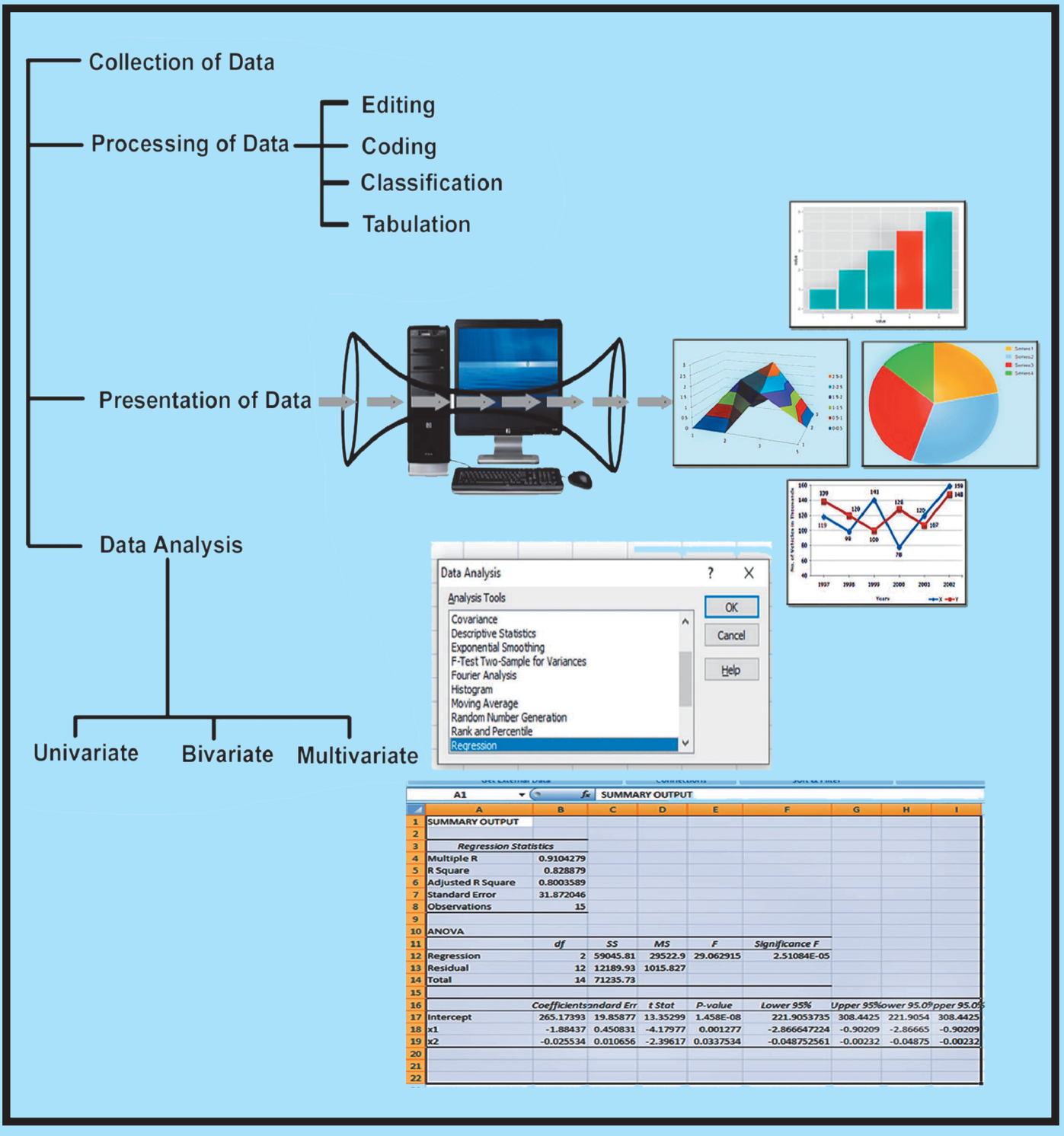
ignou
THE PEOPLE'S
UNIVERSITY

School of Social Sciences
Indira Gandhi National Open University

BECS - 184
DATA ANALYSIS



A1 SUMMARY OUTPUT

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.9104279							
5	R Square	0.828879							
6	Adjusted R Square	0.8003589							
7	Standard Error	31.872046							
8	Observations	15							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	59045.81	29522.9	29.062915	2.51084E-05			
13	Residual	12	12189.93	1015.827					
14	Total	14	71235.73						
15									
16		*Coefficients*	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	265.17393	19.85877	13.35299	1.458E-08	221.9053735	308.4425	221.9054	308.4425
18	x1	-1.88437	0.450831	-4.17977	0.001277	-2.866647224	-0.90209	-2.86665	-0.90209
19	x2	-0.025534	0.010656	-2.39617	0.0337534	-0.048752561	-0.00232	-0.04875	-0.00232
20									
21									
22									

BECS-184

Data Analysis

School of Social Sciences
Indira Gandhi National Open University

EXPERT COMMITTEE

Prof. Atul Sarma,
Former Director,
Indian Statistical Institute, New Delhi & Visiting Professor, Institute for Human Development, New Delhi

Prof. G.Pradhan
Rtd Professor of Economics
IGNOU, Maidan Garhi, Delhi

Dr. Manjula Singh,
Associate Professor of Economics,
St. Stephen's College (University of Delhi), Delhi

Prof. Narayan Prasad
Professor of Economics
IGNOU, New Delhi

Shri. B.S. Bagla
Associate Professor of Economics
PGDAV Collage, University of Delhi

Prof. K. Barik,
Professor of Economics,
IGNOU, New Delhi

Prof. B.S. Prakash,
Professor of Economics,
IGNOU, New Delhi

Shri Saugato Sen
Associate Professor of Economics
IGNOU, New Delhi

Course Coordinator : Prof. Narayan Prasad
Editor : Prof. G.K. Shukla

COURSE PREPARATION TEAM

Block 1	**Review of Mathematical and Statistical Concepts**	**Unit Writer**	**Unit Editor**
Unit 1	Mathematical Concepts	Prof. Manohar Lal, Rtd. Professor of Computer & Information Sciences, IGNOU, New Delhi	Shri Akshay Kumar, Associate professor, SOCIS, IGNOU, New Delhi
Unit 2	Statistical Concepts	Dr. J.C. Sharma, Rtd. Sr. Economic Adviser, Government of India, Ministry of New and Renewable Energy, New Delhi	Prof. Narayan Prasad & Shri. B.S. Bagla
Unit 3	Introduction to Statistical Software	Shri Akshay Kumar, Associate Professor, SOCIS, IGNOU, New Delhi	
Block 2	**Data Collection and Presentation of Data**		
Unit 4	*Data Collection: Methods and Sources	Shri S.S. Suryanarayan, Ex. Joint Advisor, Planning Commission, Renamed as Niti Ayog , New Delhi & Prof. Narayan Prasad, IGNOU, New Delhi	Shri. B.S. Bagla
Unit 5	Tools of Data Collection	Dr. Jay Kumar Ranjan, Assistant Professor in Psychology, BHU, Varanasi	Prof. Narayan Prasad
Unit 6	Data Presentation	Prof. H. S. Asthana, Professor of Psychology, BHU, Varanasi	Prof. Narayan Prasad and Shri Akshay Kumar
Block 3	**Analysis of Quantitative Data**		
Unit 7	Univariate Data Analysis	Dr. J.C. Sharma, Rtd. Sr. Economic Adviser, Government of India, Ministry of New and Renewable Energy, New Delhi	Prof. Narayan Prasad
Unit 8	Bivariate Data Analysis		Shri B.S. Bagla
Unit 9	Multivariate Data Analysis		Ms. Chetali Arora
Block 4	**Composite Index Numbers and Qualitative Data**		
Unit 10	Composite Index Numbers**	Dr. Sunil Kumar Mishra, Associate Fellow, Institute for Human Development New Delhi, Delhi	Prof. Narayan Prasad
Unit 11	Analysis of Qualitative Data	Dr. Charu Jain, Associate Fellow, NCAER, New Delhi	Prof. Narayan Prasad Shri B.S. Bagla

Print Production

Mr. Tilak Raj
Assistant Registrar (P), MPDD

Mr. Yashpal
Section Officer (P.), MPDD

Secretarial Assistance
Ms. Kamini Dogra, IGNOU, New Delhi

*Unit has been adopted from the course MEC-009
**Unit has been adopted from the course MEC-10

January, 2021
© Indira Gandhi National Open University, 2021
ISBN: 978-93-90773-19-0
All rights reserved. No part of this work may be reproduced in any form, by mimeography or any other means, without permission in writing from the Indira Gandhi National Open University.
Further information on the Indira Gandhi National Open University courses may be obtained from the University's Office at Maidan Garhi, New Delhi-110 068 or visit our website: http://www.ignou.ac.in
Printed and published on behalf of the Indira Gandhi National Open University, by Registrar MPDD, IGNOU, New Delhi.
Laser Typeset by : Tessa Media & Computers, C-206, A.F.E.-II, Okhla, New Delhi
Printed at : Dee Kay Printers, 5/37 A, Kirti Nagar Industrial Area, New Delhi - 110015

Course Contents

Block 1	Review of Mathematical and Statistical Concepts	Page No
UNIT 1	Mathematical Concepts	9
UNIT 2	Statistical Concepts	33
UNIT 3	Introduction to Statistical Software	61
Block 2	**Data Collection and Presentation of Data**	87
UNIT 4	Data Collection: Methods and Sources	89
UNIT 5	Tools of Data Collection	107
UNIT 6	Data Presentation	123
Block 3	**Analysis of Quantitative Data**	161
UNIT 7	Univariate Data Analysis	163
UNIT 8	Bivariate Data Analysis	193
UNIT 9	Multivariate Data Analysis	231
Block 4	**Composite Index Numbers and Qualitative Data**	251
UNIT 10	Composite Index Numbers	253
UNIT 11	Analysis of Qualitative Data	273
KEYWORDS		**294**
SOME USEFUL BOOKS		**304**

ignou THE PEOPLE'S UNIVERSITY

ignou
THE PEOPLE'S
UNIVERSITY

COURSE INTRODUCTION: DATA ANALYSIS (BECS-184)

In the present era of information society, information is one of the major driving forces for social, economic, political and cultural development of a country. Information is created, presented in various visual forms, related with other form of information and then used for economic and decision making activities.

One of the key ingredients of creating information is Data. The data is the raw form of information which is to be collected, cleaned and processed. Data, therefore, need to be analysed and converted to simpler presentable forms like tables and graphs so that it can be utilised in a more meaningful manner. The knowledge of tools and techniques of data collection, data presentation and data analysis by using the spreadsheet package is high in demand in almost all the sectors of economy. This skill enhancement course will equip you with the tools and techniques of collection, presentation and analysis of data through using the spreadsheet package (Excel) which, in turn, will enhance your employability. The course comprises of four blocks.

Block 1 Review of Mathematical and Statistical Concepts: A data analyst is expected to know the various mathematical and statistical concepts used frequently in the process of data analysis. Further various statistical softwares are used for the purpose of data analysis. Hence, this block covers the various mathematical and statistical concepts along with exposure to statistical software. The block has been divided into three units.

Unit 1 deals with the mathematical conceptual foundation relating to set theory, relations and functions, logic and proof techniques.

Unit 2 lays the statistical foundation by throwing light on elementary statistical concepts like statistics/ data, variable and its types, sources and types of data, measurement scale. It also covers various techniques of Uni-dimensional data analysis like percentage and measures of central tendency, measures of dispersion and measures of position.

Unit 3 enables the learner to enter and edit data in excel sheet, import and export data from one software application to another. The unit also provides means to perform simple computation and operation on data. It also enables the learners to perform different statistical analysis functions.

Block 2 on **Data Collection and Presentation of Data** deals with tools and methods of data collection and presentation. The Block is divided into three units.

Unit 4 discusses the various stages involved in planning and organising census and surveys and the various official sources of secondary data.

Unit 5 contains the various tools used for collection of data in quantitative and qualitative research.

Unit 6 covers the processes of grouping data into classes and sub-classes, general rules for constructing the tables and representing the data through various types of bar diagrams, histograms, frequency polygon and frequency curves.

Block 3 on **Analysis of Quantitative Data** sets the tone to analyse the quantitative data. It comprises of three Units.

Unit 7 throws light on inferential analysis. For this type of analysis, the unit discusses the characteristics of normal distribution, the concept of hypothesis testing, the type of errors emerge in the process, confidence level etc.

Unit 8 deals with Bivariate data analysis using regression analysis.

Unit 9 discusses the various techniques used in multivariate analysis and also provides the guidelines to determine for application and interpretation of multivariate analysis.

Block 4 on **Composite Index Numbers and Qualitative Data** comprises of two units.

Unit 10 throws light on the process of constructing the composite index and enables the learner to interpret the results derived from composite indices.

Unit 11 deals with qualitative research design and techniques of analysis of qualitative data.

ignou
THE PEOPLE'S
UNIVERSITY

Block 1
Review of Mathematical and Statistical Concepts

ignou
THE PEOPLE'S
UNIVERSITY

UNIT 1 MATHEMATICAL CONCEPTS

Structure

1.0 Objectives

1.1 Introduction

1.2 Set Theory

- 1.2.1 Concept and Representation of Set
- 1.2.2 Set Relations
- 1.2.3 Special Sets: Definitions and Properties
- 1.2.4 The Intervals as Sets
- 1.2.5 Number Sets (with Standard Notations)
- 1.2.6 Set Operations

1.3 Relation and Functions

1.4 Logic

1.5 Proof Techniques

1.6 Let Us Sum Up

1.7 References

1.8 Answers or Hints to Check your Progress Exercises

1.0 OBJECTIVES

After studying this unit, you should be able to:

- explain the concepts of set, set notations, set relations and set operations;
- define the concepts of relation, function and their properties;
- describe different number sets, relations between them, and their properties;
- appreciate the use of concepts of symbolic Logic, including logical connectors; and
- state the concept of Proof Techniques.

1.1 INTRODUCTION

We have been using mathematics in the form of numbers to count our possessions and for distributing, adding to those possessions since times immemorial. Since industrial revolution our systems of production, distribution, consumption, exchange, trade have become increasingly complex. This has prompted us to use more complex mathematical tools and formulations as well.

How Mathematics is useful for solving problems of Economics?

Let us recall some basic concepts from Economics and Mathematics:

Economics:

- Choice, trade-off, incentive, exchange, information, distribution, scarcity,
- Market (product market, labour market, capital market), goods (capital goods), service,
- Correlation, causation,
- Economy or economic system.

The concepts are used in *formulating the basic economic questions*

What to produce?

How to produce it? and

For whom it should be produced?

Mathematics:

- ***Number*** (for example, 3, −12, 8.7, 2/3,√2, $\sqrt[5]{7}$, and 7 – 3i, where 3 is an natural number,− 12 is an integer, each of 8.7 and 2/3 is a rational number, each of √2 and $\sqrt[5]{7}$ is an ***irrational number.*** Each of the numbers 3, −12, 8.7, 2/3, √2, $\sqrt[5]{7}$ is also a real number. The number 7 – 3i is a complex number, but, it is not a real number. Each of the numbers quoted here is such that the number does not have the preceding properties. For example, √2 is an irrational number, but, it is not a rational number, it is not an integer, and also it is not a natural number),
- ***Arithmetic operation*** (for example, +, −, ×, ÷ , exponent/ power e. g. 5^3),
- ***Arithmetic relation*** (for example, =, ≠ , ≤ and ≥),
- ***Variable, constant, equation*** and ***solution*** *(*for example, the variable x occurs in the equation $x^2 - 16 = 0$. Also, two constants 16 and 0 occur in the equation. The equation has two solutions, viz. x = 4 and x = − 4),
- ***Set*** (for example, {dog, cat, cow, buffalo, horse, sheep, goat} is the set of names of seven domestic animals; {4, − 4} is the set of the two solutions of the equation $x^2 - 16 = 0$; and {0, 1, 2, 3, 4, 5, 6, 7, 8, 9} is the set of all decimal digits used for representing numbers in decimal system).*Note that in representing a set only'{'and '}', called* braces*, are used here.*
- ***Interval*** (for example,
  - the interval [3, 7.5] denotes the set of all real numbers between 3 and 7.5 and it also includes both the numbers 3 and 7.5. Such an interval is called a *closed* interval.
  - the interval] 3, 7.5] *(also, denoted as (3, 7.5])* denotes the set of all real numbers between 3 and 7.5 and it includes 7.5, but not 3. Such an interval is called a *semi-closed or semi-open* interval.
  - the interval [3, 7.5[*(also, denoted as [3, 7.5))* denotes the set of all real numbers between 3 and 7.5 and it also includes 3, but not 7.5. Such an interval is also called a *semi-closed or semi-open* interval.
  - the interval]3, 7.5[*(also, denoted as (3, 7.5))*denotes the set of all real numbers between 3 and 7.5 but, it does not include any of 3 and 7.5. Such an interval is called an *open* interval.)

- ***Absolute value of a number*** $|x| = \begin{cases} -x, & x < 0 \\ x, & x \geq 0 \end{cases}$

  For example, $|5| = 5$, and $|-5| = 5$

- ***Symbol, variable, constant and parameter***

In general, a ***symbol*** is a notation which *represents or stands for something else.*

A ***variable*** is a symbol which *can be measured and which may change.* We generally use last few alphabets like p, q, r, s, x, y, z to denote variables.

A ***constant*** is, in strict sense, *not a symbol, but is a sign*, in the sense that it signifies (or stands for) itself only, For example, the sign 5.67 stands for itself and not for anything else.

A ***parameter,*** say *c*, is a sort of *generalised constant* in the sense that when it is used in a mathematical expression (say, $3x^2 + 2x + c$) or in an equation (say, $3x^2 + 2x + c = 0$), then *c* is treated *as a constant (or at most, as an independent variable)* throughout an argument/ solution-process.

In mathematics, the *beginning letters a, b, c* of Latin alphabet along with some other letters are commonly used for parameters.

The (mathematical) concepts are used in *formulating* questions mathematically**,** then also in the *processes of discovering* solutions using mathematical assets, and finally in *describing* the solutions found.

Generally, the following five-step process is adopted for solving problems from Economics by using Mathematics:

i) Define the problem from the field of Economics in terms of concepts of Economics,

ii) Translate the above definition to mathematical formulation, i.e. in terms of mathematical concepts,

iii) Use mathematical assets like techniques, algorithms/procedures, laws, principles, constructions etc. to calculate/ compute and to represent the answer or to logically derive the conclusion. The final outcome is a mathematical entity: a number, an expression, a table, a graph, or a logical statement, etc.

iv) Translate the final outcome—which is a mathematical entity—of the previous step back in terms of concepts of Economics.

v) Verify whether outcome of the previous step is justified and acceptable.

