating the value of parameter(s) but he/she is interested in testing a claim or statement or assumption about the value of population parameter(s). Such claim or statement is postulated in terms of hypothesis. In statistics, a hypothesis is a statement or a claim or an assumption about the value of a population parameter (e.g., mean, median, variance, proportion, etc.). Similarly, in case of two or more

populations a hypothesis is comparative statement or a claim or an assumption about the values of population parameters. (e.g., means of two populations are equal, variance of one population is greater than other, etc.). The plural of hypothesis is hypotheses.

8.2 Objectives

After studying this unit, you will be able to understand the following objectives:

- Studying of Statistical Analysis using R.
- Studying of a hypothesis
- Studying of the null and alternative hypotheses.
- Studying of type-I and type-II errors
- Studying of critical region and level of significance
- Studying of one-tailed and two-tailed tests
- Studying of general procedure of testing a hypothesis
- Studying of Graphs

8.3 Basic of Hypothesis

In hypothesis testing problems first of all we should be identifying the claim or statement or assumption or hypothesis to be tested and write it in the words. Once the claim has been identified then we write it in symbolical form if possible.

8.3.1 Simple and Composite Hypotheses

In general sense, if a hypothesis specifies only one value or exact value of the population parameter then it is known as simple hypothesis. And if a hypothesis specifies not just one value but a range of values that the population parameter may assume is called a composite hypothesis. Examples, the hypothesis postulated in (i) $\mu = 40$ is simple hypothesis because it gives a single value of parameter ($\mu = 40$) whereas the hypothesis postulated in (ii) $\mu > 40$ is composite hypothesis because it does not specify the exact average value. It may be 60, 50, 100 or any other.

8.3.2 Null and Alternative Hypotheses

As we have discussed in hypothesis testing problems first of all we identify the claim or statement to be tested and write it in symbolical form. After that we write the complement or opposite of the claim or statement in symbolical form. In example, the claim is $\mu = 40$ then its complement is $\mu \neq 40$ km. In (ii) the claim is $\mu > 40$ then its complement is $\mu \leq 40$. If the claim is $\mu < 40$ gm then its complement is $\mu \geq 40$ gm. The claim and its complement are formed in such a way that they cover all possibility of the value of population parameter.

Once the claim and its compliment have been established then we decide of these two which is the null hypothesis and which the alternative hypothesis is. The thump rule is that the statement containing equality is the null hypothesis. That is, the hypothesis which contains symbols $=$ or \leq or \geq is taken as null hypothesis and the hypothesis which does not contain equality i.e. contains \neq or $<$ or $>$ is taken as alternative hypothesis. The null hypothesis is denoted by H_0 and alternative hypothesis is denoted by H_1 .

The hypothesis which we wish to test is called as the null hypothesis. According to Prof. R.A. Fisher, “A null hypothesis is a hypothesis which is tested for possible rejection under the assumption that it is true.” The hypothesis which complements to the null hypothesis is called alternative hypothesis.

The alternative hypothesis has two types:

- (i) Two-sided (tailed) alternative hypothesis.
- (ii) One-sided (tailed) alternative hypothesis.

8.3.3 Critical Region

In order to test a hypothesis, the entire sample space is partitioned into two disjoint sub-spaces, say, u and $S-u$. If calculated value of the test statistic lies in u , then we reject the null hypothesis and if it lies in $S-w$, then we do not reject the null hypothesis. The region is called a “rejection region or critical region” and the region is called a “non-rejection region”. Therefore, we can say that “A region in the sample space in which if the calculated value of the test statistic lies, we reject the null hypothesis then it is called critical region or rejection region.”

8.3.4 Type-I and Type-II Errors

A test statistic is calculated on the basis of observed sample observations. But a sample is a small part of the population about which decision is to be taken. A random sample may or may not be a good representative of the population. A faulty sample misleads the inference (or conclusion) relating to the null hypothesis. For example, an engineer infers that a packet of screws is substandard when actually it is not. It is an error caused due to poor or inappropriate (faulty) sample. Similarly, a packet of screws may infer good when actually it is sub-standard. So we can commit two kinds of errors while testing a hypothesis which are summarized in Table 8.1 which is given below

Table 8.1

Decision	H ₀ True	H ₁ True
Reject H ₀	Type-I Error	Correct Decision
Do not Reject H ₀	Correct Decision	Type-II Error

8.3.5 Level of Significance

We have discussed the hypothesis, types of hypotheses, critical region and types of errors. In this part, we shall discuss very useful concept “level of significance”, which play an important role in decision making while testing a hypothesis.

The probability of type-I error is known as level of significance of a test. It is also called the size of the test or size of critical region, denoted by α .