The step (ii) above is called *mathematical modeling*. The quality of solution of a problem heavily depends on the quality of mathematical model formulated or used at this step.

The significance of use of Mathematics lies in the facts:

a) If the mathematical model is formulated carefully and appropriately, quality of the solution obtained through the five-step procedure is expected to be good,

b) Mathematics has rich assets/repertoire for step (iii) above to solve problems of mathematics, and

c) Mathematics has been world's most tried, tested and successful system for problem solving, at least, for problems from the fields of natural sciences and to some extent of social sciences.

1.2 SET THEORY

The concept of *set* has become so fundamental that now-a-days any introduction to mathematics is initiated with some discussion of *set theory.* A fundamental/basic concept of an academic discipline, including that of mathematics, is taken as *undefined*. The reason being, any attempt at defining a basic concept requires some sequence of concepts or words, where some of the new concepts/words again need to be defined or explained. The process becomes unending, or sometime circular.

***Set** is a* well-defined *collection of* distinct *objects*.

In respect of the above working definition, some clarifications are required. Let S be a set according to the working definition. The adjective *well-defined* for the noun *collection* requires that if we pick up anything, say x, in the world, it should be possible, at least theoretically, to tell whether x is in S or x is not in S. Let us have the two collections, named as C_1 and C_2, as explained below.

C_1 = collection of *all* students enrolled in the year 2017 for undergraduate courses in IGNOU

and

C_2 = collection of *some* students enrolled in the year 2017 for undergraduate courses in IGNOU.

C_1 is a well-defined collection, but, C_2 is *not* a well-defined collection. Hence, C_1 is a set, but, C_2 is *not* a set.

1.2.1 Concept and Representation of Set

The set representation notations are explained through the following examples or comments.

i) The Latin alphabet has five vowels. Let V denote the *set* of vowels: *a, e, i, o* and *u*. Then the set is denoted by the notation

V = {a, e, i, o, u}

Each of *a, e, i, o, u* is called an *element or member* of the set V.

***Notation for membership:** e being an element/ member of set V is denoted by:* $e \in V$. *The fact that the letter b does not belong to V is denoted by the notation* $b \notin V$.

Generally, the name/identifier for a *set*, in this case V, is a sequence of letters and integers generally starting with a capital letter, and the name of an *element* consists of a sequence of letters and integers, starting with a small letter. However, a name must contain at least one

letter, it cannot be a sequence of integers only. *The naming convention is not at all strict*. The names may be specified in other ways also.

ii) In set notation, *the order in which elements occur within braces does not matter*. For example, the set *V = {a, e, i, o, u}* can also be denoted correctly *as V = {o, a, i, u, e}.*

iii) In set notation, the elements of a set are *not* repeated. For example, the following is *not* proper notation for the set V discussed above.

V = {a, e, i, o, e, u},

as the element *e* is repeated twice.

iv) If the set happens to have infinite number of elements, *problem arises in mentioning each element distinctly* within braces. For example, set of all integers, each of which is square of some integer. Then its members are 0, 1, 4, 9, 16, and so on. It is not possible to mention each member of the set between braces. In such cases, the notation is slightly modified as follows:

Sq_int = {0, 1, 4, 9, …}

in which, first four members of the set are mentioned, and then these are followed by three dots/ ellipses.

Set Representation Notation

Roster or tabular form: In roster or tabular form, all the members of a set are listed, separated by commas and are enclosed within braces '{ }'. For example: Set of all integers will be represented as, Z = {…, –3, –2, –1, 0, 1, 2, 3, …}

***Set-builder form*:** Consider the set *Sq_in_less_10001*. Every element of it has two properties: (a) it is square of an integer and (b) it is less than 10001.

The *set-builder form* uses these properties to denote the set as follows:

Sq_int_less_10001= {x: x= y^2, y is an integer, and x < 10001},

The *notation above states* that the members of the set *Sq_int_less_10001* are x, where *('where' is denoted by ':' or'|')x satisfies conditions: x= y^2, y is an integer, and x* < 10001.

There may be shades of differences in the set-builder notation. For example

The set *Decimal_digits = {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}* may be denoted as

Decimal_digits = {x: x ∈ N and x < 10} OR

Decimal_digits = {x ∈ N: x < 10}, etc.

The set-builder notation is quite convenient if the properties satisfied by all the members can be easily described, even if the set has very large number or infinite number of members.

Cardinality of a set S denoted by $|S|$

The notation $|x|$ is generally used to denote absolute value of x, *if x is a number.* However, if the notation is used with a ***set***, say, X as its argument, then it denotes the ***cardinality of X, i.e.*** $|X|$ denotes the number of members/ elements in X. ***For example, if X= {1, 2, 3}, then*** $|X| = 3$***.*** Another example: for the set *Decimal_digits = {0, 1, 2, 3, 4, 5, 6, 7, 8, 9},* $|\text{Decimal_digits}|$= 10.

Infinite Set: The symbol for 'infinite number' or infinity is ∞. By definition of ∞, ∞ = ∞ *–1* = ∞ *–2* = ∞*–3* = …

A set having unlimited/ uncountably many members is called an infinite set. Each of the sets *Sq_int = {0, 1, 4, 9, … }* and *Z = { … , − 3, − 2, − 1, 0, 1, 2, 3, …}* mentioned above is an infinite set. Any set *not* having infinite number of members is called a ***finite set.***

It may be noted that $|Z|$ = ∞, and also $|\text{Sq_int}|$ = ∞, $|Sq_int_less_10001| \neq \infty$

1.2.2 Set Relations

The general concepts of *'relation between sets'* and *'relation in a set'* will be formally defined later. For the purpose of the current discussion, let us define below some relations between two sets say X and Y. The relations to be defined are called *'is a subset', 'equality of two sets',* and *'is a proper subset'*.

Subset of a Set

As sets are (special) collection of objects, it may happen that every member of one set is also a member of another set. Then former set is called ***a subset of*** the latter set.

Formally, for two sets X and Y, *if* it happens that whenever x belongs to X then x belongs to Y (i.e. whenever $x \in X$ *then* $x \in Y$ *then* X is said to be a ***subset*** of Y and the fact is denoted by the notation: $X \subseteq Y$.

For example, if *Alphabet* is set of all letters a, b, c …, w, x, y, z. i.e.

Alphabet={a,b, c …, w, x, y, z}, and V= {a, e, i, o, u}

then $V \subseteq Alphabet.$

Instead of saying X is a subset of Y, it is also said that ***set X is contained in set Y.*** The symbol ***'⊆'*** *is also called* ***inequality of containment.***

In the following, we use **the notation** $\notin$ denoting '***does not belong to***'.

Proper Subset: Further, in addition to whenever $x \in X$ then $x \in Y$, it is also true that for at least one $y \in Y$ *with* $y \notin X$, then X is called a *proper subset* of Y. The fact is denoted by **X⊂ Y**, and then it is said that ***X is a proper subset of Y*** or ***X is properly contained in Y.***

Remarks: There is no complete standardisation of the notation of subset. Some people use '⊂' to denote a subset, which may or may not be proper

subset. Then, they use the symbol $\subsetneq$ for proper subset. We could have used these symbols. But, we would continue with symbols '$\subseteq$' and '$\subset$' denoting respectively 'subset' and 'proper subset'.

For example, if

Alphabet = {a, b, c …, w, x, y, z}, and V = {a, e, i, o, u}

Then, in view of the fact that $b \in Alphabet$ and $b \notin V$

$V \subset Alphabet$

Notation $\not\subset$: The fact that X is *not* a subset of Y is denoted by $\mathbf{X \not\subset Y}$.

Equality of two sets

Two sets X and Y are said to equal *if* $\boldsymbol{X \subseteq Y}$ and $\boldsymbol{Y \subseteq X}$. *The notation for equality of two sets X and Y is* $\boldsymbol{X = Y}$.

For example: The two sets *Decimal-digits* = {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}, and X = {x $\in N$: $x < 10$} are equal, i.e. *Decimal-digits* = X.

Inequality of two sets X and Y, denoted by $\boldsymbol{X \neq Y}$: Two sets X and Y are not equal (or are unequal) if at least one of the following happens: (i) there exists a $x \in X$ *such that* $x \notin$ Y or (ii) there exists a $x \in Y$ *such that* $x \notin$ X (or both (i) and (ii)).

Remarks: In mathematics and in logic, the word *or,* as in the phrase *A or B,* means *'either A, or B or both A and B'.*

Check Your Progress 1

1) Which of the following collections are sets? Justify your answer.

i) The collection of all the months of a year beginning with the letter M.

ii) The collection of all the months of a year beginning with the letter Z.

iii) The collection of ten most talented living writers of the world.

iv) A team of eleven most renowned football players of the world.

v) {1, 2, 2, 3}

vi) {1, 2, {2,3}, 3}.

2) For each of the sets consisting of following members, write (i) it in roster/ tabular form(ii) its cardinality:

a) Positive prime integer factors of 180

b) solutions of the quadratic equation $x^2 - 3x - 10 = 0$

3) Tell which of the following is a set, and if it is a set, then (i) write it in set-builder form (ii) tell its cardinality

a) {6, 9, 12, 15, 18, 21, 24, 27}

b) {2,4, 8, 16…}

c) {{1}, {1, 2}, {1, 2, 3}… {1, 2, 3, 4…10}}.

1.2.3 Special Sets: Definitions and Properties

Universal Set

While using mathematics for solving problems from some domain of human experience, some sets of entities from the domain are considered as and when required for facilitating arguments.

A *sufficiently big set of entities* from the problem domain, *which includes all possible entities* that may be required any time during the discussion or solution-process of solving the problems of the type under consideration, *is called the* ***Universal set*** *and is generally denoted by* ***U****.*

The Empty or Null set

It may appear strange, at least when heard for the first time, that a set may have no member at all, or the number of elements in a set may be Zero. According to our definition of a set, the following is a set

$$S = \{x: x \text{ is an even integer, and } x^2 = 9\}.$$

The collection S is well-defined, and hence, is a set. But, there is no member which satisfies the conditions of its membership. Therefore, the number of members of S is zero. There are many combinations of conditions, which will make a set to have no member. *A* ***null set*** *is denoted by the standard notation* $\emptyset$ or by {}.

Remarks: $\emptyset$ is a subset of every set X, i.e. $\emptyset \subseteq X$, for every set X.

1.2.4 The Intervals as Sets

The notation for intervals has already been discussed before. It may be stated that an interval—whether closed, open or semi-open—is a subset of real numbers such that if x and y are two members of interval I, and $x \leq y$, then every real number between x and y is also a member of interval I.

However, in the notation of an interval, instead of braces, i.e. '{'and '}',the brackets, i.e. '[' and ']' and/or parentheses, i.e. '(' and ')' are used as follows:

For real numbers a and b with a< b

- the interval [a, b] denotes the set of all real numbers between a and b and it also includes both a and b. Such an interval is called a *closed* interval.
- the interval] a, b] *(also, denoted as* ***(a, b]****)* denotes the set of all real numbers between a and b and it includes b, but not a. Such an interval is called a *semi-closed* or *semi-open* interval.
- the interval [a, b[*(also, denoted as* ***[a, b)*** *)* denotes the set of all real numbers between *a* and *b* and it also includes a, but not b. Such an interval is also called a *semi-closed* or *semi-open* interval.
- the interval]a, b[*(also, denoted as* ***(a, b)*** *)*denotes the set of all real numbers between a and b but, it does not include any of a and b. Such an interval is called an *open* interval.

1.2.5 Number Sets (With Standard Notations)

Next, we briefly discuss some frequently used number sets and their names/ notations.

N = {1, 2, 3…} is called the set of natural/ counting numbers.

W= {0, 1, 2, 3 …} is called the set of whole numbers.

It may be noted that for quite some time now, even 0 is being treated as a counting/ natural number. Thus N=W, and there is no need for the concept of whole number and for a separate notation W.

Z = {…,– 3,– 2, – 1, 0, 1, 2, 3…} is called the set of integers.

Q ={x| $x=p/q$ with q ≠ 0, and $p, q \in$ **Z**} is called the set of rational numbers.

Remarks: It can be easily seen that **∅⊂N ⊂W ⊂Z⊂Q.**

Except Z⊂Q, other set-inequalities given above are obvious. For showing Z⊂Q, let p $\in Z$, then p = (p/1)$\in Q$. *Further, 3/4* $\in Q$, *but 3/4*$\notin$ Z

R, the set of real numbers

The Set of Real Numbers is denoted by R, where R = {x| x is a real number}.

It can be easily seen that the length of hypotenuse of a square of length one unit is √2 units. But, it has been proved that there is no rational number p/q which equals to √2. Such a real number, which is not a rational number, is called an **irrational number.** It has been proved that number of irrational real numbers is much more than number of the rational real numbers. The set of irrational numbers may be denoted as **R–Q** or, as **R ~ Q.**

Remarks: From the discussion above, the containment relations '**⊂**' between various number sets can be further extended to include R as follows:

∅⊂N ⊂W ⊂Z⊂Q ⊂R.

The truth of these statements can be easily verified.

C, the set of complex numbers

The real number were required to be extended further, in context of solving equations of the type

$x^2 + 1 = 0$ or equivalently $x^2 = -1$

In view of the fact that for any real number, its square is a positive real number, the equation above cannot be solved for any real number.

The set of complex numbers were developed in context of attempts at solving equations of the type $x^2+1 =0$, for which no real number solution exists. A new type of number *i* was introduced for which $i^2 = -1$ was assumed. Then C is defined as

$$C = \{z: z = x + i\, y,\ x, y \in R\}$$

*(**Note:** x + i y is also written as x + y i)*

Then inequality $\boldsymbol{R \subset C}$ *follows from the following argument:* Let x ∈ R, then x = x + i 0 ∈ C. Also, there are many complex numbers including, say, 1 + i 2 such that 1 + i 2∉ R. (Also, alternatively, as discussed above, i ∉ R, but, i = 0 +1 i ∈ ***C***). Hence, the proof follows.

Also, using the earlier set inequalities, we get

$$\emptyset \subset N \subset W \subset Z \subset Q \subset R \subset C$$

Remarks**:** The notations **∅, N, W, Z, Q, R,** ***C*** are *standard* and can be used without giving their respective definitions/ contents. However, the other set names like *Sq_int* are *not* standard; we could have used some other name instead of *Sq_int*. For all such set names, in every new context, its definition has to be given afresh.

Check Your Progress 2

1) For the pair of sets X and Y, state whether X = Y or not, along with explanation for your response.

   i) X = {2, 4, 6, 8, 10}, Y = {*x*: *x* is positive even integer and $x \leq 10$}

   ii) X = {*x*: *x* is a multiple of 10}, Y = {10, 15, 20, 25, 30 . . .}

2) Find the correct relation out of $X \subseteq Y$, $Y \subseteq X$, $X \subset Y$, $Y \subset X$ or $X \not\subset Y$ etc. for the following pairs of sets X and Y:

   i) X={*x*: *x* is a student enrolled in IGNOU and having Economics as one of the courses},

      Y={*y*: *y* is a student enrolled in IGNOU}

   ii) X= {*x*: *x* is a triangle in a plane}, Y= {*y*: *y* is a rectangle in the plane}

3) For X = {1, {2, 3}, 3, 4} which one of the following is/are true statement(s):

   (i) $2 \in X$ (ii) $\emptyset \in X$ *(iii)* $\emptyset \not\subset X$ (iv) $\{2,3\} \in X$. *(v)* $\{2,3\} \subset X$

1.2.6 Set Operations

We know how new numbers are obtained from some already given numbers through number-operations. For example, for a given number, say 2, we get (– 2) as negative (or additive inverse) of 2; 2^{-1} = (1/2) as multiplicative inverse of 2, $\sqrt{2}$ as square-root of 2, etc. Each of the additive inverse, multiplicative inverse and square-root is a *unary* operation on numbers. Each of these operations is called *unary operation*, as each takes a *single* number (in the above discussion, it is 2), and associates some (generally new) number e.g. (1/2) in the case of multiplicative inverse. Similarly, for given *two* numbers, say, 2 and – 5, through the binary operations plus (+), minus (–), multiplication (×) , division (÷) , and exponent/ power , we obtain respectively –3 (= 2– 5), 7 (= 2– (–5)), –10 (= 2× (– 5)), –2/5 (= 2/ (– 5)) and 2^{-5} (= 2 raised to power (– 5)). Each of plus, minus, multiplication, division, and exponent is a *binary* operation as each requires *two* numbers (called *arguments*) to perform its role and return the result.

Similarly, unary and binary operations have been defined for sets, so that a set operation after taking sufficient number of sets as arguments, returns a (generally new) set as result of its action.

We will consider only two types of set operations: (i) *unary* set operation and (ii) *binary* set operations. First, we will discuss the *binary* operations. Then we will discuss *only one unary* operation, as this is a special case of some binary operations.

For *two* given sets X and Y, the *binary* set operations, to be discussed, are (i) *Union* of sets, denoted by the symbol '$\cup$' as in $X \cup Y$, (ii) *intersection* of sets denoted by the symbol $\cap$ as in $X \cap Y$, (iii) *difference* of (two) sets denoted by the symbol '– ' or '~' as in X – Y (or X~Y), and (iv) the *symmetric difference* operator Δ as in $X \Delta Y$. Finally, we will discuss the *unary operation, complement of a set denoted by ' ' ', as in 'X'*.

Note: Symbol '$\cup$' denotes union of sets, while letter 'U' denotes the *universal* set.

The pattern of discussion for each operation is as follows: First, the definition of the operation is given, which is then illustrated by an example. Finally, some properties of the operation are stated.

For the discussion of the properties, in the following, the sets X, Y and Z are arbitrary sets.

a) ***Union of two sets X and Y ($X \cup Y$) is defined as***

$$X \cup Y = \{x: x \in X \textbf{ or } x \in Y\}$$

For example, if X= {1, 2, 3, 4} and Y = {2, 3, 5, 8, 9}, then

$$X \cup Y = \{1, 2, 3, 4, 5, 8, 9\},$$

Elements occurring in both X and Y appear only once in $X \cup Y$, as elements are not repeated in representation of a set.