If calculated value of the test statistic lies in rejection (critical) region, then we reject the null hypothesis and if it lies in non-rejection region, then we do not reject the null hypothesis. Also we note that when H₀ is rejected then automatically the alternative hypothesis H₁ is accepted. Now, one point of our discussion is that how to decide critical value(s) or cut-off value(s) for a known test statistic.

8.3.6 One-Tailed and Two-Tailed Tests

We have seen that rejection (critical) region lies at one-tail or two-tails on the probability curve of sampling distribution of the test statistic its depend upon the form of alternative hypothesis. Similarly, the test of testing the null hypothesis also depends on the alternative hypothesis. A test of testing the null hypothesis is said to be two-tailed test if the alternative hypothesis is two-tailed whereas if the alternative hypothesis is one-tailed then a test of testing the null hypothesis is said to be one-tailed test. For example, if our null and alternative hypothesis are

$$H_0: \theta = \theta_0 \text{ and } H_1: \theta \neq \theta_0$$

Then the test for testing the null hypothesis is two-tailed test because the alternative hypothesis is two-tailed that means, the parameter θ can take value greater than θ_0 or less than θ_0 .

If the null and alternative hypotheses are $H_0: \theta \leq \theta_0$ and $H_1: \theta > \theta_0$

Then the test for testing the null hypothesis is right-tailed test because the alternative hypothesis is right-tailed. Similarly, if the null and alternative hypotheses are

$$H_0: \theta \geq \theta_0 \text{ and } H_1: \theta < \theta_0$$

Then the test for testing the null hypothesis is left-tailed test because the alternative hypothesis is left-tailed.

8.4 General Procedure of Testing a Hypothesis

Testing of hypothesis is a huge demanded statistical tool by many disciplines and professionals. It is a step-by-step procedure as you will see in next three units through a large number of examples. The aim of this section is just given you flavor of that sequence which involves following steps:

First of all, we have to setup null hypothesis H_0 and alternative hypothesis H_1 . Suppose, we want to test the hypothetical / claimed / assumed value θ_0 of parameter θ . So, we can take the null and alternative hypotheses as

$$H_0: \theta = \theta_0 \text{ and } H_1: \theta \neq \theta_0 \quad \{\text{for two tail test}\}$$

Or

$H_0: \theta \leq \theta_0$ and $H_1: \theta > \theta_0$ {for one tail test}

$H_0: \theta \geq \theta_0$ and $H_1: \theta < \theta_0$

In case of comparing same parameter of two populations of interest, say, θ_1 and θ_2 , then our null and alternative hypotheses would be

$H_0: \theta_1 = \theta_2$ and $H_1: \theta_1 \neq \theta_2$ {for two tail test}

Or

$H_0: \theta_1 \leq \theta_2$ and $H_1: \theta_1 > \theta_2$ {for one tail test}

$H_0: \theta_1 \geq \theta_2$ and $H_1: \theta_1 < \theta_2$

After setting the null and alternative hypotheses, we establish criteria for rejection or non-rejection of null hypothesis, that is, decide the level of significance (α), at which we want to test our hypothesis. Generally, it is taken as 5% or 1% ($\alpha = 0.05$ or 0.01).

We choose an appropriate test statistic under H_0 for testing the null hypothesis. After that, specify the sampling distribution of the test statistic preferably in the standard form like Z (standard normal), χ^2 , t, F or any other well-known.

Calculate the value of the test statistic on the basis of observed sample observations.

Obtain the critical (or cut-off) value(s) in the sampling distribution of the test statistic and construct rejection (critical) region of size α . Generally, critical values for various levels of significance are putted in the form of a table for various standard sampling distributions of test statistic such as Z-table, χ^2 -table, t-table, etc.

After that, compare the calculated value of test statistic obtained, with the critical value(s) obtained and locates the position of the calculated test statistic, that is, it lies in rejection region or non-rejection region.

In testing of hypothesis ultimately, we have to reach at a conclusion. It is done as explained below

- (i) If calculated value of test statistic lies in rejection region at α level of significance then we reject null hypothesis. It means that the sample data provide us sufficient

evidence against the null hypothesis and there is a significant difference between hypothesized value and observed value of the parameter.

- (ii) If calculated value of test statistic lies in non-rejection region at α level of significance, then we do not reject null hypothesis. It means that the sample data fails to provide us sufficient evidence against the null hypothesis and the difference between hypothesized value and observed value of the parameter due to fluctuation of sample.

Example: Suppose the manufacturer claims that the mean lifetime of a light bulb is more than 10,00 hours. For a sample of 20 light bulbs, the mean lifetime turns out to be only 9,900 hours. Assume the population standard deviation to be 12 hours. At 5% level of significance level, can we reject the claim by the manufacturer?

The null hypothesis is that $\mu \geq 10000$.