Some Properties of Union operation

For all sets X, Y and Z, the following set identities are true

i) $X \cup Y = Y \cup X$ (*Commutative* law or commutativity of sets *w.r.t. union operation*)

ii) $(X \cup Y) \cup Z = X \cup (Y \cup Z)$ (*Associative* law *w.r.t.* union operation)

iii) $X \cup \emptyset = X$ (the null set $\emptyset$ is the *identity* for union operation)

iv) $X \cup X = X$ (*Idempotent* law *w.r.t.* union operation)

v) $U \cup X = U$.

b) ***The Intersection of two sets X and Y ($X \cap Y$) is defined as***

$$X \cap Y = \{x: x \in X \text{ and } x \in Y\}$$

For example, if X= {1, 2, 3, 4} and Y = {2, 5, 8, 3, 9}, then $X \cap Y = \{2, 3\}$.

Some Properties of Intersection operation

For all sets X, Y and Z, the following set identities are true

i) X∩Y = Y∩X (Commutative law or commutativity of sets *w.r.t.* intersection operation)

ii) (X ∩Y) ∩Z = X ∩(Y ∩Z) (Associative law for intersection operation)

iii) X ∩ U = X (the universal set U is the *identity* for intersection operation)

iv) X∩X = X (*Idempotent* law *w.r.t.* intersection operation)

v) Ø ∩X = Ø

c) ***The Difference X – Y of two sets X and Y is defined as***

$X - Y = \{x: x \in X \text{ and } x \notin Y\}$

For example, if X= {1, 2, 3, 4} and Y = {2, 5, 8, 3, 9}, then

$$X - Y = \{1, 4\} \quad \text{and } Y - X = \{5, 8, 9\}$$

Some Properties of Difference operation

i) *Difference operator is* not *commutative,* i.e. *for at least one pair* of X and Y, X – Y ≠ Y –X

It is shown just above by pair of X and Y.

Another simpler example:

Let X= {1, 2, 3, 4} and Y =Ø, then X – Y = X ≠ Ø = Y –X

ii) *Difference operator is* not *associative, i.e. for at least one triple of sets X, Y and Z,*
$(X - Y) - Z \neq X - (Y - Z)$

Let X= {1, 2, 3, 4} and Y = {2, 3, 5, 8, 9} and Z = {1, 5, 7,8}

Then, from above, $X - Y = \{1, 4\}$, *hence* $(X - Y) - Z = \{4\}$

And $Y - Z = \{2, 3, 9\}$. Hence, $X - (Y - Z) = \{1,4\}$

Thus, $(X - Y) - Z = \{4\} \neq \{1, 4\} = X - (Y - Z)$

Another simpler example:

Let X= {1, 2, 3, 4} and Y = Ø, and Z = {1} then(X – Y) – Z ={1, 2, 3, 4} – {1} = {2, 3, 4}

(Y – Z) = Ø, and hence, X – (Y – Z) = {1, 2, 3, 4}– Ø ={1, 2, 3, 4}.

iii) There is no both-sided identity *w.r.t.* 'difference operator', however, Ø serves as one-sided identity, as X– Ø = X, but Ø – X = Ø, and not X

iv) *Idempotent Law does not hold w.r.t. difference operator, as* X– X = Ø

d) *Another well-known binary operator* ***'symmetric difference'*** *for two given sets X and Y, is denoted by* **X Δ Y,** and is defined as follows:

X Δ Y = (X– Y) ∪ (Y–X)

Another way of defining ***Δ*** *is*

X Δ Y = (X ∪ Y) – (X ∩Y).

For example, X= {1, 2, 3, 4} and Y= {2, 4, 5, 7}

Then X Δ Y = (X ∪ Y) – (X ∩Y) = {1, 2, 3, 4, 5, 7} – $\{2, 4\} = \{1, 3, 5, 7\}$.

The following are some of the notable identities w.r.t. (binary) set operation Δ for all sets X, Y, Z, we have

i) X Δ Y = Y Δ X, (Δ is commutative/ symmetric)

ii) *(*X Δ Y) Δ Z = X Δ (Y Δ Z), (Δ is associative)

iii) X Δ Ø = Ø Δ X = X (Ø is identity for Δ)

iv) X Δ X' = X' Δ X = U, the universal set (as X∪X' = U and X ∩ X' = Ø)

e) ***Unary Set operator 'Complement',*** denoted as **X'** for a set X, is defined as

X' = U – X, where U is the universal set.

One notable property w.r.t. complement is (X')' = X

Properties/identities of set operations, when two or more than two operations are involved

i) Z∪ (X ∩Y)= (X ∩Y) ∪ Z = (X∪Z) ∩(Y ∪ Z)

ii) (Y ∪ Z) ∩ X = X ∩ (Y ∪ Z) = (X ∩Y) ∪(X ∩Z)

iii) (X ∩Y)' = X' ∪ Y' (De Morgan's Law)

iv) (X ∪ Y)' = X' ∩Y' (De Morgan's Law)

1.3 RELATION AND FUNCTIONS

We have some informal idea about the concept of 'relation' through our discussions in sub-section 1.2.2 and the 'Remarks' above. For our purpose below, a relation takes two (otherwise, may also take one or more) arguments *and returns a 'true' or a 'false'*.

Many of the (formal) concepts in mathematics and other academic disciplines have their roots in some aspects, entities or phenomena encountered routinely, and observed casually or even unconsciously. 'Relation' to be defined formally in this section is one such concept. We know the relations of *'father'*, 'daughter', 'friend'. What significance any of these has for mathematics? The significance is that each can be considered as a (mathematical) *set* by the method explained below. For this purpose,

first, we need the concept of *'ordered set'* and its special case *'ordered pair',* and also the concept of *'Cartesian product'.*

Ordered Pair/ Set: We know from co-ordinate geometry that the point (1, 2) is different from the point (2, 1), i.e., (1, 2) ≠ (2,1). Hence, order of occurrence of numbers is taken into consideration in the pairs representing co-ordinates of points. On the other hand, sets {1, 2} = {2,1}, as the order of elements is not taken into consideration in the representation of a *general* set.

Thus, an ***ordered set*** is a set in which *order of occurrence of elements* is required to be taken into consideration. If a, b, c … are elements of an ordered set, then the ***ordered set is denoted as (a, b, c, …)***.

Cartesian-product or Cross-product of two *non-empty* sets X and Y, denoted as X × Y, is the set of all ordered pairs (x, y), i.e.

$$X \times Y = \{(x, y): x \in X \text{ and } y \in Y\}.$$

If (at least) one of X and Y is Ø, the empty set, then, by definition X × Y is Ø.

Remarks: *It may be noted that each (x, y) of (X × Y) is an ordered set, but, X × Y itself need* not *be an ordered set.*

Remarks: $|X \times Y| = |X| . |Y|$, i.e. *cardinality-of (or number of elements in)* X × Y is equal to product of the cardinalities $|X|$ and $|Y|$.

We now attempt to define the mathematical concept of 'relation'. An intuitive notion of some relation can be mapped to a mathematical (formal) concept of relation.

A (mathematical) relation from a non-empty set X to a non-empty set Y is some subset of X × Y*. Conversely, every subset of X × Y gives rise to a relation from the set X to the set Y.*

An example: we know that every person has (at a particular time) some fixed height. This concept of HEIGHT may be considered as a mathematical relation as follows:

Let X be set of all persons, and Y be the set of positive real numbers, given by {x: $x \in R$, and $x > 0$}

Then, the relation HEIGHT is a subset of X × Y, such that

HEIGHT = {(x,y) : $x \in X$, $y \in Y$ }. Here, y represent the height of x in inches.

We have already encountered ***arithmetic relations*** (e.g. =, ≠ , <, ≤ and ≥) and ***set relations*** (e.g. ⊆, =, ⊂, ⊄, ≠). We reconsider some of these in terms of the formal definition of relation, given just above.

Let us discuss the relation < on N, the set of Natural numbers. By definition of a relation, it is a subset of N × N, which is given by

Less-than-on-N = { (1,2), (1,3), (2,3), (1,4), (2,4), (3,4), (1,5), (2,5), (3,5), (4,5), …}

Or, even as

'<'= {(x, y): x, y ∈ N, and x < y}.

Remarks: *It may be reminded that the second symbol from L.H.S. i.e. '=' here is not in the sense of 'is equal to', but, '=' stands for 'L.H.S. is defined as the R.H.S'.*

Image of an element, Domain, Range and Codomain of a relation:

For any relation R from a set X to a set Y, and for any ordered pair (a, b) in R, the element *b* is called ***image*** of element *a* .

The *set* of *all* first elements of the ordered pairs in a relation R from a set X to a set Y is called the ***domain*** of the relation R.

The *set* of all second elements in a relation R from a set X to a set Y is called the ***range*** of the relation R. The whole set Y is called the ***co-domain*** of the relation R. Note that for any relation R, ***range of R⊆ co-domain of R***.

For example, for the relation

Air-distance-from-Delhi = {(Chennai,1759), (Mumbai, 1153), (Kolkata, 1305), (Hyderabad, 1259), (Bangaluru, 1744)},

1759 is the ***image*** of Chennai, 1259 is ***image*** of Hyderabad, etc. ***Domain*** of the relation *Air-distance-from-Delhi* is the set {Chennai, Mumbai, Kolkata, Hyderabad, Bangaluru} and its ***range*** is the set {1759, 1153, 1305, 1259, 1744}.

Functions: The usage of the concept of *function* in its current sense began probably sometime at the end of 17^{th} century by Gottfried Leibnitz in his letters to Johann Bernoulli. However, visualisation of function as a special type of *relation* came sometime in the beginning of 20^{th} century.

For two given *non-empty* sets *X* and *Y*, ***a function f from X to Y***, denoted as

$$f : X \to Y$$

is a *rule* which associates to *each* element, say, *x* of *X* a *unique* element , say, *y* of *Y*.

Then, X is called the ***domain*** *of f and Y is called the* ***co-domain*** *of f.* Also, if f(*x*) = *y*, then *y* is called ***image*** of *x*, and *x* is called ***pre-image*** of *y*.

The names ***'map'*** *and* ***'mapping'*** *are also used instead of the name* ***'function'.***

If *Y* = *R*, the function *f*: *X* → *R* is called ***real-valued*** function. If *Y*= *C*, then *f*: *X* → *C* is called *complex-valued* function. Similarly, by taking Y, the co-domain, as Z or Q, we get respectively *integer-valued* and *rational-valued* functions.

For example,

i. Let *X* = *Y* = *N*, the set of natural numbers, then
f: *N* →*N*, such that *f*(*x*) = 2*x* + 1, ∀x∈*N*.

Clearly, for each *x*∈*N*, 2*x* + 1 exists in *N* and is unique for each *x*. Hence, *f*: *N*→*N* is a function.

ii. Let *f*: *Z*→*N*, such that $f(x) = x^2 + 3$,
then for each *x*∈*Z*, $x^2 + 3 \in N$ and is unique, therefore, *f* is a function

iii. Let $f: Z \to N$, such that $f(x) = x^2 - 3$.
Then for $x = 0$, $x2 - 3 = -3$ N, Hence, $f: Z \to N$ is not a function.

iv. Let $f: Z \to Z$, such that $f(x) = x^2 - 3 \in Z$,
then $x^2 - 3 \in Z$ (codomain), $\forall x \in Z$ (domain). Hence, $f: Z \to Z$ is a function.
(**Note:** Change of codomain has converted a non-function to a function)

Check Your Progress 2

1) For the two sets X = {1, {2, 3}, {4, 5}, 6} and Y = {1,{2, 4}, {5, 6, 7}}, find the value of each of

   (i) $X \cup Y$, (ii) $X \cap Y$, (iii) $X - Y$ (iv) $X \Delta Y$

2) Let X = {2, 3, 4, 5, 6, 8}. Let R be the relation on X defined by

   {(x, y): $x, y \in$ X, y is divisible by x *and* $x \neq y$} then

   (i) Write R in roster form, (ii) Find the domain of R and (iii) Find the range of R.

3) Determine the domain and range of the relation R defined by

   R = {(x, x + 5): $x \in$ {0, 1, 2, 3, 4, 5}, y$\in$ { 5, 6, 7, and $x^2 <$ y }.

1.4 LOGIC

We initiate our discussion of the topic with the questions: Why (we study) Logic? and, What is Logic?

We start with simple short answers to the questions and then gradually expand the answers to meet our requirement.

At the outset, we say that *Logic* is about *reasoning, rather reasoning properly/ correctly. It tells and trains us in how we* should *reason properly so that*

i) we always arrive at (the same) correct conclusion, every time, from the initially assumed/ given facts, and

ii) The process of arriving at the conclusion from the initially assumed facts is convincing enough like mathematical calculations, so that hardly any possibility of dispute remains between involved (reasonable) parties.

In fact, logic investigates and provides tools and techniques of inferring and deriving new facts (conclusions) only by starting with some given facts/ information. However, what constitutes initial assumptions of an academic field is the subject matter of that academic discipline, whether it is Physics, Economics or Political Science, etc and not of the logic. *For the sake of emphasis, it is repeated that logic is used and applied only for deriving correct conclusions from the given assumptions.* For example,

i) only economists, and not logicians, decide the *basic concepts* of Economics like scarcity, choice, trade-off, incentive, exchange, information, distribution, etc., and the fundamental principles or laws of Economics such as 'law of diminishing returns', etc.

ii) only mathematicians decide that (a) basic assumptions for geometry should be Euclid's postulates, and (b) basic assumptions for numbers should be Peano's axioms.

Some basic concepts of logic are(i) logical constants, *viz.* True and False, also written as 0 and 1 (ii) statement (iii) argument, rather, argument form – logical connectives and rules of inference. Next, we start with the basic concept of argument.

'Argument' may be considered as *the central* concept of logic, as *atom* is of physics, in the sense that discussion of more *complex* concepts like molecule, etc. on one hand, and electron and proton, etc. on the other hand starts with *atom.*

Argument: The concept of argument in logic *is different from* the notion of argument in everyday language. *An argument in logic* is a set, or a sort of a sequence of (n+1) *statements* say, S_1, S_2 … S_n, S_{n+1}, with n $\in$ N (in some cases *n* may be even 0 also), such that S_{n+1} is concluded from the other statements.

Such an argument is generally denoted (vertically) as

S_1,
S_2,
…
S_n

S_{n+1}

Or, (horizontally) as

$S_1, S_2 \ldots S_n$

S_{n+1}

to indicate that the statement S_{n+1} *(the statement below the horizontal line)* is inferred or concluded from *given/ assumed* statements S_1, S_2 … S_n *(all above the horizontal line). 'Inferring'* means that *if* we *assume* all the statements above the horizontal line as True, *then, through inference*, we take the statement below horizontal line also to be True.

The quality/ reliability of conclusion depends, in addition to the assumed statements, on the *approach* used for making a conclusion, or inferring the conclusion. Some well-known approaches include (i) deduction, (ii) induction, and (iii) abduction. We will come back to the discussion of the approaches later. Next, we discuss the concept of '*statement*', the constituent concept of argument.

Statement: A *statement,* also called a **Proposition**, is a *declarative* sentence*, i.e.* a *sentence* for which it is possible to say that it is 'True' or 'False'. For example, each of the following three sentences is a statement: (i) Sugar is sweet (ii) Sun rises in the west (iii) Peter is a graduate. The statement (i) is

True, the statement (ii) is False and the statement (iii) may not be known at present, but, has exactly one of the two possible values, *viz.* True and False. Each of the values 'True' and 'False' associated with a statement is called *truth value* of the statement.

In order to clarify the difference between a sentence and a statement, let us consider below some sentences which are *not* statements:

– Who was the first Prime Minister of India?*(Interrogative sentence)*

– Please, give me that book. *(Imperative sentence)*

– You must exercise regularly. *(Imperative, rather Deontic)*

– Hurrah! We have won the trophy.*(Exclamatory sentence)*

Symbolic Language and Logical Connectives

Why symbolic language?

a) The symbols are crisp and very short for long statements; for example, we may use the symbol '***CMRain'***, or symbol **CR**, or even symbol P for the statement: *Cherrapunji/ Chirapoonji, a town in the north-east India, received maximum average rainfall in the world during the years 1901–2000.*

b) In a logical argument, we are, generally, interested in the logical truth 'True' or 'False' of a statement and not in its content or structure. The symbols *abstract away* the other aspects of a statement to highlight the logical content better than the corresponding English statements do.

c) In symbolic representation, a *single* symbol may represent *many* equivalent forms of a statement. For example, 'Kalidas wrote Meghaduta', ' Meghaduta was written by Kalidas', ' Meghaduta is one of the works of Kalidas', etc. are distinct but almost equivalent statements, however, which may not be recognised as such, at least by some computer system. But, symbolically, we may just say that *Composition (Kalidas, Meghaduta)* is 'True'.

Symbolic Language: A symbolic language for logic consists of the following types of symbols:

i) ***Symbolic names*** for statements, as *P, Q, R, CMRain* and *EF*.

*(only **atomic statements** are assigned **symbolic names**, where, an **atomic** statement is a statement which does not contain (i) any logical connective and (ii) any quantifier*)*

ii) A symbolic name in logic is a ***logical variable*** in the sense that at any time, it may be assigned exactly one of the two possible logical constant values *viz*. T or F(discussed under (iii) below)

iii) **Symbolic Constants:** The two truth values *'True'* and **'***False***'** are two *constants* of symbolic logic**.** The two values are also denoted as '*T*' and '*F*, or even *'1'* and *'0'*. We will generally use the symbols 'T' and 'F' for the two logical constants*.*

***Quantifiers**

In sub-section 1.1.2 under quantifiers and logical operators, we have already introduced the quantifiers, viz. (i) ($\forall$x) denoting the phrase 'for all x$\in$X', (ii) ($\exists$ x) denoting the phrase 'for same x $\in$ X' and (iii) ($\nexists$ x $\in$ X) denoting the phrase 'for none of x $\in$ X' etc.

iv) ***Logical connectives, quantifiers, Symbolic Expressions, Well Formed Formula (in short wff)*** each of which will be discussed in detail.

v) ***Statement forms (Symbolic parameters),*** generally, denoted by letters in some non-normal font or even by Greek letters Φ, Ψ, etc. for representing identities like De Morgan's Laws: ***~ (Φ ∨ Ψ) ≡ ~ Φ ∧ ~ Ψ,*** where, each of Φ and Ψ may be replaced by a ***wff***.

Let us discuss (iv) and (v) in detail.

Logical Connectives:

Five of the frequently used logical connectives are (with their names in parentheses):

~ (negation, 'not'), ∧ (conjunction, 'and'), ∨ (disjunction, 'or'),
→ (conditional or implication), ↔ (bi-conditional or bi-implication)

Each of these is a function to the set {T, F} as defined below:

i) The function ' ~ ' (negation) is a *unary* logical operator, from *{T, F} to {T, F} and is* defined as: ~ (T) = F , and, ~ (F) =T

Truth Table: is a notation for a logical function/ functions in tabular form

For example Truth table of negation operator is

A	~A
T	F
F	T

Where, A represents a statement.

*The remaining four are **binary** operations that maps {T, F} x {T, F} to {T, F},* these are defined as

ii) The function '∧'(conjunction): (T∧ T) = T, (T∧ F) = F, (F∧T) = F and (F∧F) = F,

iii) The function '∨ ' (disjunction) is defined as: (T∨ T) = T, (T∨ F) = T, (F∨ T) = T and (F ∨ F) = F,

iv) The function '→' (material implication/conditional) is defined as: (T→ T) = T, (T→F) =F, (F→T) = T and (F → F) =T,

v) The function '↔'((bi-conditional) is defined as: (T↔ T) = T, (T↔ F) =F, (F↔ T) = F and (F↔ F) = T.

The truth table for the four *binary* logical operators (to be read column-wise):

A	B	$A \wedge B$	$A \vee B$	$A \rightarrow B$	$A \leftrightarrow B$
T	T	T	**T**	T	**T**
T	F	F	**T**	F	**F**
F	T	F	**T**	T	**F**
F	F	F	**F**	T	**T**

Remarks: The symbols for various logical operators are *not standard/unique* in the sense that other symbols are also used in practice for logical connectives. For example: (i) ¬, ~ and ! are used for *negation* (ii) $\wedge$, and & are also used for *conjunction*, (iii) $\forall$, + and // are used for *disjunction*. Similarly, there is variation in symbols for each of conditional and bi-conditional.

Well-formed Formula (in short, *wff, or statement)*

Like any *natural* language, there are rules which determine whether an expression (of symbols) constitutes a proper form of representation or not. For example, in infix notation (a notation for which the operator is inserted between terms) for binary logical operators, the following is proper representation $P \rightarrow (S \vee U)$, but, ***none*** of the following is in proper form: $\rightarrow P (S \vee U)$, $P \rightarrow \vee S U$, and $\rightarrow \vee P SU$, *etc.*

Each of formal expression in proper form is called ***Well-formed formula*** *(in short,* ***wff or statement****). We will not go in more detail of this topic.*

Interpretation and Evaluation of a wff

As mentioned above, a logical symbol can assume at any time exactly one of two possible values 'T' and 'F'. If we have a logical expression E involving two symbols, then, there are four possibilities of values which can be assigned to the two symbols in E. Each of the four possible assignments is called an ***interpretation*** of symbols occurring in E, and then a logical value of E can be found corresponding to this interpretation. For example, the wff E given by involves two symbols A and B. Let us assign the truth values F and T respectively to A and B.

$$E: \sim ((A \vee (\sim B)) \rightarrow (A \leftrightarrow B))$$

In other words, we assume the *interpretation, say I*, where I assigns F to A and T to B. *I is also written as {~ A, B}.* The truth value of expression E is obtained by replacing A and B respectively by F and T in E. Thus, we get

$$\sim ((F \vee (\sim T)) \rightarrow (F \leftrightarrow T))$$

Using the Truth-table given above, we successively obtain

$$\sim ((F \vee F) \rightarrow F)$$

$$\sim (F \rightarrow F)$$

~ (T)

F

Thus, F is the Truth value of E under the interpretation which assigns F to A and T to B.

Logically Equivalent Formulas (or Logical Identities): Two formulas G1 and G2 are said to be (logically) *equivalent* if for *each* interpretation (*i.e., truth assignment to all the atoms that occur in either G1 or G2*); the truth values of G1 and G2 are identical. In other words, two formulas G1 and G2 are said to be (logically) *equivalent if and only if* for each interpretation, G1 is True if G2 is True, and , for each interpretation, G1 is False if G2 is False.

Example

We can verify that the formula G_1: ~ $(A \rightarrow B)$ is equivalent to the formula G_2: $A \wedge \sim B$ by examining the following truth table. The corresponding values in the last two columns are identical.

A	B	~ B	(A → B)	~ (A → B)	A ∧ ~ B
T	T	F	T	F	F
T	F	T	F	T	T
F	T	F	T	F	F
F	F	T	T	F	F

Table of (well-known) Equivalences of Formulae

As per our convention, each of E, G and H denotes a *statement form or a general wff*, and not necessarily an atom or a particular wff

(1.1) $E \leftrightarrow G = (E \rightarrow G) \wedge (G \rightarrow E)$

(1.2) $E \rightarrow G = \sim E \vee G$

(1.3)(a) $E \vee G = G \vee E$ (commutativity of $\vee$)

(1.3)(b) $E \wedge G = G \wedge E$ (commutativity of $\wedge$)

(1.4)(a) $(E \vee G) \vee H = E \vee (G \vee H)$ (associativity of $\vee$)

(1.4)(b) $(E \wedge G) \wedge H = E \wedge (G \wedge H)$ (associativity of $\wedge$)

(1.5)(a) $(E \vee (G \wedge H)) = (E \vee G) \wedge (E \vee H)$ (Distributivity of $\vee$ over $\wedge$)

(1.5)(b) $(E \wedge (G \vee H)) = (E \wedge G) \vee (E \wedge H)$ (Distributivity $\wedge$ of over $\vee$)

(1.6)(a) $E \vee$ False = E (False is identity for $\vee$)

(1.6)(b) $E \wedge$ True = E (True is identity for $\wedge$)

(1.7)(a) $E \vee$ True = True

(1.7)(b) $E \wedge$ False = False

(1.8)(a) $E \vee \sim E$ = True (Law of excluded middle)

(1.8)(b) $E \wedge \sim E$ = False

(1.9)(a) $E \wedge E = E$ ($\wedge$ is idempotent operation)

(1.9)(b) $E \vee E = E$ ($\vee$ is idempotent operation)

(1.10) $\sim (\sim E) = E$

(1.11)(a) $\sim (E \vee G) = \sim E \wedge \sim G$ (De Morgan's Law)

(1.11)(b) $\sim (E \wedge G) = \sim E \vee \sim G$ (De Morgan's Law)

1.5 PROOF TECHNIQUES

Following are the well-known techniques for proof in Mathematics:

i) Direct proof and counter examples

ii) Proof by contradiction and Proof by contra positive

iii) Proof by cases

iv) Existence proof

v) Proof by Mathematical Induction

vi) Proof of equivalences

Generally, for different types of problems, different techniques may be more appropriate, though, for some problem types, more than one techniques may be equally appropriate.

A problem requiring a proof consists explicitly of two parts, viz. (i) what is given as a set of assumed facts/axioms/ postulates, etc., all taken together to be called ***Hypotheses*** *and* (ii) what is to be proved (to be called **Conclusion**). Apart from these, *implicitly,* a number of definitions, tools/ techniques/ strategies, earlier proved theorems, etc. commonly used in such type of proofs, *are also* assumed to be given.

For example, suppose that we are given the problem: *Prove that for all integers m and n, if m and n are odd, then their product (m. n) is odd.*

In this problem, *the explicitly given/assumptions* include: (i) each of m and n is an integer and (ii) each of m and n is an odd integer. Here, (ii) subsumes (i), i.e. whatever is said under (i) is implied by (ii). *What is to be proved* is: product *'m. n is odd'*. *Implicitly*, a number of appropriate definitions, appropriate properties of the involved elements, the valid rules of inference, and proof techniques are also assumed to be given, along with some other assumptions. Appropriate choice of rules of inference and technique(s) play significant role in solving the problem convincingly and efficiently.

A detailed discussion on these proofing techniques are beyond the scope of this Unit.

1.6 LET US SUM UP

This being the first Unit of this course introduces to some of the basic mathematical concepts. Some of these concepts may be very useful in the future. Set theory is one of the most basic and most applicable in many problem situations. This Unit has introduced you to the concept and representation of set, set notations, operations on the sets, and some important number sets. The concept of mathematical relation, function, logic and proofing techniques are also introduced. You must study the truth tables of the logic connectors as they will be very useful in creating logical expressions.

1.7 REFERENCES

1) K. Sydsaeter and P. Hammond, 2012. *Mathematics for Economic Analysis,* Fourth Edition, Pearson, Educational Asia: Delhi.
2) NCERT Chapters 1, NCERT Math class XI, NCERT publication 2018.
3) NCERT Chapters 1& 2, NCERT Math class IX, NCERT publication 2018.
4) IGNOU Unit 2/ Block 1& Unit 1 of Block2, MCSE-003: Artificial Intelligence and Knowledge Management, IGNOU, (2008).
5) R. Johnsonbaugh, 2009. *Discrete Mathematics* (Seventh Edition) Pearson Education Asia.
6) Thomas J. Mckay, 1989. *Modern Formal Logic* MacMillan Publishing Company.
7) Copi I., Cohen C. McMahon K. 2014. *Introduction to Logic* (Fourteenth Edition) Pearson.

1.8 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Each of (i), (ii), (vi) being well-defined collection is a set. On the other hand, each of (iii) and (iv) is not well-defined, hence is not a set. (v) is not a set, as repetition of elements in set notation is not allowed.

2) (a) {2, 3, 5}, cardinality=3, (b) {5, –2}, cardinality =2

3) (a) it is a set; {x: $x \in N$, *3 divides x, and* $6 \leq x \leq 27$}, cardinality =8 (b) it is a set, {x: $x= 2^n$, n $\in N$ }, cardinality =infinite (*or,* ∞) (c) it is a set, {X: X= {1, …,i}, i $\in$ N }, cardinality = 10.

Check Your Progress 2

1) (i) X=Y, (ii) X ≠ Y, as 15 $\in Y$, *and* 15 $\notin$X

2) (i) $X \subset Y$ (ii) X$\not\subset Y$ *and* Y$\not\subset X$

3) (i) False (ii) False (iii) False (iv) True (v) False

Check Your Progress 3

1) (a) X $\cup$ Y = {1, {2, 3}, {4, 5}, 6, {2, 4}, {5, 6, 7}}; X $\cap$ Y = {1}; (X – Y) = {{2, 3}, {4, 5}, 6} X Δ Y = {{2, 3}, {4, 5}, 6, { 2, 4}, {5, 6, 7}}

2) R = { (2,4), (2,6), (2,8), (3, 6), (4, 8)}, Domain of R = {2, 3, 4}, and Range of R = {4, 6, 8}

3) Domain of R = {0, 1, 2} and Range of R = {5, 6, 7}.

UNIT 2 STATISTICAL CONCEPTS

Structure

2.0 Objectives

2.1 Introduction

2.2 Some Elementary Concepts

- 2.2.1 Statistics/ Data
- 2.2.2 Variable and Its Types
- 2.2.3 Source of Data
- 2.2.4 Types of Data
- 2.2.5 Measurement Scales
- 2.2.6 Classification of Data

2.3 Descriptive Statistics

2.4 Quantitative Data – Percentages and Measures of Central Tendency

- 2.4.1 Use of Percentages
- 2.4.2 Measures of Central Tendency
- 2.4.3 Empirical Relationship between Averages
- 2.4.4 Which Measure to Choose?

2.5 Quantitative Data – Measures of Dispersion

- 2.5.1 Range and Coefficient of Range
- 2.5.2 Mean Deviation and Coefficient of Mean Deviation
- 2.5.3 Standard Deviation and Variance
- 2.5.4 Coefficient of Variation
- 2.5.5 Distribution Shapes
  - 2.5.5.1 Skewness
  - 2.5.5.2 Kurtosis

2.6 Quantitative Data – Measures of Position

- 2.6.1 Standard Scores
- 2.6.2 Percentiles
- 2.6.3 Quartiles and Deciles

2.7 Let Us Sum Up

2.8 References

2.9 Answers or Hints to Check Your Progress Exercises

2.0 OBJECTIVES

After going through this unit, you will be able to:

- explain the concept of data, variable, its classification and measurements;
- delineate the concept of central tendency and variability in the distribution;
- compute the measures of central tendency and measures of variability;

- know how to select the best measure of the Central tendency;
- discuss the Measures of Position;
- state the Process of Exploratory Data Analysis;
- learn how to get the results of percentage, mean, mode, median, skewness and kurtosis by using the spread sheet package.

2.1 INTRODUCTION

In the process of data analysis, various mathematical and statistical concepts are used. Various mathematical concepts like number, set, relation, functions etc. have already been explained in Unit 1. Before undertaking any Data Analysis, we would like to know: what do we mean by data? What are its various types? What are sources of data? What kinds of variables we may have in the course of data analysis? What are the various measurement scales? How can we summarise or describe the characteristics of data? We shall explain all these issues in this Unit. Further, in a situation to summarise or describe the characteristics of variable (s) we tend to look at central tendency, distribution, dispersion and location. All these methods are put under the category of descriptive statistics. Being these methods basic or foundation in nature, we shall cover all these methods in this Unit. Let us begin with explaining the concept of Statistics/Data.

2.2 SOME ELEMENTARY CONCEPTS

2.2.1 Statistics/ Data

The word statistics is used to convey different meanings in singular and plural sense. Therefore, it can be defined in two different ways. In singular sense 'Statistics' refers to what is called statistical methods. When 'Statistics' is used in plural sense, it refers to 'data'. In this context, it is described as a branch of science which deals with the collection of data, their classification, analysis and interpretations of statistical data. Put in other words, Statistics is concerned with scientific methods for collecting, organising, summarising, presenting and analysing data, as well as drawing valid conclusions and making reasonable decisions on the basis of this analysis. It is concerned with the systematic collection of numerical data and its interpretation. This systematic collection of data distinguishes statistics from other kinds of information.

2.2.1 Variable and Its Types

It is worthwhile to define the term **'variable'**. Variables can be classified as qualitative or quantitative.

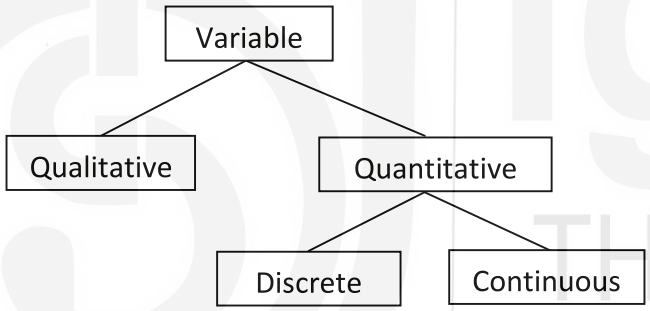
Qualitative variables are variables that can be placed into distinct categories, according to some characteristic or attribute. For example, if subjects are classified according to gender (male or female), then the variable gender is qualitative.

Quantitative variables are numerical and can be ordered or ranked. For example, the variable age is numerical, and people can be ranked in order according to the value of their ages. Other examples of quantitative variables are heights, weights, and body temperatures. Quantitative variables can be further classified into two groups: discrete and continuous.

Discrete variables can be assigned values such as 0, 1, 2, 3 and are said to be countable. Examples of discrete variables are the number of children in a family, the number of students in a classroom, and the number of calls received by a switchboard operator each day for a month. *Discrete variables assume values that can be counted*.

Continuous variables, by comparison, can assume an infinite number of values in an interval between any two specific values. Temperature, for example, is a continuous variable, since the variable can assume an infinite number of values between any two given temperatures. *Continuous variables can assume an infinite number of values between any two specific values.* They are obtained by measuring. They often include fractions and decimals. Variables may be classified as nominal, ordinal, or interval.

The classification of variables can be summarised as follows:



In addition to being classified as qualitative or quantitative, variables can be classified by how they are categorised, counted, or measured.

2.2.2 Source of Data

Data is a set of values of qualitative or quantitative variables. Pieces of data are individual pieces of information. While the concept of data is commonly associated with scientific research, data is collected by a huge range of organisations and institutions, including businesses (e.g., sales data, revenue, profits, stock price), governments (e.g., crime rates, unemployment rates, literacy rates) and non-governmental organisations.

Primary and Secondary Source: Gathering data can be accomplished through a primary source (the researcher is the first person to obtain the data) or a secondary source (the researcher obtains the data that has already been collected by other sources, such as data disseminated in a scientific journal).

2.2.3 Types of Data

The data may be of three types: Time series, Cross-section and Pooled.

1) **Time Series Data:** It is a set of observations on the values that a variable takes at different times. Such data may be collected at regular time intervals such as daily (i.e. consumer price index etc.) quarterly (i.e. GDP), annually (i.e. government budget etc.).

2) **Cross Section Data:** Cross-section data are data on one or more variables collected at the same point of time. For example, the data on the census of population collected by the Registrar General of India.

3) **Pooled Data:** In pooled data, the elements of both time series and cross section are clubbed. For example, over a period of time say from 2000 to 2013, we have data on saving, investment and GDP across Indian states.

4) **Panel, Longitudinal or Micro Panel Data:** This is a special type of pooled data in which the same cross-sectional (say a family or firm) is surveyed overtime.

2.2.4 Measurement Scales

- The first level of measurement is called the nominal level of measurement. A sample of college instructors classified according to subject taught (e.g., English, history, psychology, or mathematics) is an example of nominal-level measurement. Classifying survey subjects as male or female is another example of nominal-level measurement. No ranking or order can be placed on the data. The nominal level of measurement classifies data into mutually exclusive (non-overlapping), exhausting categories in which no order or ranking can be imposed on the data.

- The next level of measurement is called the *ordinal level.* Data measured at this level can be placed into categories, and these categories can be ordered, or ranked. For example, from student evaluations, guest speakers might be ranked as superior, average, or poor. *Note that precise measurement of differences in the ordinal level of measurement* does not *exist.* Such categories could be converted to a numerical scale for further analysis.

- The third level of measurement is called the interval level. This level differs from the ordinal level in that precise differences do exist between units. For example, many standardised psychological tests yield values measured on an interval scale. IQ is an example of such a variable. There is a meaningful difference of 1 point between an IQ of 109 and an IQ of 110. Temperature is another example of interval measurement, since there is a meaningful difference of 1°F between each unit, such as 72 and 73°F. One property is lacking in the interval scale: There is no true zero. For example, IQ tests do not measure people who have no intelligence. For temperature, 0°F does not mean no heat at all.

- The final level of measurement is called the ratio level. Examples of ratio scales are those used to measure height, weight, area, and number of phone calls received. Ratio scales have differences between units (1 inch, 1 pound, etc.) and a true zero. In addition, the ratio scale

contains a true ratio between values. For example, if one person can lift 200 pounds and another can lift 100 pounds, then the ratio between them is 2 to 1. Put another way, the first person can lift twice as much as the second person. The ratio level of measurement possesses all the characteristics of interval measurement, and there exists a true zero. In addition, true ratios exist when the same variable is measured on two different members of the population.

The definition and classification of variables assumes significance as it will determine the techniques to be used in univariate, bivariate or multivariate data analysis.