We begin with computing the test statistic.

```
> xbar = 9900          # sample mean
> mu0 = 10000         # hypothesis
> sigma = 12          # population sd
> n = 20              # sample size
> z = (xbar-mu0)/(sigma/sqrt(n))
> z                  # test statistic
[1] -3.72678
```

We then compute the critical value at 5% level of significance.

```
> alpha = .05
> z.alpha = qnorm(1-alpha)
> -z.alpha          # critical value
[1] -1.6449
```

The test statistic -3.72678 is less than the critical value of -1.6449. Hence, at .05 significance level, we reject the claim that the mean lifetime of the light bulbs is above 10,000 hours

Example 2: Suppose the food label on a cookie bag states that there are at most 2 grams of saturated fat in a single cookie. In a sample of 35 cookies, it is found that there are 2.1 grams of

saturated fat per cookie on average. Assume the population standard deviation to be 0.25 grams. At .05 significance of level, can we reject the claim on food label?

The null hypothesis is that $\mu \leq 2$.

We begin with computing the test statistic.

```
> xbar = 2.1 # sample mean
> mu0 = 2 # hypothesis
> sigma = 0.25 # population sd
> n = 35 # sample size
> z = (xbar-mu0)/(sigma/sqrt(n))
> z # test statistic
```

```
[1] 2.3664
```

We then compute the critical value at .05 significance of level.

```
> alpha = .05
> z.alpha = qnorm(1-alpha)
> z.alpha # critical value
```

```
[1] 1.6449
```

Conclusion: The test statistic 2.3664 is greater than the critical value of 1.6449. Hence, at .05 Significance of level, we reject the claim that there are at most 2 grams of saturated fat in a cookie.

8.5 Statistical Analysis using R

R is a reliable programming language for Statistical Analysis. It has a wide range of statistical library support like T-test, linear regression, logistic regression, and time-series data analysis. R comes with very good data visualization features supporting potting and graphs using graphical packages like ggplot2.

8.6 Descriptive Statistics

In this section, we will look at measures of central tendency, variability, and distribution shape for continuous variables. For illustrative purposes, we'll use several of the variables from the Motor Trend Car Road Tests (mtcars) dataset we will focus will be on miles per gallon (mpg), horsepower (hp), and weight (wt):

```
> myvars <- c("mpg", "hp", "wt")
> head(mtcars[myvars])
mpg hp wt
Mazda RX4 21.0 110 2.62
Mazda RX4 Wag 21.0 110 2.88
Datsun 710 22.8 93 2.32
Hornet 4 Drive 21.4 110 3.21
Hornet Sportabout 18.7 175 3.44
Valiant 18.1 105 3.46
```

First, we'll look at descriptive statistics for all 32 cars. Then we'll examine descriptive statistics by transmission type (am) and number of cylinders (cyl). Transmission type is a dichotomous variable coded 0=automatic, 1=manual, and the number of cylinder can be 4, 5, or 6.

```
> myvars <- c("mpg", "hp", "wt")
> summary(mtcars[myvars])
mpg hp wt
Min. :10.4 Min. : 52.0 Min. :1.51
1st Qu.:15.4 1st Qu.: 96.5 1st Qu.:2.58
Median :19.2 Median :123.0 Median :3.33
Mean :20.1 Mean :146.7 Mean :3.22
3rd Qu.:22.8 3rd Qu.:180.0 3rd Qu.:3.61
Max. :33.9 Max. :335.0 Max. :5.42
```

The summary() function provides the minimum, maximum, quartiles, and mean for numerical variables and frequencies for factors and logical vectors. You can use the apply() or sapply() function from chapter 5 to provide any descriptive statistics you choose. For the sapply() function, the format is sapply(x, FUN, options) Where x is the data frame (or matrix)

and FUN is an arbitrary function. If options are present, they're passed to FUN. Typical functions that you can plug in here are `mean()`, `sd()`, `var()`, `min()`, `max()`, `median()`, `length()`, `range()`, and `quantile()`. The function `fivenum()` returns Tukey's five-number summary (minimum, lower-hinge, median, upper-hinge, and maximum).

Surprisingly, the base installation doesn't provide functions for skew and kurtosis, but you can add your own. The example in the next listing provides several descriptive statistics, including skew and kurtosis.

```
> mystats <- function(x, na.omit=FALSE)
{
  if (na.omit)
  x <- x[!is.na(x)]
  m <- mean(x)
  n <- length(x)
  s <- sd(x)
  skew <- sum((x-m)^3/s^3)/n
  kurt <- sum((x-m)^4/s^4)/n - 3
  return(c(n=n, mean=m, stdev=s, skew=skew, kurtosis=kurt))
}
> myvars <- c("mpg", "hp", "wt")
> sapply(mtcars[myvars], mystats)
mpg hp wt
n 32.000 32.000 32.0000
mean 20.091 146.688 3.2172
stdev 6.027 68.563 0.9785
skew 0.611 0.726 0.4231
kurtosis -0.373 -0.136 -0.0227
```

In regression models for predicting a quantitative response variable from quantitative predictor variables, but there's no reason that we couldn't have included nominal or ordinal factors as predictors as well. When factors are included as explanatory variables, our focus usually shifts from prediction to understanding group differences, and the methodology is

referred to as analysis of variance (ANOVA). ANOVA methodology is used to analyze a wide variety of experimental and quasi-experimental designs. This chapter provides an overview of R functions for analyzing common research designs. First we'll look at design terminology, followed by a general discussion of R's approach to fitting ANOVA models. Then we'll explore several examples that illustrate the analysis of common designs. Along the way, you'll treat anxiety disorders; lower blood cholesterol levels, help pregnant mice have fat babies, assure that pigs grow long in the tooth, facilitate breathing in plants, and learn which grocery shelves to avoid. In addition to the base installation, you'll be using the `car`, `gplots`, `HH`, `rrcov`, `multicomp`, `effects`, `MASS`, and `mv` outlier packages in the examples. Be sure to install them before trying out the sample code.

Often, in practice, one is required to compare more than two population means. In this unit, we shall study a statistical procedure, called analysis of variance, which allows one to test a hypothesis comparing several normal population means.

The problem of comparing several populations means arises quite naturally in practice. For instance, on the basis of sample data, one might wish to decide whether there is any real difference between three teaching methods of a foreign language. Quite often in agriculture, an experimenter is interested in comparing the yielding abilities of several varieties of a crop, say wheat. Similarly, there may be four different drugs for the control of blood pressure and it is of interest to know whether these four drugs are equally efficient in the control of blood pressure.

In each of the above examples, we have several populations (one for each method of teaching, each crop variety, and each drug) and the hypothesis that is required to be tested is whether the means of these populations are equal. Analysis of variance, written in short as ANOVA is useful in such situations. We shall assume that each of the populations have a normal distribution with possibly different means but having the same variance.

We shall start this unit by acquainting you with real life problem in which you need to compare means of several populations simultaneously. We shall introduce the method of analysis of variance through this example. In this unit we shall confine our attention to study the effect of a single factor on a variable under study. Such study leads to one-way classification data. We shall then formulate the model for one-way classification for the problem considered and use this

model to explain the procedure to carry out test of hypothesis in the give situation and draw the conclusion.

8.7 Fitting and Interpreting ANOVA Type Models

Some followings are:

8.7.1 One-way ANOVA

In a one-way ANOVA, you're interested in comparing the dependent variable means of two or more groups defined by a categorical grouping factor. This example comes from the `cholesterol` dataset in the `multcomp` package, taken from Westfall, Tobia, Rom, & Hochberg (1999). Fifty patients received one of five cholesterol-reducing drug regimens (`trt`). Three of the treatment conditions involved the same drug administered as 20 mg once per day (`1time`), 10mg twice per day (`2times`), or 5 mg four times per day (`4times`). The two remaining conditions (`drugD` and `drugE`) represented competing drugs. Which drug regimen produced the greatest cholesterol reduction (`response`)? The analysis is provided in the following listing.

```
> library(multcomp)
> attach(cholesterol)
> table(trt)
trt
1time 2times 4times drugD drugE
10 10 10 10 10
> aggregate(response, by=list(trt), FUN=mean)
Group.1 x
1 1time 5.78
2 2times 9.22
3 4times 12.37
4 drugD 15.36
5 drugE 20.95
> aggregate(response, by=list(trt), FUN=sd)
```

```

Group.1 x
1 1time 2.88
2 2times 3.48
3 4times 2.92
4 drugD 3.45
5 drugE 3.35
> fit <- aov(response ~ trt)
> summary(fit)
Df Sum Sq Mean Sq F value Pr(>F)
trt 4 1351 338 32.4 9.8e-13 ***
Residuals 45 469 10
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> library(gplots)
> plotmeans(response ~ trt, xlab="Treatment", ylab="Response",
main="Mean Plot\nwith 95% CI")
> detach(cholesterol)

```

Looking at the output, you can see that 10 patients received each of the drug regimens b. From the means, it appears that drug E produced the greatest cholesterol reduction; whereas 1time produced the least c. Standard deviations were relatively constant across the five groups, ranging from 2.88 to 3.48 d. The ANOVA F test for treatment (trt) is significant ($p < .0001$), providing evidence that the five treatments are not all equally effective e. The plotmeans() function in the gplots package can be used to produce a graph of group means and their confidence intervals f. A plot of the treatment means, with 95% confidence limits, is provided in figure 9.1 and allows you to clearly see these treatment differences.