2.2.5 Classification of Data

After data collection we face the problem of arranging them into a format from which we will be able to draw some conclusions. The arrangement on the basis of some similarities is known as classification. The classification reduces trouble and complexities involved in working with raw scores by grouping them into some classes, which provides a comprehensive insight into the data. It presents data in a condensed form so that they can be easily analysed in an economical way with due precision; explain the affinities and diversities of data; and facilitate comparisons.

It can be said that classification is an essential and foremost step that enables the investigator to summarise the results in a meaningful and convenient way.

You have been introduced to the art and craft of classification in Unit 6. There we did talk about the arrays, discrete and continuous frequency distributions, grouped frequency distributions. In the present unit, we are introducing you to the techniques of summarising the essential features of a given data set.

2.3 DESCRIPTIVE STATISTICS

Most of the observations in this universe are subject to variability, especially observations related to human behaviour. In order to make a sensible decision on the basis of such data, it is necessary to express them in a precise manner, that is, the observations need to be expressed as a single estimate which summarises the observations. 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.

The branch of statistics devoted to the summarisation and description of data, called descriptive statistics, includes the construction of graphs, charts, and tables, and the calculation of various descriptive measures such as averages, measures of variation, and percentiles etc. It is thus clear that

descriptive statistics is much more than just the tabulation of numbers and the graphical presentation of these tabulated numbers.

Descriptive statistics are used to describe the basic features of the data in a study. With descriptive statistics we simply describe what is or what the data shows. Descriptive statistics help us to simplify large amounts of data in a sensible way. Each descriptive statistic reduces lots of data into a simpler summary. For instance, consider a simple number used to summarise how well a batter is performing in baseball, the batting average. This single number is simply the number of hits divided by the number of times at bat (reported to three significant digits). A batter who is hitting .333 is getting a hit one time in every three at bats. One batting .250 is hitting one time in four. The single number describes a large number of discrete events. Or, consider the Grade Point Average (GPA), which is a single number describes the general performance of a student across a potentially wide range of course experiences. Descriptive statistics are most often used to examine:

- **Central tendency** (location) of data, i.e. where data tend to fall, as measured by the mean, median, and mode.
- **Dispersion** (variability) of data, i.e. how spread out the data are, as measured by the variance and its square root, the standard deviation.
- **Skewness** (symmetry) of data, i.e. how concentrated data are at the low or high end of the scale, as measured by the skew index.
- **Kurtosis** (peaked-ness) of data, i.e. how concentrated data are around a single value, as measured by the kurtosis index.

Check Your Progress 1

1) Explain the term 'Univariate Data Analysis'?

.....

.....

.....

.....

2) Define Descriptive Statistics and its fundamental contribution towards data analysis?

.....

.....

.....

.....

3) Define the term 'quantitative variable' and its various levels of measurement ?

.....

.....

.....

.....

2.4 QUANTITATIVE DATA – PERCENTAGES AND MEASURES OF CENTRAL TENDENCY

2.4.1 Use of Percentages

A commonly used statistics, percentage expresses the information as a proportion of a whole. Percentages tend to be easy to interpret. For example, it is more understandable to say that 70% of the total students appeared have passed the examination instead of saying that 280 out of 400 have passed the exams. Percentages are also good way to show the relationships and comparisons – either between the categories of respondents or between categories of responses. Percentages are also useful when we want to show a frequency distribution of grouped data. Although computing the percentage appears to be a simple process, but following cautions need to be applied to avoid the possibility for making errors:

- *Use the correct base*: The base is the number from which the percentage is calculated. It is important to use the right base and indicate the base you have used. In above example, does 70% means 70% of all the students, or 70% of the students appeared in examination or the 70% of the students to whom the question was put? To elaborate further, for a certain query, 10 people did not respond, 70 answered 'yes' and 20 answered 'no'. If we use 100 as a base or divisor, we show 70% answered 'yes' but if we use 90 as the base who actually answered the question, it will be 78% who reported 'yes' as we do not know whether those who did not respond would have been 'yes' or 'no'.
- *Adding percentages*: percentages are added only when the categories are mutually exclusive (do not overlap). This not the case in multiple choice questions where the respondent may select several answers. For example, in a question asking respondents indicate a source of information for use of fertilizers in wheat production; the respondent might use one or multiple answers like dealer, fellow farmer, advertisement in newspaper etc. These answers are not mutually exclusive there percentages should not be added.
- *Averaging percentages*: Avoid the error of adding percentages and taking an average of summed percentages. This will provide erroneous answer as reflected in the following table:

District	Projects Completed	Percent
A	31,000	74
B	8,000	60
C	12,000	65
D	26,000	75
E	11,000	50
F	28,000	72
TOTAL	**1,16,000**	**Mean 66 (incorrect)**

In this table, a researcher has wrongly reported an average of six district percentages. Rather, the percentage should have been calculated by dividing the completed projects (116000) by the total approved projects (165000) to obtain the total mean percentage, i.e. 70.3% which reports the district-wise completed projects out of 165000 projects sanctioned.

2.4.2 Measures of Central Tendency

When we work with numerical data, it seems apparent that in most set of data there is a tendency for the observed values to group themselves about some interior values; some central values seem to be the characteristics of the data. This phenomenon is referred to as central tendency. For a given set of data, the measure of location we use depends on what we mean by middle; different definitions give rise to different measures. We shall consider some more commonly used measures, namely arithmetic mean, median and mode. The formulas in finding these values depend on whether they are ungrouped data or grouped data.

The *mean,* also known as the *arithmetic average,* is found by adding the values of the data and dividing by the total number of values. For example, the mean of 3, 2, 6, 5, and 4 is found by adding 3 + 2 + 6 + 5 + 4 and dividing by 5; hence, the mean of the data is 20 / 5 = 4.

The **mean** is the sum of the values, divided by the total number of values. The symbol $\bar{X}$ represents the sample mean.

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} = \frac{\sum X}{n}$$

where *n* represents the total number of values in the sample.

For a population, the Greek letter μ(mu) is used for the mean.

$$\mu = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N} = \frac{\sum X}{N}$$

where *N* represents the total number of values in the population.

In statistics, Greek letters are used to denote parameters, and Roman letters are used to denote statistics.

The steps for finding the mean for grouped data are summarised in the following table:

Finding the Mean for Grouped Data

Step 1 : Make a table as shown

A	B	C	D
Class	**Frequency *f***	**Midpoint X_m**	***f*.X_m**

Step 2 : Find the midpoints of each class and place them in column C.

Step 3 : Multiply the frequency by the midpoint for each class, and place the product in column D.

Step 4 : Find the sum of the column D.

Step 5 : Divide the sum obtained in column D by the sum of the frequencies obtained in column B.

The formula for the mean is

$$\bar{X} = \frac{\sum f.X_m}{n}$$

[**Note:** The symbol $\sum f.X_m$ mean to find the sum of the product of the frequency (f) and the midpoint (X_m) for each class.]

For calculating simple mean, we suppose that all the values or the sizes of items in the distribution have equal importance. But, in practical life this may not be so. In case some items are more important than others, a simple average computed is not representative of the distribution. Proper weightage has to be given to the various items. For example, to have an idea of the change in cost of living of a certain group of persons, the simple average of the prices of the commodities consumed by them will not do because all the commodities are not equally important, e.g. rice, wheat and pulses are more important than tea, confectionery etc., It is the weighted arithmetic average which helps in finding out the average value of the series after giving proper weight to each group. It is defined as the average whose component items are being multiplied by certain values known as "weights" and the aggregate of the multiplied results are being divided by the total sum of their "weight".

If x_1, x_2…x_n be the values of a variable x with respective weights of w_1, w_2… w_n assigned to them, then

Weighted A.M. = $\overline{x_w} = \frac{w_1x_1+w_2x_2+\cdots+w_nx_n}{w_1+w_2+\cdots+w_n} = \frac{\sum w_1x_1}{\sum w_1}$

Median is the value in the middle of the data set when the measurements are arranged in order of magnitude. For example, if 11 individuals were weighed and their weights arranged in ascending or descending order, the sixth value is the median since five values fall both above and below the sixth value. Median family income is often used in statistics because this value represents the exact middle of the data better than the mean. Fifty percent of families would have incomes above or below the median. In case of the number of measurements is even, the median is obtained by taking the average of the middle. In terms of formula:

$$\textbf{Median = Md= } \left(\frac{n+1}{2}\right)^{\text{th}} \textbf{ item}$$

In a grouped data, values are associated with frequencies. Grouping can be in the form of a discrete frequency distribution or a continuous frequency distribution. Whatever may be the type of distribution, cumulative frequencies have to be calculated to know the total number of items. Cumulative frequency of each class is the sum of the frequency of the class and the frequencies of the pervious classes, i.e. adding the frequencies successively, so that the last cumulative frequency gives the total number of items. The median in a grouped data is calculated using the following steps:

Step 1: Find cumulative frequencies.

Step 2: Find $\left(\frac{N}{2}\right)$

Step 3: See in the cumulative frequency the value first greater than $\left(\frac{N}{2}\right)$. Then the corresponding class interval is called the Median class. Then apply the formula

Median = $l + \frac{\frac{N}{2} - m}{f} \times c$

Where l = Lower limit of the median class

m = cumulative frequency preceeding the median

c = width of the median class

f = frequency in the median class

N = Total frequency

Note: If the class intervals are given in inclusive type convert them into exclusive type and call it as true class interval and consider lower limit in this.

Mode is the value occurring most often in the data. According to Croxton and Cowden *"The mode of a distribution is the value at the point around which the items tend to be most heavily concentrated. It may be regarded at the most typical of a series of values"*. It shows the centre of concentration of the frequency in around a given value. Therefore, where the purpose is to know the point of the highest concentration, it is preferred. It is, thus, a positional measure. If the largest group of people in a sample measuring age were 25 years old, then 25 would be the mode. The mode is the least commonly used measure of central tendency, particularly in large data sets. However, the mode is still important for describing a data set, especially when more than one value occurs frequently. In this instance, the data would be described as bimodal or multimodal, depending on whether two or more values occur frequently in the data set. In a continuous data, you first find out the highest frequency then the corresponding value of class interval is called the modal class. Then apply the formula to calculate the mode:

$$\textbf{Mode} = M_0 = l + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times c$$

l = Lower limit of the model class

$\Delta_1 = f_1 - f_0$

$\Delta_2 = f_1 - f_2$

f_1 = frequency of the modal class

f_0 = frequency of the class preceding the modal class

f_2 = frequency of the class succeeding the modal class

$$\text{Mode} = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times c$$

Remarks:

1) If $(2f_1 - f_0 - f_2)$ comes out to be zero, then mode is obtained by the following formula taking absolute differences within vertical lines.

2) $M_0 = l + \frac{(f_1 - f_0)}{|f_1 - f_0| + |f_1 - f_2|} \times c$

3) If mode lies in the first class interval, then f_0 is taken as zero.

2.4.3 Empirical Relationship between Averages

In a symmetrical distribution the three simple averages mean = median = mode. For a moderately asymmetrical distribution, the relationship between them is brought by Prof. Karl Pearson as Mode = 3 median – 2 mean.

2.4.4 Which Measure to Choose?

Researchers and statisticians must know which measure of central tendency is being used and when to use each measure of central tendency. The properties and uses of the measures of central tendency are summarised below, highlighting their usages in a given situation:

The Mean

- The mean is found by using all the values of the data.
- The mean varies less than the median or mode when samples are taken from the same population and all three measures are computed for these samples.
- The mean is used in computing other statistics, such as the variance.
- The mean for the data set is unique and not necessarily one of the data values.
- Mean, especially arithmetic mean, is most suitable for algebraic treatment like calculating combined means of two or more than two series.
- The mean cannot be computed for the data in a frequency distribution that has an open-ended class.
- The mean is affected by extremely high or low values, called outliers, and may not be the appropriate average to use in these situations.

The Median

- The median is used to find the center or middle value of a data set.
- The median is used when it is necessary to find out whether the data values fall into the upper half or lower half of the distribution.
- The median is used for an open-ended distribution.
- The median is affected less than the mean by extremely high or extremely low values.

The Mode

- The mode is used when the most typical case is desired.
- The mode is the easiest average to compute.
- The mode can be used when the data are nominal, such as religious preference, gender or political affiliation.

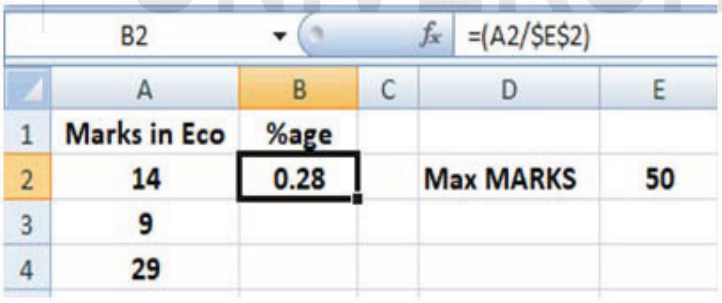
- The mode is not always unique. A data set can have more than one mode, or the mode may not exist for a data set.

Activity 1: Measure of Central Tendency

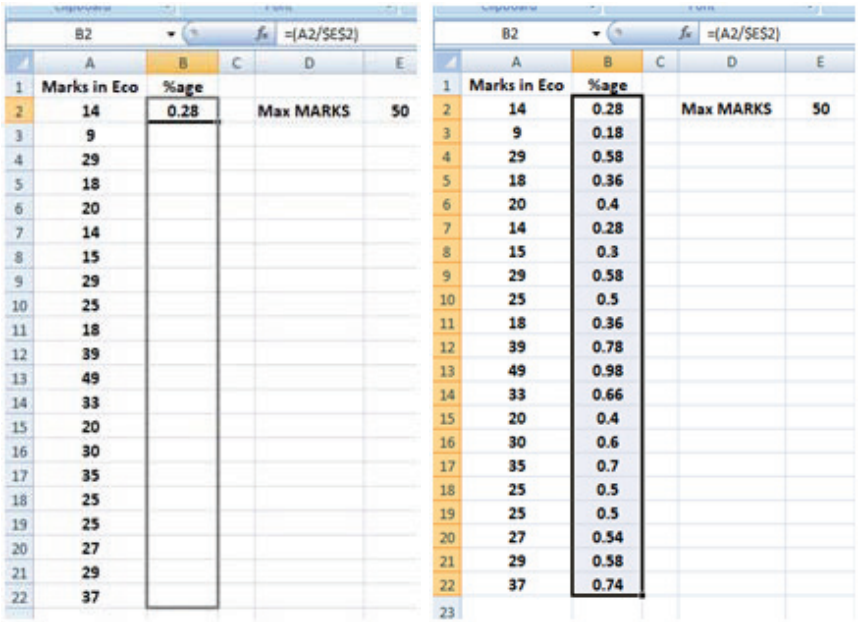
The marks of 21 students in Economics subjects are 14, 9, 29, 18, 20, 14, 15, 29, 25, 18, 39, 49, 33, 20, 30, 35, 25, 25, 27, 29, 37 (Marks are from a Maximum of 50 marks). Find the percentage of marks of each student, mean, median and mode for the data using a spreadsheet package.

How to do it?

- Insert the Marks in the spreadsheet in the cells A2 to A22. In cell A1 you may insert the heading "Marks in Eco".
- To calculate % of marks for each student we require maximum marks of the subject (in our example it is 50). To calculate %, you can simply divide the marks obtained by the Maximum marks and format the cell to percentage format. In this example, the Maximum Marks are stored in a cell which can be used for the division. This has been done to demonstrate the use of absolute addressing. Therefore, enter text "Max Marks" in cell D2 and 50 in cell E2. Also enter text "%age" in cell B1.
- You will calculate %age of marks in cells A2 to A22 into cells B2 to B22 respectively. To do so perform the following steps:
  - First enter the formula ***= (A2/$E$2)*** in cell B2 and press Enter key. The value 0.28 will be displayed in cell B2 and the formula bar will display the entered formula (refer to following figure). Please note that cell A2 contains marks of the first student, it is divided by Maximum marks which is stored in cell E2. But in the formula cell *E2* has been written as *$E$2.* Why? This is to make sure that when you copy the formula from cell B2 to B3 to B22, then this cell will remain the same that is E2 only. This is called absolute addressing.



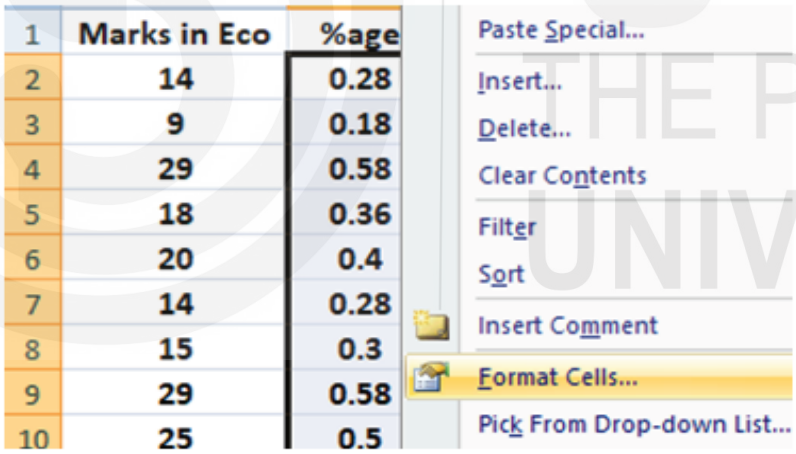
  - To copy the formula from Cell B2 to cells B3 to B22, bring your mouse pointer to Right-bottom side of cell B2. It will convert to a plus sign. Press the mouse button at this point and drag it down till cell B22 as shown in the following figure:



	A	B	C	D	E
1	Marks in Eco	%age			
2	14	0.28		Max MARKS	50
3	9				
4	29				
5	18				
6	20				
7	14				
8	15				
9	29				
10	25				
11	18				
12	39				
13	49				
14	33				
15	20				
16	30				
17	35				
18	25				
19	25				
20	27				
21	29				
22	37				

	A	B	C	D	E
1	Marks in Eco	%age			
2	14	0.28		Max MARKS	50
3	9	0.18			
4	29	0.58			
5	18	0.36			
6	20	0.4			
7	14	0.28			
8	15	0.3			
9	29	0.58			
10	25	0.5			
11	18	0.36			
12	39	0.78			
13	49	0.98			
14	33	0.66			
15	20	0.4			
16	30	0.6			
17	35	0.7			
18	25	0.5			
19	25	0.5			
20	27	0.54			
21	29	0.58			
22	37	0.74			
23					

- Release the Mouse and you will find the screen changes as shown above in the right. Please check the formula in different cells. For example, the formula in cell B10 will be ***=(A10/$E$2)***
- You may notice that all the data shown in column B2 to B22 is in decimal. To convert it into percentage, select cells B2 to B22 and click the Right button on your mouse. From the resulting Menu Select "Format Cells…" option as shown below:



- A dialog box for Format Cells will be displayed. Select the format "Percentage" from the options. You may select decimal places to 0 and press the OK button.

- To find the Mean, Median and mode, first you can define a name for the range that contains the student's marks. This range is (A2 to A22). This can be done by performing the following steps:
  - Select the Range A2 to A22 using clicking and holding mouse left button on cell A2 and while holding the left button scroll down the mouse pointer till A22. This range will be highlighted. (Refer to Fig. 2.4.4.1)

- Select the tab *Formula*s and Select *Define Name* from the Ribbon and in out a name EcoMarksMax50 in the resulting New Name dialog box (refer to Fig. 2.4.4.1)
- Press Ok button in the New Name dialog box. You have now created a new Range Name EcoMarksMax50 for the Range A2..A22. The range is being shown as *Refers to =Sheet1!$A$2:$A$22*in the New Name dialog box, as it follows absolute addressing.

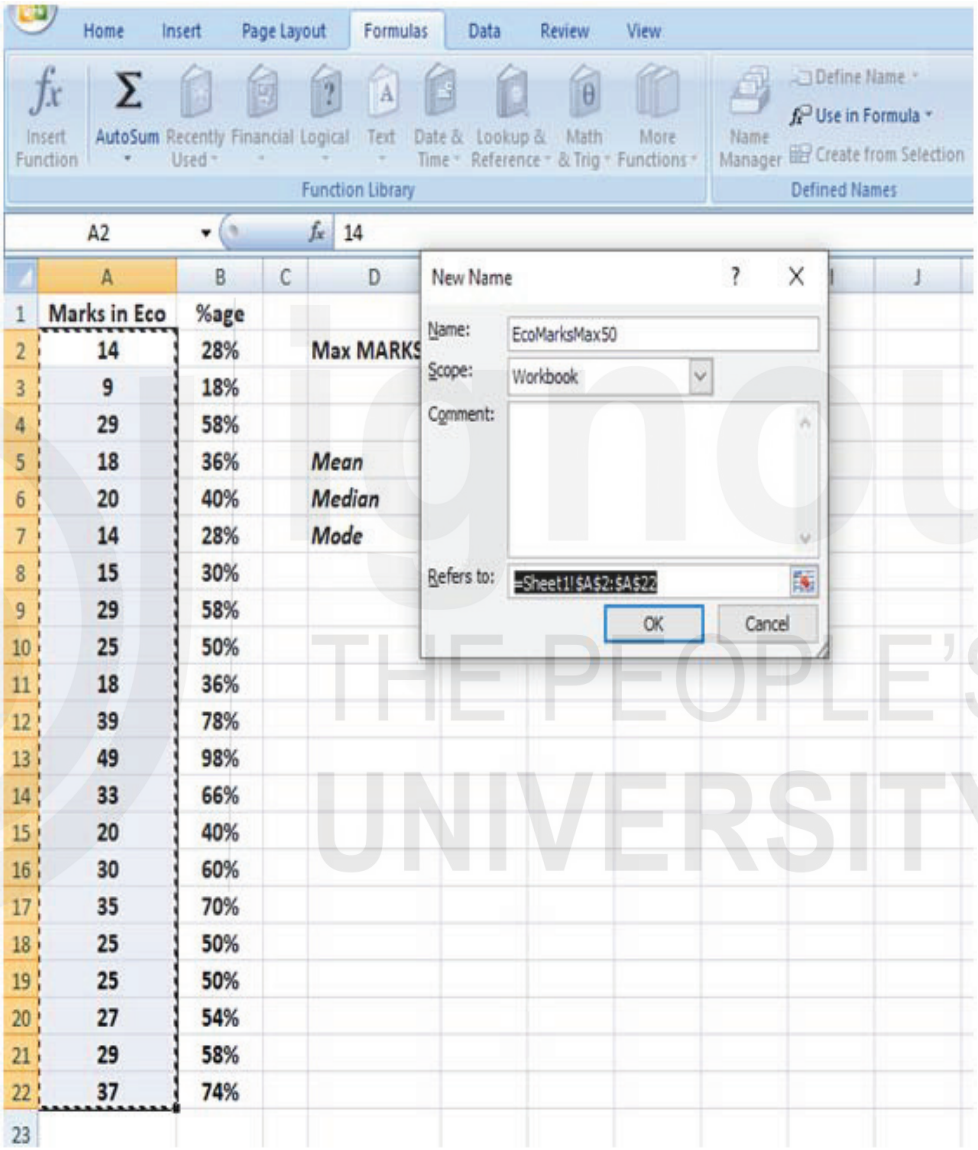
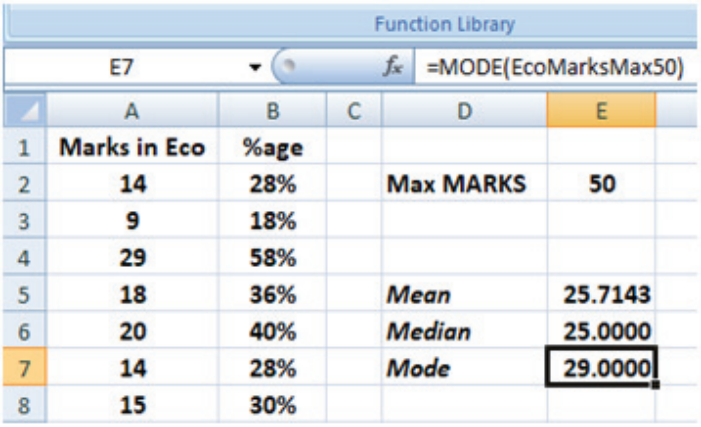


Fig. 2.4.4.1: Defining Name to a Range

You can now insert Labels - *Max MARKS, Mean, Median* and *Mode* as shown in the figure above. We have entered maximum marks value (50) in cell ***E2.***

Now enter the formulas for Mean: **=AVERAGE(EcoMarksMax50)**in Cell ***E5;*** and enter the formulas for Median: **=MEDIAN(EcoMarksMax50)**in Cell ***E6;*** and enter the formulas for Mode: **=Mode(EcoMarksMax50)**in Cell ***E7.***

You will be seeing the calculated values on the screen:



	A	B	C	D	E
1	Marks in Eco	%age			
2	14	28%		Max MARKS	50
3	9	18%			
4	29	58%			
5	18	36%		*Mean*	25.7143
6	20	40%		*Median*	25.0000
7	14	28%		*Mode*	29.0000
8	15	30%			

Check Your Progress 2

1) Define the term 'percentage' and precautions to be taken in its calculation and usage?

……

……

……

……

……

2) Which are the central measures of a univariate data series?

……

……

……

……

……

3) Explain the specific situations where a specific measure of central tendency could be used?

……

……

……

……

……

2.5 QUANTITATIVE DATA – MEASURES OF DISPERSION

The measures of central tendency serve to locate the center of the distribution, but they do not reveal how the items are spread out on either side of the center. This characteristic of a frequency distribution is commonly referred to as dispersion. In a series all the items are not equal. There is difference or variation among the values. The degree of variation is evaluated by various measures of dispersion. Small dispersion indicates high uniformity of the items, while large dispersion indicates less uniformity. An

ideal measure of dispersion should be rigidly defined, based on all the items along with simple to understand and easy to compute, not to be unduly affected by extreme items and capable of lending itself for algebraic manipulation.

The various absolute and relative measures of dispersion are listed below.

Absolute measure (indicates the amount of variation in terms of units of observations)	Relative measure (Unit free)
Range	Co-efficient of Range
Mean deviation	Co-efficient of Mean deviation
Standard deviation	Co-efficient of variation

2.5.1 Range and Coefficient of Range

Range is the simplest possible measure of dispersion and is defined as the difference between the largest and smallest values of the variable. In symbols, Range = L – S, where L = Largest value and S = Smallest value. In individual observations and discrete series, L and S are easily identified. In continuous series, either L = Upper boundary of the highest class and S = Lower boundary of the lowest class or L = mid value of the highest class and S = mid value of the lowest class.

Co-efficient of Range:

Co-efficient of Range = $\frac{L-S}{L+S}$

2.5.2 Mean Deviation and Coefficient of Mean Deviation

Mean deviation is the arithmetic mean of the deviations of a series computed from any measure of central tendency; i.e., the mean, median or mode, all the deviations are taken as positive i.e., signs are ignored. According to Clark and Schekade, “Average deviation is the average amount scatter of the items in a distribution from either the mean or the median, ignoring the signs of the deviations”. We usually compute mean deviation about any one of the three averages mean, median or mode. Sometimes mode may be ill defined and as such mean deviation is computed from mean and median. Median is preferred as a choice between mean and median. But in general practice and due to wide applications of mean, the mean deviation is generally computed from mean. M.D can be used to denote mean deviation.

In an individual series, steps involved in calculating the mean deviation are mentioned below:

1) Calculate the average mean, median or mode of the series.
2) Take the deviations of items from average ignoring signs and denote these deviations by $|D|$.
3) Compute the total of these deviations, i.e., $\sum|D|$.
4) Divide this total obtained by the number of items.

Symbolically: M.D. = $\frac{\sum|D|}{n}$

In case of discrete or continuous series of grouped data, multiply the deviation of each value by its respective frequency and find out the total $\sum f|D|$ and divide it by *n* to find out the MD.

Coefficient of mean deviation: Mean deviation calculated by any measure of central tendency is an absolute measure. For the purpose of comparing variation among different series, a relative mean deviation is required. The relative mean deviation is obtained by dividing the mean deviation by the average used for calculating mean deviation, i.e.

$$\text{Coefficient of mean deviations} = \frac{\text{Mean deviation}}{\text{Mean or Median or Mode}}$$

If the result is desired in percentage, the coefficient of mean deviation =

$$\frac{\text{Mean deviation}}{\text{Mean or Median or Mode}} \times 100$$

2.5.3 Standard Deviation and Variance

Karl Pearson introduced the concept of standard deviation in 1893. It is the most important measure of dispersion and is widely used in many statistical formulae. Standard deviation is also called Root-Mean Square Deviation. The reason is that it is the square root of the mean of the squared deviation from the arithmetic mean. It provides accurate result. As a parameter for a population, the standard deviation is denoted by the Greek letter σ (sigma). Square of standard deviation is called Variance.

The steps involved in calculating the standard deviation are indicated below:

1. Find out the actual mean of the series ($(\bar{X})$).
2. Find out the deviation of each value from the mean (x = X – $\overline{X}$).
3. Square the deviations and take the total of squared deviations $\sum x^2$.
4. Divide the total ($\sum x^2$) by the number of observation $\left[\frac{\sum x^2}{n}\right]$

The square root of $\left[\frac{\sum x^2}{n}\right]$ is standard deviation.

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

To calculate the variance and standard deviation for grouped, firs step is to make the table shown below and fill up columns as suggested:

	A	B	C	D	E	F
	Class	Frequency (f)	Mid-point X_m	d = X_m – (Assumed Mean)	f.d	$f.d^2$

Taking deviations from fractional value would be a very difficult and tedious task. To save time and labour, you may apply short-cut method in which deviations are taken from an assumed mean. The formula is:

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

Where d = X – A, and A is assumed mean, i.e. any value from distribution could be picked up as an assumed mean. In case of grouped data, d needs to be multiplied by the corresponding frequencies.

Sample Variance and Standard Deviation:

When computing the variance for a sample, formula mentioned above is not usually used, as in most cases the purpose of calculating the statistic is to estimate the corresponding parameter. For example, the sample mean is used to estimate the population mean. The expression does not give the best estimate of the population variance because when the population is large and the sample is small (usually less than 30), the variance computed by this formula usually underestimates the population variance. Therefore, instead of dividing by n, find the variance of the sample by dividing by n – 1, giving a slightly larger value and an unbiased estimate of the population variance.

The forumula for the sample variance, denoted by s^2, is

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

Where

$\bar{X}$ = Sample mean

n = Sample size

In case of standard deviation or variance, few things need to be noted:

- First, the square root of the variance gives the standard deviation; and vice versa, squaring the standard deviation gives the variance.
- Second, the variance is actually the average of the square of the distance that each value is from the mean. Therefore, if the values are near the mean, the variance will be small. In contrast, if the values are far from the mean, the variance will be large.
- Third, main reason for taking squared distances instead of the actual distances is that the sum of the distances from actual mean will always be zero. When each value is squared, the negative signs are eliminated.
- Fourth, why is it necessary to take the square root? The reason for taking the square root is that since the distances were squared, the units of the resultant numbers are the squares of the units of the original raw data. Finding the square root of the variance puts the standard deviation in the same units as the raw data.
- Finally, when you are finding the square root, always use its positive or principal value, since the variance and standard deviation of a data set can never be negative.

2.5.4 Coefficient of Variation

The Standard deviation is an absolute measure of dispersion. It is expressed in terms of units in which the original figures are collected and stated. The

standard deviation of heights of students cannot be compared with the standard deviation of weights of students, as both are expressed in different units, i.e. heights in centimeter and weights in kilograms. Therefore the standard deviation must be converted into a relative measure of dispersion for the purpose of comparison. The relative measure is known as the coefficient of variation. The coefficients of variation is obtained by dividing the standard deviation by the mean and multiply it by 100. Symbolically,

$$\text{Coefficient of variation (C.V.)} = \frac{\sigma}{\bar{X}} \times 100$$

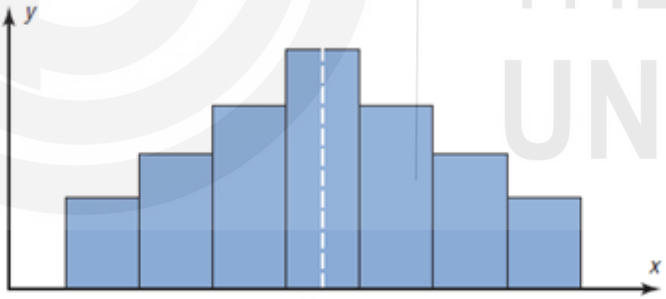
If we want to compare the variability of two or more series, we can use C.V. The series or groups of data for which the C.V. is greater indicate that the group is more variable, less stable, less uniform, less consistent or less homogeneous. If the C.V. is less, it indicates that the group is less variable, more stable, more uniform, more consistent or more homogeneous.

2.5.5 Distribution Shapes

2.5.5.1 Skewness

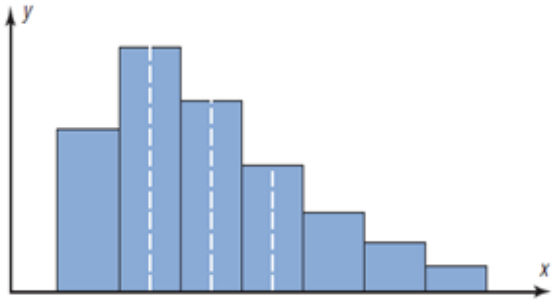
Skewness means 'lack of symmetry'. We study skewness to have an idea about the shape of the curve which we can draw with the help of the given data. If in a distribution mean = median = mode, then that distribution is known as symmetrical distribution. If in a distribution mean, median and mode are not equal, then it is not a symmetrical distribution and it is called a skewed distribution and such a distribution could either be positively skewed or negatively skewed.

Symmetrical distribution:



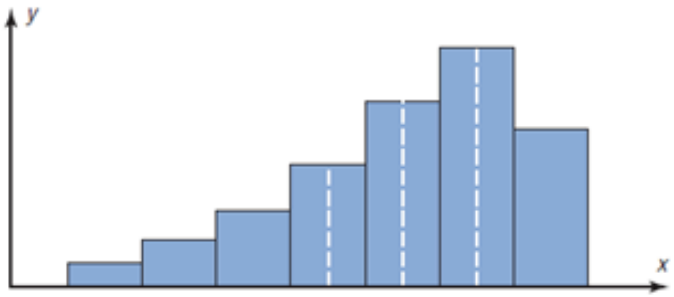
Mean
Median
Mode

It is clear from the above diagram that in a symmetrical distribution the values of mean, median and mode coincide. The spread of the frequencies is the same on both sides of the center point of the curve.



Mode Median Mean

(a) Positively Skewed or right-skewed



Mean Median Mode

(b) Negatively Skewed or left-skewed

It is clear from the above diagrams that in both positively and negatively skewed distributions, the value of the mode is highest and that of the mean is least, the median lies in between the two. In the positively skewed distribution the frequencies are spread out over a greater range of values on the right hand side than they are on the left hand side. In the negatively skewed distribution the frequencies are spread out over a greater range of values on the left hand side than they are on the right hand side.

Measures of skewness:

The important measures of skewness are Karl Pearson's coefficient of skewness and Bowley's coefficient of skewness.

According to Karl Pearson, the absolute measure of skewness = mean – mode. This measure is not suitable for making valid comparison of the skewness in two or more distributions because the unit of measurement may be different in different series. To avoid this difficulty use relative measure of skewness called Karl Pearson's coefficient of skewness given by:

$$\text{Karl Pearson's Coefficient of Skewness} = \frac{\text{Mean}-\text{Mode}}{\text{S.D.}}$$

In case of mode is ill-defined, the coefficient can be determined by the formula:

$$\text{Coeffiecient of skewness} = \frac{3\,(\text{Mean}-\text{Median})}{\text{S.D.}}$$

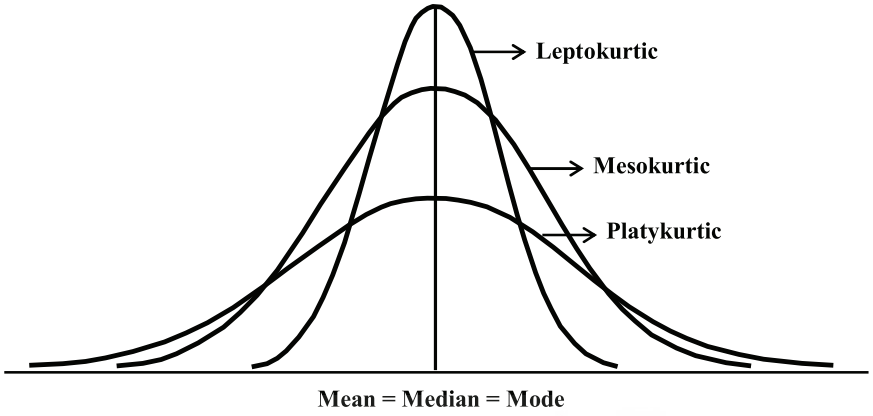
In Karl Pearson's method of measuring skewness, the whole of the series is needed. Prof. Bowley has suggested a formula based on relative position of quartiles. In a symmetrical distribution, the quartiles are equidistant from the value of the median; ie., Median – Q_1 = Q_3 – Median. But in a skewed distribution, the quartiles will not be equidistant from the median. Hence Bowley has suggested the following formula:

$$\text{Bowley's Coefficient of skewness (sk)} = \frac{Q_3+Q_1-2\,\text{Median}}{Q_3-Q_1}$$

2.5.5.2 Kurtosis

Kurtosis is another measure of the shape of a distribution. Whereas skewness measures the lack of symmetry of a distribution around the measure of central tendency, kurtosis is a measure of the relative peakedness or flatness of a distribution compared to the normal

distribution. In other words, Kurtosis is a parameter that measures the degree to which a curve of the frequency distribution is peaked or flat-topped. Various frequency curves can be divided into three categories depending upon the shape of their peak. The three shapes are termed as Leptokurtic, Mesokurtic (also called the Normal distribution) and Platykurtic as shown in the Figure below.