8.7.2 Multiple Comparisons

The ANOVA F test for treatment tells you that the five drug regimens aren't equally effective, but it doesn't tell you which treatments differ from one another. You can use a multiple comparison procedure to answer this question. For example, the TukeyHSD() function provides a test of all pair wise differences between group means, as shown next.

```

> TukeyHSD(fit)
Tukey multiple comparisons of means
95% family-wise confidence level
Fit: aov(formula = response ~ trt)
$trt
diff lwr upr p adj
2times-1time 3.44 -0.658 7.54 0.138
4times-1time 6.59 2.492 10.69 0.000
drugD-1time 9.58 5.478 13.68 0.000
drugE-1time 15.17 11.064 19.27 0.000
4times-2times 3.15 -0.951 7.25 0.205
drugD-2times 6.14 2.035 10.24 0.001
drugE-2times 11.72 7.621 15.82 0.000
drugD-4times 2.99 -1.115 7.09 0.251
drugE-4times 8.57 4.471 12.67 0.000
drugE-drugD 5.59 1.485 9.69 0.003
> par(las=2)
> par(mar=c(5,8,4,2))
> plot(TukeyHSD(fit))

```

8.7.3 Two-Way Factorial ANOVA

In a two-way factorial ANOVA, subjects are assigned to groups that are formed from the cross-classification of two factors. This example uses the `ToothGrowth` dataset in the base installation to demonstrate a two-way between-groups ANOVA. Sixty guinea pigs are randomly assigned to receive one of three levels of ascorbic acid (0.5, 1, or 2 mg) and one of two delivery methods (orange juice or Vitamin C), under the restriction that each treatment combination has 10 guinea pigs. The dependent variable is tooth length. The following listing shows the code for the analysis.

```

attach(ToothGrowth)
> table(supp, dose)

```

```

dose
supp 0.5 1 2
OJ 10 10 10
VC 10 10 10
> aggregate(len, by=list(supp, dose), FUN=mean)
Group.1 Group.2 x
1 OJ 0.5 13.23
2 VC 0.5 7.98
3 OJ 1.0 22.70
4 VC 1.0 16.77
5 OJ 2.0 26.06
6 VC 2.0 26.14
> aggregate(len, by=list(supp, dose), FUN=sd)
      Group.1      Group.2 x
1 OJ  0.5      4.46
2 VC  0.5      2.75
3 OJ  1.0      3.91
4 VC  1.0      2.52
5 OJ  2.0      2.66
6 VC  2.0      4.80
> dose <- factor(dose)
> fit <- aov(len ~ supp*dose)
> summary(fit)
Df Sum Sq Mean Sq F value Pr(>F)
supp 1 205 205 15.57 0.00023 ***
dose 2 2426 1213 92.00 < 2e-16 ***
supp:dose 2 108 54 4.11 0.02186 *
Residuals 54 712 13
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> detach(ToothGrowth)

```

The table statement indicates that you have a balanced design (equal sample sizes in each cell of the design), and the `aggregate` statements provide the cell means and standard deviations. The `dose` variable is converted to a factor so that the `aov()` function will treat it as a grouping variable, rather than a numeric covariate. The ANOVA table provided by the `summary()` function indicates that both main effects (`supp` and `dose`) and the interaction between these factors are significant. We can visualize the results in several ways. You can use the `interaction.plot()` function to display the interaction in a two-way ANOVA.

8.8 Using R to Model Basic Experimental Designs

In an experimental study, various treatments are applied to test subjects and the response data is gathered for analysis. A critical tool for carrying out such analysis is the **Analysis of Variance** (ANOVA). It enables a researcher to differentiate treatment results based on easily computed statistical quantities from the treatment outcome.

The statistical process is derived from estimates of the population variance via two separate approaches. The first approach is based on the variance of the sample means, and the second one is based on the mean of the sample variances. Under the ANOVA assumptions as stated below, the ratio of the two statistical estimates follows the F distribution. Hence, we can test the null hypothesis on the equality of various response data from different treatments via estimates of critical regions.

The assumptions of ANOVA are:

The treatment responses are independent of each other.

The response data follow the normal distribution.

The variances of the response data are identical.

In the following tutorials, we demonstrate how to perform ANOVA on a few basic experimental designs.

8.8.1 Completely Randomized Design

In a **Completely Randomized Design**, there is only one primary factor under consideration in the experiment. The test subjects are assigned to treatment levels of the primary factor at random.

Example: A fast-food franchise is test marketing 3 new menu items. To find out if they the same popularity, 18 franchisee restaurants are randomly chosen for participation in the study. In accordance with the completely randomized design, 6 of the restaurants are randomly chosen to test market the first new menu item, another 6 for the second menu item, and the remaining 6 for the last menu item.

Problem: Suppose the following table represents the sales figures of the 3 new menu items in the 18 restaurants after a week of test marketing. At .05 level of significance, test whether the mean sales figures of the 3 new menu items are all equal.

	Item1	Item2	Item3
	22	52	16
	42	33	24
	44	8	19
	52	47	18
	45	43	34
	37	32	39

Solution

Step 1: Copy the sales figure above into a table file named "fastfood-1.txt" with a text editor.