Measure of Kurtosis:

A measure of kurtosis is given by Karl Pearson who defined a distribution's degree of Kurtosis as $\eta = \beta_2 - 3$, where β_2 is given by $\frac{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^4}{\left[\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2\right]^2}$, where N is the size of the population or the sample size in sampling distribution; μ represents the population mean, the sample counterpart of which is given by the sample mean ($\bar{X}$). According to the above formula, the three types of curves can be defined as:

1) Mesokurtic (or Normal) curve: A curve which is neither flat nor peaked, such as given by the Normal distribution. For such a type of curve, we have β_2= 3 and thus η = 0.
2) Leptokurtic curve: A curve which is more peaked than the normal curve. For such a type of curve, we have β_2> 3 and thus η > 0.
3) Platykurtic curve: A curve which is relatively flatter than the normal curve. For such a type of curve, we have β_2< 3 and thus η < 0.

Check Your Progress 3

1) Explain the terms 'standard deviation' and 'variance', their relationship and their usage?

   ……

   ……

   ……

2) Which measure of dispersion ensures highest degree of reliability and why?

   ……

   ……

   ……

   ……

3) For a negatively skewed distribution what is relationship between mean, mode, median and quartiles?

……

……

……

……

4) How will you infer about the shape of the distribution by using skewness and kurtosis?

……

……

……

2.6 QUANTITATIVE DATA – MEASURES OF POSITION

In addition to measures of central tendency and measures of variation, there are measures of position or location. These measures include standard scores, percentiles, deciles, and quartiles. They are used to locate the relative position of a data value in the data set. For example, if a value is located at the 80th percentile, it means that 80% of the values fall below it in the distribution and 20% of the values fall above it. The *median* is the value that corresponds to the 50th percentile, since one-half of the values fall below it and one half of the values fall above it. This section discusses these measures of position.

2.6.1 Standard Scores

Suppose that a student scored 90 on a music test and 45 on an English exam. Direct comparison of raw scores is impossible, since the exams might not be equivalent in terms of number of questions, value of each question, and so on. However, a comparison of a relative standard similar to both can be made. This comparison uses the mean and standard deviation and is called a standard score or *z* score. A standard score or *z* score tells how many standard deviations a data value is above or below the mean for a specific distribution of values. If a standard score is zero, then the data value is the same as the mean. It is to be noted that if the z score is positive, the score is above the mean. If the z score is 0, the score is the same as the mean. And if the z score is negative, the score is below the mean.

A z score or standard score for a value is obtained by subtracting the mean from the value and dividing the result by the standard deviation. The symbol for a standard score is z. The formula is

$$z = \frac{\text{Value} - \text{mean}}{\text{Standard deviation}}$$

For samples, the formula is

$$z = \frac{X - \bar{X}}{s}$$

For populations, the formula is

$$z = \frac{X - \mu}{\sigma}$$

The z score represents the number of standard deviations that a data value falls above or below the mean.

Example:

A student scored 65 on a calculus test that had a mean of 50 and a standard deviation of 10; she scored 30 on a history test with a mean of 25 and a standard deviation of 5. Compare her relative positions on the two tests.

Solution:

First, find the z scores. For calculus the z score is

$$z = \frac{X - \bar{X}}{s} = \frac{65 - 50}{10} = 1.5$$

For history the z score is

$$z = \frac{30 - 25}{5} = 1.0$$

Since the z score for calculus is larger, her relative position in the calculus class is higher than her relative position in the history class.

2.6.2 Percentiles

Percentiles are position measures used in educational and health-related fields to indicate the position of an individual in a group. Percentiles divide the data set into 100 equal groups.

Percentiles are symbolised by

$P_1, P_2, P_3, \ldots\ldots, P_{99}$

And divide the distribution into 100 groups.



Percentiles are not the same as percentages. That is, if a student gets 72 correct answers out of a possible 100, she obtains a percentage score of 72. There is no indication of her position with respect to the rest of the class. She could have scored the highest, the lowest, or somewhere in between. On the other hand, if a raw score of 72 corresponds to the 64th percentile, then she did better than 64% of the students in her class.

Several mathematical methods exist for computing percentiles for data. These methods can be used to find the approximate percentile rank of a data value or to find a data value corresponding to a given percentile. When the data set is large (100 or more), these methods yield better results.

$$Percentile = \frac{(number\ of\ values\ below\ X) + 0.5}{total\ number\ of\ values} .100$$

Example:

A teacher gives a 20-point test to 10 students. The scores are shown here. Find the percentile rank of a score of 12.

18, 15, 12, 6, 8, 2, 3, 5, 20, 10

Arrange the data in order from lowest to highest.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

Then substitute into the formula.

$$Percentile = \frac{(number\ of\ values\ below\ X) + 0.5}{total\ number\ of\ values} .100\%$$

Since ther are six values below a score of 12, the solution is

$$Percentile = \frac{6+0.5}{10} .100\% = 65^{th} \text{ percentile}$$

Thus, a student whose score was 12 did better than 65% of the class.

Similarly, one can find the data value corresponding to a given percentile, the procedure for which is given below:

Finding a Data Value Corresponding to a Given Percentile.

Step 1: Arrange the data in order from lowest to highest.

Step 2: Substitute into the formula

$$c = \frac{n.p}{100}$$

Where

n = total number of values

p = percentile

Step 3A: If c is not a whole number, round up to the next whole number. Starting at the lowest value, count over to the number that corresponds to the rounded-up value.

Step 3B: If c is a whole number use the value halfway between the c^{th} and $(c + 1)^{st}$ values when counting up from the lowest value.

In the above example, find out the value corresponding to 25^{th} percentile

Step 1: Arrange the data in order from lowest to highest.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

Step 2: Compute

$$c = \frac{n.p}{100}$$

Where

n = total number of values

p = percentile

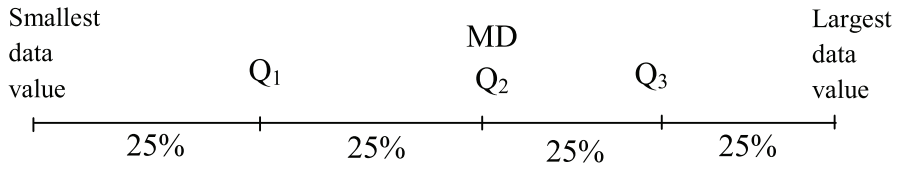
Thus,

$$c = \frac{10 \cdot 25}{100} = 2.5$$

Step 3: If c is not a whole number, round it up to the next whole number; in this case, c =3. (If c is a whole number, see Example 3 – 35). Start at the lowest value and count over to the third value, which is 5. Hence, the value 5 corresponds to the 25th percentile.

2.6.3 Quartiles and Deciles

Quartiles divide the distribution into four groups, separated by Q_1, Q_2, Q_3. Note that Q_1 is the same as the 25thpercentile; Q_2 is the same as the 50th percentile, or the median; Q_3 corresponds to the 75thpercentile, as shown:



Quartiles can be computed by using the above formula given for computing percentiles. For Q_1 use p = 25. For Q_2 use p = 50. For Q_3 use p = 75. However, an easier method for finding quartiles is given below:

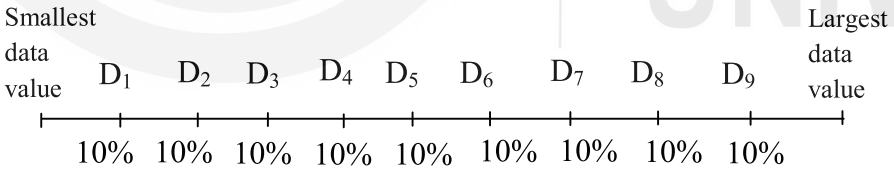
Step 1: Arrange the data in order from lowest to highest.

Step 2: Find the median of the data valuees. This is the value for Q_2.

Step 3: Find the median of the values that fall below Q_2. This is the value for Q_1.

Step 4: Find the median of the data valuees that fall above Q_2. This is the value for Q_3.

Deciles divide the distribution into 10 groups, as shown. They are denoted by D_1, D_2, etc.



Note that D_1 corresponds to P_{10}; D_2 corresponds to P_{20}; etc. Deciles can be found by using the formulas given for percentiles. Taken altogether then, these are the relationships among percentiles, deciles, and quartiles.

Deciles are denoted by D_1, D_2, D_3. . ..D_9, and they correspond to P_{10}, P_{20}, P_{30}, . . . , P_{90}.

Quartiles are denoted by Q_1, Q_2, Q_3 and they correspond to P_{25}, P_{50}, P_{75}. The median is the same as P_{50} or Q_2 or D_5.

Activity 2: Dispersion of Data

Find the Skewness, Kurtosis, Variance, Standard deviation, Quartiles for the data given in Activity 1.

How to do it?

By now you are familiar about how to enter formula. Fig. 2.6.1 shows the final output that has been generated for Activity 2.

E13 =STDEV(MarksInEcoOutOf50)

	A	B	D	E	G	H	I	J	K	M	N
1	Marks in Eco	%age			Z-Score	Percentile	Quartile	Marks		Sorted Marks	
2	14	28%	Max MARKS	50	-1.2043	5%	25%	18	P_{25}	9	Minimum
3	9	18%			-1.7183	0%	50%	25	P_{50}	14	
4	29	58%			0.3378	60%	75%	30	P_{75}	14	
5	18	36%	*Mean*	25.7143	-0.7931	20%				15	
6	20	40%	*Median*	25.0000	-0.5875	30%		Five Point Summary		18	
7	14	28%	*Mode*	29.0000	-1.2043	5%		*Minimum*	9	18	Quartile 1 at 25%
8	15	30%	Skewness	0.4429	-1.1015	15%		*Q1*	18	20	
9	29	58%	Kurtosis	0.1916	0.3378	60%		*Median*	25	20	
10	25	50%			-0.0734	40%		*Q3*	30	25	
11	18	36%			-0.7931	20%		*Maximum*	49	25	
12	39	78%			1.3659	95%				25	Median (Quartile 2) at 50%
13	49	98%	Std Dev	9.7270	2.3939	100%				27	
14	33	66%	Variance	94.6143	0.7490	80%				29	
15	20	40%			-0.5875	30%				29	
16	30	60%			0.4406	75%				29	
17	35	70%			0.9546	85%				30	Quartile 3 at 75%
18	25	50%			-0.0734	40%				33	
19	25	50%			-0.0734	40%				35	
20	27	54%			0.1322	55%				37	
21	29	58%			0.3378	60%				39	
22	37	74%			1.1602	90%				49	Maximum

Fig. 2.6.1: Output of Activity 2

Fig. 2.6.2(a) and (b) shows various formulas that have been entered to obtain Fig. 2.6.1

I1 Quartile

	D	E	G	H
1			Z-Score	Percentile
2	Max MARI	50	=STANDARDIZE(A2,$E$5,$E$13)	=PERCENTRANK(MarksInEcoOutOf50,A2)
3			=STANDARDIZE(A3,$E$5,$E$13)	=PERCENTRANK(MarksInEcoOutOf50,A3)
4			=STANDARDIZE(A4,$E$5,$E$13)	=PERCENTRANK(MarksInEcoOutOf50,A4)
5	*Mean*	=AVERAGE(MarksInEcoOutOf50)	=STANDARDIZE(A5,$E$5,$E$13)	=PERCENTRANK(MarksInEcoOutOf50,A5)
6	*Median*	=MEDIAN(MarksInEcoOutOf50)	=STANDARDIZE(A6,$E$5,$E$13)	=PERCENTRANK(MarksInEcoOutOf50,A6)
7	*Mode*	=MODE(MarksInEcoOutOf50)	=STANDARDIZE(A7,$E$5,$E$13)	=PERCENTRANK(MarksInEcoOutOf50,A7)
8	Skewness	=SKEW(MarksInEcoOutOf50)	=STANDARDIZE(A8,$E$5,$E$13)	=PERCENTRANK(MarksInEcoOutOf50,A8)
9	Kurtosis	=KURT(MarksInEcoOutOf50)	=STANDARDIZE(A9,$E$5,$E$13)	=PERCENTRANK(MarksInEcoOutOf50,A9)
10			=STANDARDIZE(A10,$E$5,$E$13)	=PERCENTRANK(MarksInEcoOutOf50,A10)
11			=STANDARDIZE(A11,$E$5,$E$13)	=PERCENTRANK(MarksInEcoOutOf50,A11)
12			=STANDARDIZE(A12,$E$5,$E$13)	=PERCENTRANK(MarksInEcoOutOf50,A12)
13	Std Dev	=STDEV(MarksInEcoOutOf50)	=STANDARDIZE(A13,$E$5,$E$13)	=PERCENTRANK(MarksInEcoOutOf50,A13)
14	Variance	=VAR(MarksInEcoOutOf50)	=STANDARDIZE(A14,$E$5,$E$13)	=PERCENTRANK(MarksInEcoOutOf50,A14)
15			=STANDARDIZE(A15,$E$5,$E$13)	=PERCENTRANK(MarksInEcoOutOf50,A15)
16			=STANDARDIZE(A16,$E$5,$E$13)	=PERCENTRANK(MarksInEcoOutOf50,A16)
17			=STANDARDIZE(A17,$E$5,$E$13)	=PERCENTRANK(MarksInEcoOutOf50,A17)
18			=STANDARDIZE(A18,$E$5,$E$13)	=PERCENTRANK(MarksInEcoOutOf50,A18)
19			=STANDARDIZE(A19,$E$5,$E$13)	=PERCENTRANK(MarksInEcoOutOf50,A19)
20			=STANDARDIZE(A20,$E$5,$E$13)	=PERCENTRANK(MarksInEcoOutOf50,A20)
21			=STANDARDIZE(A21,$E$5,$E$13)	=PERCENTRANK(MarksInEcoOutOf50,A21)
22			=STANDARDIZE(A22,$E$5,$E$13)	=PERCENTRANK(MarksInEcoOutOf50,A22)

Fig. 2.6.2 (a): Formulas of Column E, G and H

	I	J	K
1	**Quartile**	**Marks**	
2	*0.25*	*=PERCENTILE(MarksInEcoOutOf50,I2)*	P_{25}
3	*0.5*	*=PERCENTILE(MarksInEcoOutOf50,I3)*	P_{50}
4	*0.75*	*=PERCENTILE(MarksInEcoOutOf50,I4)*	P_{75}
5			
6		**Five Point Summary**	
7		*Minimum*	=MIN(MarksInEcoOutOf50)
8		*Q1*	=QUARTILE(MarksInEcoOutOf50,1)
9		*Median*	=QUARTILE(MarksInEcoOutOf50,2)
10		*Q3*	=QUARTILE(MarksInEcoOutOf50,3)
11		*Maximum*	=MAX(MarksInEcoOutOf50)
12			

Fig. 2.6.2 (b): Formulas of Column J and K

Please note that in the Fig. 2.6.2(a) and (b) the name for the Range A2..A22 is *Marks In Economics (out of 50).*

Please also note that in Fig. 2.6.1 Marks have been sorted in Column M so that various Quartiles can be shown.

Check Your Progress 4

1) Explain the term Measures of Position?

……

……

……

2) Which of the measures of position is widely used and why?

……

……

……

3) Explain the relationship between various Measures of Position?

……

……

……

2.7 LET US SUM UP

Various statistical concepts used in summarising the data have been discussed in this Unit. These include measures of central tendency, measures of variation or dispersion, and measures of position. The three most commonly used measures of central tendency are – mean, median, and mode. The three most commonly used measurements of variation are range, variance, and standard deviation. The most common measures of position are percentiles, quartiles, and deciles. Light has been thrown on how data values are distributed around the central value of the distribution and the empirical rule. The coefficient of variation is used to describe the

standard deviation in relationship to the mean. These methods are commonly called traditional statistical methods and are primarily used to confirm various conjectures about the nature of the data. The central tendency measures and dispersions are widely used in solving the practical problems in real life.

The Unit also enables us to get the results of percentage, mean, mode, median, skewness, kurtosis from the given data by using the spread sheet package (Excel).

2.8 REFERENCES

1) Blaikie, N. (2003). *Analyzing quantitative data: From description to explanation*. Sage.
2) Bohrnstedt, G. W., & Knoke, D. (1994). Statistics for social data analysis.
3) Bryman, A., & Cramer, D. (1994). *Quantitative data analysis for social scientists (rev*. Taylor & Frances/Routledge).
4) Cramer, D. (2003). *Advanced quantitative data analysis*. McGraw-Hill International.
5) Creswell, J. W. (2002). *Educational research: Planning, conducting, and evaluating quantitative data*. Prentice Hall.
6) Suen, H. K., & Ary, D. (2014). *Analyzing quantitative behavioral observation data*. Psychology Press.

2.9 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) See Section 2.2
2) See Section 2.3
3) See Section 2.3

Check Your Progress 2

1) See Sub-section 2.4.1
2) See Sub-section 2.4.2
3) See Sub-section 2.4.4

Check Your Progress 3

1) See Sub-section 2.5.3
2) See Sub-section 2.5.4
3) See Sub-section 2.5.5
4) Skewness is a measure of relative size of the two tails whereas kurtosis is a measure of combined size of the two tails.

Check Your Progress 4

1) See Section 2.6
2) See Sub-section 2.6.1 & 2.6.2
3) See Sub-section 2.6.3

UNIT 3 INTRODUCTION TO STATISTICAL SOFTWARE

Structure

3.1 Introduction
3.2 Need of Statistical Software
3.3 Data Handling
  3.3.1 Data Entry
  3.3.2 Validating Data
3.4 Use of Formula and Functions
3.5 Making Charts
3.6 Activating Data Analysis Tab
3.7 Let Us Sum Up
3.8 Exercises
3.9 References
3.10 Answers or Hints to Check Your Progress Exercises

3.1 INTRODUCTION

In the present era of information society, information is one of the major driving force for economic development. Information is created, presented in various visual forms, related with other form of information and then used for economic and decision making activities.