Step 2: Load the file into a data frame df1 with the function read.table. Since the first line in the file contains the column names, we enable the header option.

```
> df1 = read.table(  
+ "fastfood-1.txt",  
+ header=TRUE)  
> df1
```

	Item1	Item2	Item3
1	22	52	16
2	42	33	24
3	44	8	19
4	52	47	18
5	45	43	34
6	37	32	39

Step 3: Concatenate the data rows of df1 into a single vector r. Note the use of the function t for matrix transpose.

```
> r = c(t(as.matrix(df1))) # response
```

```
> r
```

```
[1] 22 52 16 42 33 ....
```

Step 4: Assign new variables for the treatment levels and number of observations.

```
> f = c("Item1", "Item2", "Item3")
```

```
> k = 3 # number of treatments
```

```
> n = 6 # data per treatment
```

Step 5: Create a vector of treatment factors that corresponds to members of r in step 3.

```
> tm = gl(k, 1, n*k, factor(f))
```

```
> tm
```

```
[1] Item1 Item2 Item3 Item1 ...
```

Step 6: Apply the function aov to a formula that describes the response r by the treatment factor tm.

```
> av = aov(r ~ tm)
```

Step 7: Print out the ANOVA table with the function summary.

```
> summary(av)
```

```
Df Sum Sq Mean Sq F value Pr(>F)
```

```
tm 2 745 373 2.54 0.11
```

```
Residuals 15 2200 147
```

Answer: Since the p-value of 0.11 is greater than the .05 significance level, we do not reject the null hypothesis that the mean sales figures of the new menu items are all equal.

Exercise: Create the response data in step 3 above along *vertical* columns instead of horizontal rows. Adjust the factor levels in step 5 accordingly.

Solution of Exercise: We load the data into a data frame as before:

```
> df1 = read.table("fastfood-1.txt", header=TRUE)
```

Then we concatenate the column vectors by *skipping* the use of the function t.

```
> r = c(as.matrix(df1)) # response
```

```
> r
```

```
[1] 22 42 44 52 45 ...
```

```
>
```

Now we create the factor values `tm` by setting arguments in the function `gl` according to how the response data is arranged.

```
> f = c("Item1", "Item2", "Item3")
```

```
> k = 3 # treatment levels
```

```
> n = 6 # data per treatment
```

```
> tm = gl(k, n, n*k, factor(f))
```

```
> tm
```

```
[1] Item1 Item1 Item1 Item1 ...
```

Finally, we apply the function `aov` and print out the summary.

```
> av = aov(r ~ tm)
```

```
> summary(av)
```

```
Df Sum Sq Mean Sq F value Pr(>F)
```

```
tm 2 745 373 2.54 0.11
```

```
Residuals 15 2200 147
```

8.8.2 Randomized Block Design

In a randomized block design, there is only one primary factor under consideration in the experiment. Similar test subjects are grouped into blocks. Each block is tested against all treatment levels of the primary factor at random order. The intention is to eliminate possible extraneous influence by other factors.

Example: A fast-food franchise is test marketing 3 new menu items. To find out if they have the same popularity, 6 franchised restaurants are randomly chosen for participation in the study. In accordance with the randomized block design, each restaurant will be testing marketing all 3 new menu items. Furthermore, a restaurant will test market only one menu item per week, and it takes 3 weeks to test market all menu items. The testing order of the menu items for each restaurant is randomly assigned as well.

Problem: Suppose each row in the following table represents the sales figures of the 3 new menus in a restaurant after a week of test marketing. At .05 level of significance, test whether the mean sales figures of the 3 new menu items are all equal.

Item1	Item2	Item3
31	27	24
31	28	31
45	29	46
21	18	48
42	36	46
32	17	40

Solution

Step 1: Copy the sales figure above into a table file named "fastfood-2.txt" with a text editor.

Step 2: Load the file into a data frame df2 with the function read.table. Since the first line in the file contains the column names, we enable the header option.

```
> df2 = read.table(  
+ "fastfood-2.txt",  
+ header=TRUE)  
> df2
```

	Item1	Item2	Item3
1	31	27	24
2	31	28	31
3	45	29	46
4	21	18	48
5	42	36	46
6	32	17	40

Step 3: Concatenate the data rows of df2 into a single vector r. Note the use of the function t for matrix transpose.

```
> r = c(t(as.matrix(df2)))  
> r  
[1] 31 27 24 31 28 ...
```

Step 4: Assign new variables for the treatment levels and number of control blocks.

```
> f = c("Item1", "Item2", "Item3")
> k = 3 # treatment levels
> n = 6 # data per treatment
```

Step 5: Create a vector of treatment factors that corresponds to members of r in step 3.

```
> tm = gl(k, 1, n*k, factor(f))
> tm
[1] Item1 Item2 Item3 Item1 ...
```

Step 6: Similarly, create a vector of blocking factors for members of the response data r.

```
> blk = gl(n, k, k*n)
> blk
[1] 1 1 1 2 2 2 3 3 3 4 4 4 ...
Levels: 1 2 3 4 5 6
```

Step 7: Apply the function aov to a formula that describes the response r by both the treatment factor tm and the block control blk.

```
> av = aov(r ~ tm + blk)
```

Step 8: Print out the ANOVA table with the function summary.

```
> summary(av)
Df Sum Sq Mean Sq F value Pr(>F)
tm 2 539 269 4.96 0.032 *
blk 5 560 112 2.06 0.155
Residuals 10 543 54
```

Answer

Since the p-value of 0.032 is less than the .05 significance level, we reject the null hypothesis that the mean sales figures of the new menu items are all equal.

8.8.3 Factorial Design

In a factorial design, there are more than one factors under consideration in the experiment. The test subjects are assigned to treatment levels of every factor combinations at random.

Example: A fast food franchise is test marketing 3 new menu items in both East and West Coasts of continental United States. To find out if they have the same popularity, 12 franchisee restaurants from each coast are randomly chosen for participation in the study. In accordance with the factorial design, within the 12 restaurants from East Coast, 4 are randomly chosen to test market the first new menu item, another 4 for the second menu item, and the remaining 4 for the last menu item. The 12 restaurants from the West Coast are arranged likewise.

Problem: Suppose the following tables represent the sales figures of the 3 new menu items after a week of test marketing. Each row in the upper table represents the sales figures of 3 different East Coast restaurants. The lower half represents West Coast restaurants. At .05 level of significance, test whether the mean sales figures of the new menu items are all equal. Decide also whether the mean sales figures of the two coastal regions differ.

East Coast:

=====

	Item1	Item2	Item3
E1	25	39	36
E2	36	42	24
E3	31	39	28
E4	26	35	29

West Coast:

=====

	Item1	Item2	Item3
W1	51	43	42
W2	47	39	36
W3	47	53	32
W4	52	46	33

Solution

Step 1: Save the sales figure into a file named "fastfood-3.csv" in CSV format as follows:

	Item1,	Item2,	Item3
E1,	25,	39,	36
E2,	36,	42,	24

E3,	31,	39,	28
E4,	26,	35,	29
W1,	51,	43,	42
W2,	47,	39,	36
W3,	47,	53,	32
W4,	52,	46,	33

Step 2: Load the file into a data frame df3 with the function read.csv.

```
> df3 = read.csv("fastfood-3.csv")
```

Step 3: Concatenate the data rows of df3 into a single vector r. Note the use of the function t for matrix transpose.

```
> r = c(t(as.matrix(df3)))
```

```
> r
```

```
[1] 25 39 36 36 42 ...
```

Step 4: Assign new variables for the treatment levels and number of observations.

```
> f1 = c("Item1", "Item2", "Item3")
```

```
> f2 = c("East", "West")
```

```
> k1 = length(f1)
```

```
> k2 = length(f2)
```

```
> n = 4 # data per treatment
```

Step 5: Create a vector that corresponds to the first treatment level of the response data r in step 3.

```
> tm1 = gl(k1, 1, n*k1*k2,
```

```
+ factor(f1))
```

```
> tm1
```

```
[1] Item1 Item2 Item3 Item1 ...
```

Step 6: Similarly, create a vector that corresponds to the second treatment level of the response data r in step 3.

```
> tm2 = gl(k2, n*k1, n*k1*k2, factor(f2))
```

```
> tm2
```

```
[1] East East East East East ...
```

Step 7: Apply the function `aov` to a formula that describes the response `r` by the two treatment factors `tm1` and `tm2` with interaction.

```
> av = aov(r ~ tm1 * tm2)
```

Step 8: Print out the ANOVA table with the function `summary`.

```
> summary(av)
```

```
Df Sum Sq Mean Sq F value Pr(>F)
```

```
tm1 2 385 193 9.55 0.0015 **
```

```
tm2 1 715 715 35.48 1.2e-05 ***
```

```
tm1:tm2 2 234 117 5.81 0.0113 *
```

Answer

Since the p-value of 0.0015 for the menu items is less than the .05 significance level, we reject the null hypothesis that the mean sales figures of the new menu items are all equal. Moreover, the p-value of 1.2e-05 for the east-west coasts comparison is also less than the .05 significance level. It shows there is a difference in overall sales figures between the coasts. Finally, the last p-value of 0.0113 (< 0.05) indicates that there is a possible interaction between the menu item and coast location factors, *i.e.*, customers from different coastal regions probably have different tastes.

8.9 Population Mean between Two Matched Samples

Two data samples are said to be matched if they come from repeated observations of the same subject. Here, we assume that the data populations follow the normal distribution. Using the paired t-test, we can obtain an interval estimate of the difference of population means. In the built-in data set named `immer`, we can find barley yields of six selected locations in each year of 1931 and 1932. The yield data are presented in the data frame columns `Y1` and `Y2`.

```
library(MASS)
```

```
> head(immer)
```

	Loc	Var	Y1	Y2
1	UF	M	81.0	80.7
2	UF	S	105.4	82.3

Assuming that the data in immer follows the normal distribution, find the 95% confidence interval estimate of the difference between the mean barley yield in years 1931 and 1932.

We apply the function `t.test` to compute the difference in means of the matched samples. As it is a paired test, we set the argument `paired` as `TRUE`.

```
> t.test(immer$Y1, immer$Y2, paired=TRUE)
```

Paired t-test

data: immer\$Y1 and immer\$Y2

t = 3.324, df = 29, p-value = 0.002413

alternative hypothesis: true difference in means is not equal to 95 percent confidence interval:

6.122 25.705

sample estimates: mean of the differences **15.913**

8.10 Population Mean between Two Independent Samples

Two data samples are called independent if they come from unrelated populations and the samples do not affect each other. Here, we assume that the data populations follow the normal distribution. Using the unpaired t-test, we can obtain an interval estimate of the difference between two population means.

Example: In the data set `mtcars`, the data frame column `mpg` contains gas mileage statistics of various 1974 U.S. automobiles

```
mtcars$mpg
```

```
[1] 21.0 21.0 22.8 21.4 18.7 ...
```

Meanwhile, another data column in `mtcars`, named `am`, indicates the transmission type of the automobile model (0 = automatic, 1 = manual).

```
> mtcars$am
```

```
[1] 1 1 1 0 0 0 0 ...
```

We can regard the gas mileages for manual and automatic transmissions as two independent data populations

Problem: Assuming that the data in `mtcars` follows the normal distribution, find the 95% confidence interval estimate of the difference between the mean gas mileages of manual and automatic transmissions.

As discussed in our previous tutorial on data frame row slice, the gas mileage for automatic transmission can be expressed as follows

```
> L = mtcars$am == 0
> mpg.auto = mtcars[L,]$mpg
> mpg.auto
[1] 21.4 18.7 18.1 14.3 24.4 ...
```

By applying the negation of the vector L, we can find the gas mileage for manual transmission.

```
> mpg.manual = mtcars[!L,]$mpg
> mpg.manual
[1] 21.0 21.0 22.8 32.4 30.4 ...
```

We can now apply the function t-test to compute the difference in means of the two-sample data.

```
> t.test(mpg.auto, mpg.manual)
Welch Two Sample t-test
data: mpg.auto and mpg.manual
t = -3.7671, df = 18.332, p-value = 0.001374
Alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-11.2802      -3.2097
Sample estimates:
Mean of x      mean of y
17.147         24.392
```

Example: Assuming that the data in quine follows the normal distribution, find the 95% confidence interval estimate of the difference between the female proportion of Aboriginal students and the female proportion of non-Aboriginal students, each within their own ethnic group.

We apply the function prop.test to compute the difference in female proportions. The Yates's continuity correction is disabled for pedagogical reasons.

```
> prop.test(table(quine$Eth, quine$Sex), correct=FALSE)
2-sample test for equality of proportions without continuity correction
data: table(quine$Eth, quine$Sex)
```

X-squared = 0.0041, df = 1, p-value = 0.949

Alternative hypothesis: two. sided

95 percent confidence interval: -0.15642 0.16696

Sample estimates: prop 1 prop 2

0.55072 0.54545

8.11 Summary

This unit reviewed the Basic of Hypothesis which provides advanced methods based on Statistical Analysis using R. The package is designed to provide you with a complete and comprehensive alternative to the native Statistical Analysis using R. It offers methods for creating Descriptive Statistics that are difficult to generate in other ways. You should now have a General Procedure of Testing a Hypothesis using R software. If a picture is worth a thousand words and R provides a thousand Fitting and interpreting ANOVA model.

8.12 Self-Assessment Question

1. Suppose the following table represents the sales figures of the 3 new menu items in the 18 restaurants after a week of test marketing at .05 level of significance test whether the mean sales figures of the 3 new menu items are all equal.

Item1	Item2	Item3
22	52	16
42	33	24
44	8	19
52	47	18
45	43	34
37	32	39

2. Select ten random numbers between one and three

8.13 References

- Introduction to Statistics and Data Analysis with Exercises, Solutions and Applications in R. Christian Heumann, Michael Schomaker and Shalabh, Springer, (2022).
- Auguie,B.(2012).gridExtra:functions in Grid graphics. <http://CRAN.R-project.org/package=gridExtra>, R package.
- <http://docs.ggplot2.org>
- nptl.ac.in
- swayam.gov.in
- r-tutor.com.