How do we create information? Well one of the key ingredient of creating information is Data. Data needs to be analysed and converted to simpler presentable forms like tables and graphs so that it can be utilised in a more meaningful manner. Thus, data is the raw form of information, which is to be collected, cleaned and processed to create meaningful, presentable information.

A statistical software is a tool that can help converting raw data into information. This Unit introduces you to basic features that must be supported by any statistical software that needs to be used for the purpose of data analysis. As we are discussing about the use of software, we must use one example to explain the concepts involved. For the purpose of this Unit, MS-Excel has been chosen as the vehicle to provide conceptual information in a more practical way. However, you are free to choose any other open source or proprietary software to learn these concepts. Please note that this Unit will not elaborate all the concepts used in Excel except the basic essential ones required from the point of view of data processing.

3.2 NEED OF STATISTICAL SOFTWARE

A statistical software is a tool that allows you to create and process data to produce meaningful and presentable form of data. The following are some of the basic functionalities that are provided by most statistical packages:

- an interface to enter and edit data in a simple and meaningful way with suitable headings.
- a simple way to import or export data from one software application to another.
- help in data definition and data validations.
- provide means to perform simple computation and operation on data such as sorting and filtering the data for analysis.
- support for building different kinds of charts easily.
- have functions to perform different statistical analysis functions.

There are large number of statistical packages available that can be used for performing data analysis These tools include specialised languages that has been primarily designed for data analysis - for example Python. Python is an Object Oriented (not pure) language that can be used for preparing data for analysis, data visualisation and data analysis.

In addition, a large number of Open Source software are also available for performing statistical analysis. R is a free statistical environment which has become one of the popular tool for statistical analysis. PSPP, Greti, MicroOsiris, GNU Octave and Dmelt and many more are open source tools for the purpose of statistical analysis.

There are a number of proprietary software packages that are very useful for statistical analysis. These include SPSS, EViews, SAS, MATLAB, MS Excel and many more are useful in statistical analysis.

We will be discussing more about MS-Excel in this Unit, however, you may use Libreoffice Calc to learn almost similar functionality.

3.3 DATA HANDLING

First step of any data analysis task is to create, edit/ validate data. We will explain it with the help of an activity which is a part of the next sub-section.

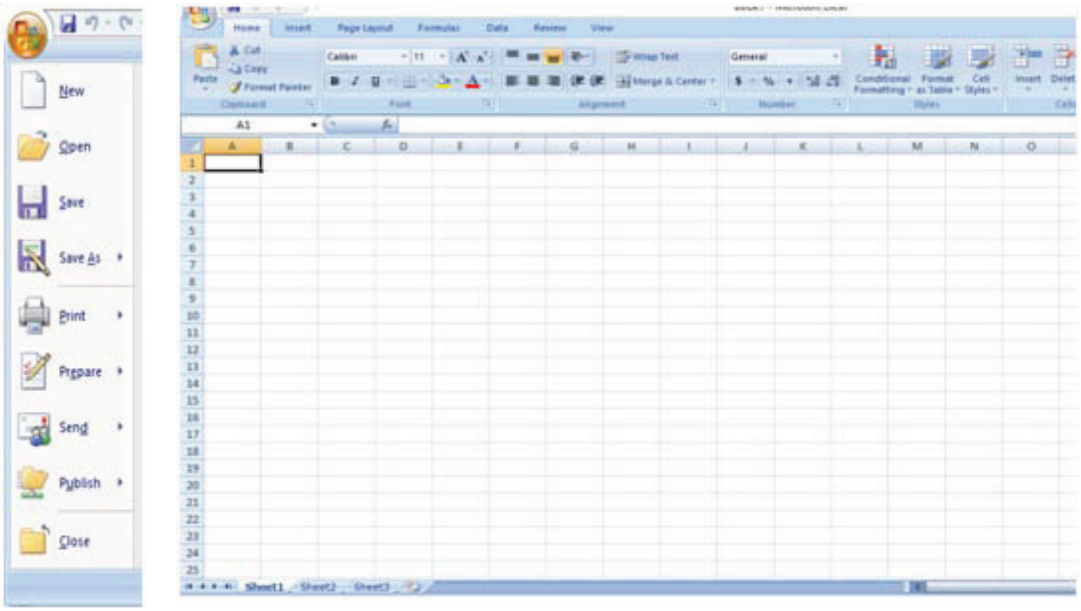
3.3.1 Data Entry

The following activity illustrates the process of simple data entry.

Activity 1: Input the Marks data of 20 students: 3, 30, 29, 17, 25, 58, 23, 48, 16, 25, 25, 29, 21, 23, 19, 22, 20, 29, 37, 32 in the Economics subject in an Excel worksheet.

How to do the activity?

To perform this activity, you must first Open the Excel software, Select the Office Button in the Top left corner of your screen and Select New Option as shown in Fig. 3.1(a).



(a) New Option **(b) A Blank Spread Sheet**

Fig. 3.1: Creating a New Worksheet

Please note that in Excel, intersection of Row and Column is called a cell. Also notice that in Fig. 3.1 (b) the selected Cell is A1, which is displayed in the Name Box just above the column marker.

Now you are ready for entering the data. You can enter the data in column A. It may be a good idea to write the Label about data in the cell A1 as **Marks**. Assume the maximum marks are 50. You may like to enter this information in cells B1 and C1 as **Max Marks** and **50** respectively. (refer to Fig. 3.2). You may perform centering of labels and make them bold. (do it yourself)

A1 Marks

	A	B	C	D
1	Marks	Max Marks	50	
2	3			
3	30			
4	29			
5	17			
6	25			
7	58			
8	23			
9	48			
10	16			
11	25			
12	25			
13	29			
14	21			
15	23			
16	19			
17	22			
18	20			
19	29			
20	37			
21	32			
22				

Fig. 3.2: Marks data in a spreadsheet

You can save this data by selecting Save option from the Office Button Menu options (refer to Fig. 3.1(a)) followed by giving a file name (we have given it

a name Unit 3 Activity) to the new file in the resulting dialog box followed by selecting the Save button.

Notice that in cell A7, you have a value 58 which is larger than the Maximum Marks. Can you find this error using Excel? The next section explains this process.

3.3.2 Validating Data

You have already entered the data and have save it as a file Unit 3 Activty.xls, now is the time for you to verify your data. You want to verify that all the values entered are less than or equals to 50. For this purpose, you can use Data tab and Data Verification from Data Tools in the ribbon, however, we will show a simpler approach using Conditional Formatting.

Conditional Formatting

Conditional formatting allows you to format data as per some condition like greater than some value or less than and so on. This is a very useful way of displaying cells that may contain data that may be different from other data set.

Activity 2: Use Conditional Formatting to display the data that is in error (more than the maximum marks).

How to do the activity?

Our objective is that any cell that contains data value more than that of maximum marks value given in cell C1 should be displayed in Red colour. You may perform the following steps:

In the **Home** tab, Select the Range for which the stated rule format is to be created (for the current example, this range is cells A2 to A21, and Select **Conditional Formatting** from the **Styles** ribbon, you will see a list of options, select **Highlight Cells Rules** and from the resulting list, Select Greater Than… (refer to Fig. 3.3).

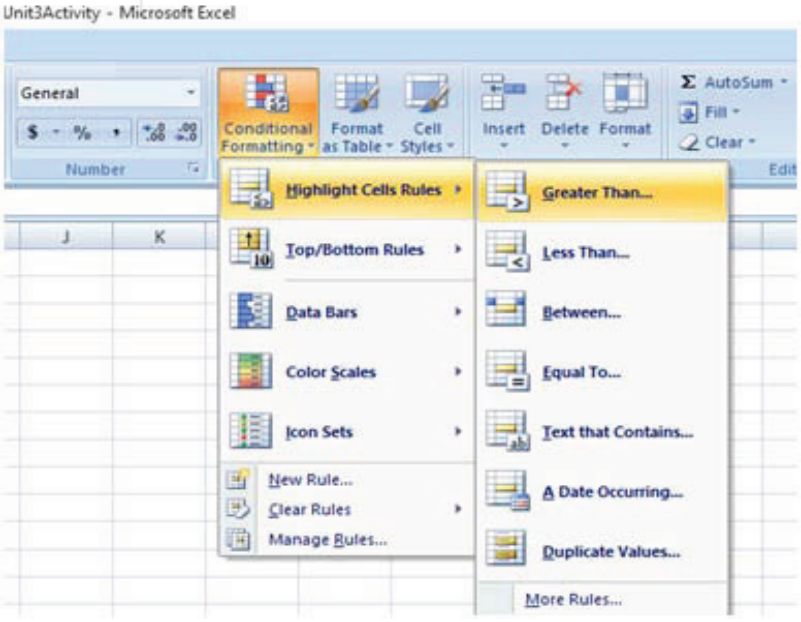


Fig. 3.3: Highlighting the cell using Greater than rule

A dialog box titled Greater than will be displayed (as shown in Fig. 3.4).

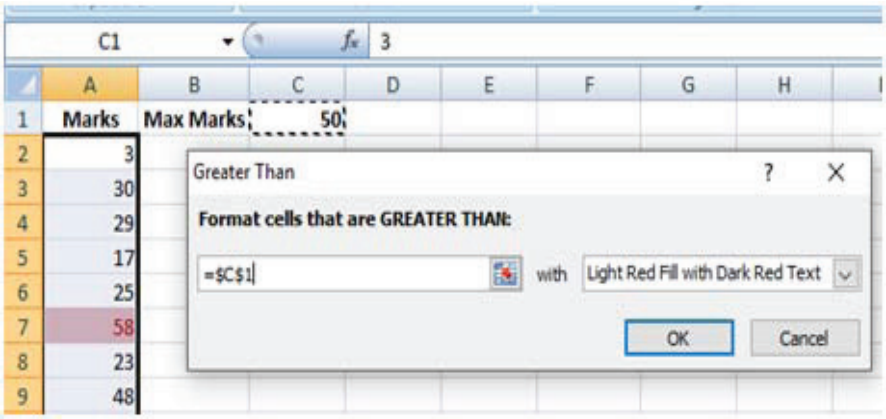


Fig. 3.4: Greater than Dialog Box

Please notice that in the text box below the label ***Format cells that are GREATER THAN:*** cell C1 is selected which contains the maximum marks. Please also note that selection of cell C1 is typed (or selected from spreadsheet) as ***=$C$1*** which is a formula with absolute addressing style of MS Excel (We will discuss about the these in more details later in this Unit). Please also notice that in the selected format of display is Light Red Fill with Dark Red Text, you can change it from the drop down list, if desired. The effect of this conditional formatting is immediately being displayed in cell C7 that contains a value more than 50 (an error of data entry). Press OK in the dialog box and the cell C7 will be highlighted. You can correct this data error.

Data Validation

Other way of validating data may be selecting a data format or condition such that you not allowed to enter incorrect data (as per rule only).

Activity 3: Create a rule for data validation for the cells A2 to A21. The rule should check that the values entered in this range is between 0 and Maximum marks 50 (given in cell C1). Further, assume that values are already entered in the cells A2 to A8 (it means that the error value of 58 is already in the entered data).

How to do the activity?

You can first create the data validation rule in cells A2 to A21 using the following steps:

Select the Range for which the stated validation is to be created (range A2 to A21), select ***Data*** tab, and Select ***Data Validation*** from the ***Data Tools*** ribbon, you will get a ***Data Validation*** dialog box, Select ***Whole number*** in the *Allow:* list, ***between*** in *Data:* list, ***0*** in *Minimum:* text box and ***=$C$1*** in the *Maximum:* text box as shown in Fig. 3.5(a).

Select the ***Error Alert*** tab in the Data Validation window and input data in Title: text box and Error Message: text area as shown in Fig. 3.5(b). Those messages will be displayed when an error will be encountered during input of values in cells C2 to C21.

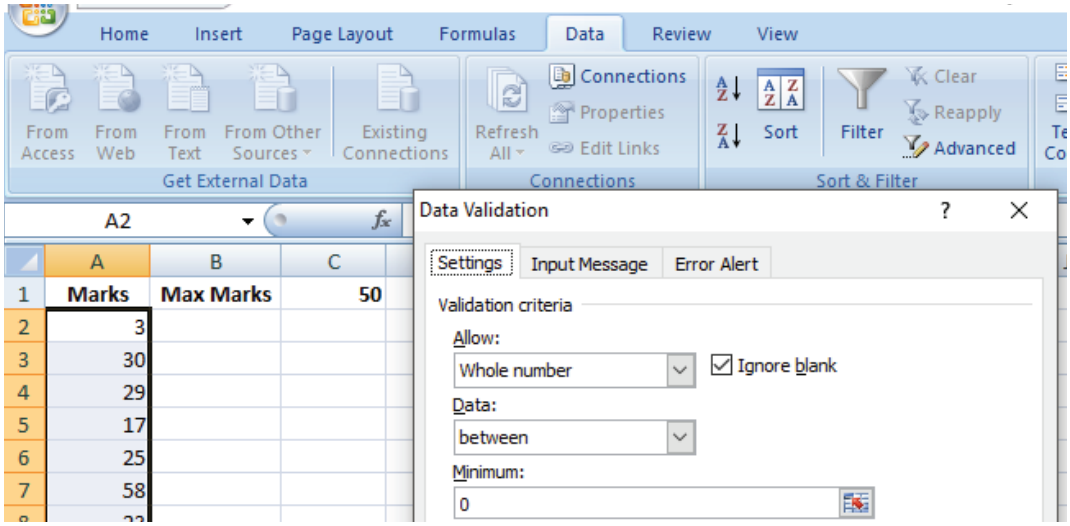


Fig. 3.5 (a): Data Validation - Settings tab

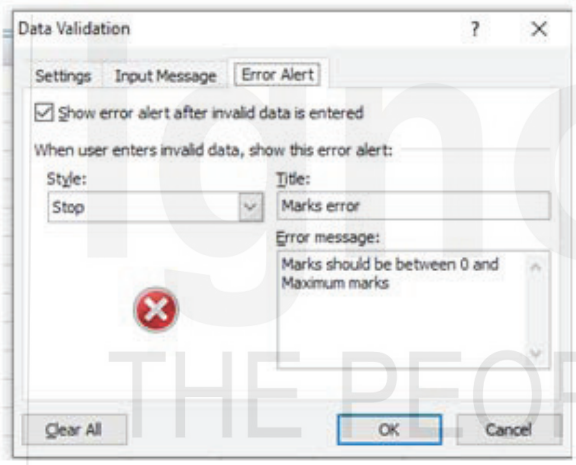
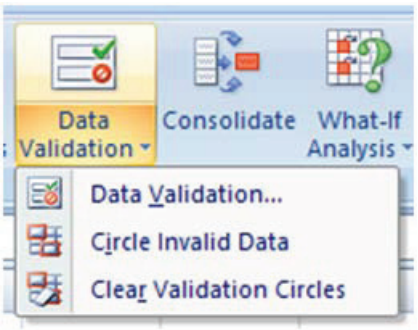


Fig. 3.5 (b): Data Validation - Error Alert tab

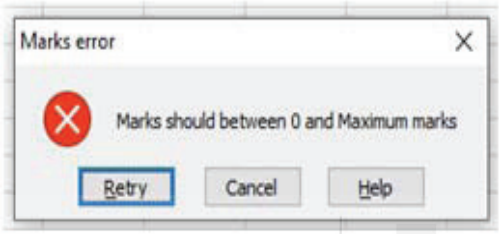
Now press OK button. You will notice nothing seems to have happened. Even the value 58 which is an error value (greater than maximum 50) is not highlighted). To highlight the error in already existing data, you need to press the text portion of ***Data Validation*** button. You will see options as shown below:



Select the Circle Invalid data Option from the displayed options. You will find that the wrong data of cell will be circled as shown below:

4	29
5	17
6	25
7	58
8	23

Please notice in Fig. 3.5(a), you have entered data till cell C8 only. Please recollect that you have entered the validation checks for cell A2 to A21. Therefore, the data validation will remain valid for cells A9 to A21 which at present contains blank value. Now, try entering value 52 and –2 in cells A9 and A10 respectively. What happens in both the cases you will see an error message alert as:



Check the title and error message of this alert box, they are same as what you have entered (please refer to Fig. 3.5(b)).

Thus, data validation not only helps in identifying cells having data validation errors but also does not allow incorrect data entry.

Check Your Progress 1

1) Create a list of 5 useful statistical software packages. Also list five important features of them.

……

……

……

……

……

2) What are the advantages of using data validations?

……

……

……

……

……

3) Create the following data using any statistical software:

Monthly Income of 10 employees of an Organisation is:

25000, 33000, 43500, 12000, 55000,

104000, 60000, 29000, 31507, 39201

Create the following validation:

Minimum income should be 15000 and maximum 100000 only. Circle all the cases which does not fulfill this validation. Format all the cells that have income more than 50000. These cells should have the fill colour as RED.

……

……

……

……

3.4 USE OF FORMULA AND FUNCTIONS

Most of the spreadsheet and statistical packages provide calculations using formula and functions. As an example, in this section MS Excel will be used to demonstrate the importance of formula and functions. It may be noted that the logic for using formula and functions may remain the same for most statistical software though the method of their use may differ.

As every statistical software has vast repository of functions, so rather than explaining majority of functions, we will explain how to use them with the help of an example. You can find out more about functions from the **help** of the statistical software package that you are using.

Activity 4: Calculate %age marks and grades of a group of students having the following total marks data:

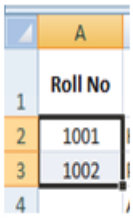
Roll No	Student Name	Marks Obtained Out of 200
1001	KAMAL JAIN	57
1002	PREM LATA SINGH	91
1003	ADRIAN	155
1004	SAPNA JHA	150
1005	ARIF KHAN	152
1006	ARJUNAN R	136
1007	SWAPN PATNAIK	173
1008	DEVENDRA MISHRA	154
1009	UMESH CHANDRA	83
1010	NAVEEN KUMAR	101

The grades are to be calculated using the following rule:

Grade Calculation Rule	
>=75%	A
>=60% but <75%	B
>= 40% but <60%	C
< 40%	D

How to do the activity?

Input of Data: As a first step, you must enter all the data in the worksheet, first enter the headings, make them bold using bold button marked as ***B*** button in the ***Fonts*** ribbon. You can also align the data to Alignment options. You can now, enter the names and marks. While entering the text please make sure that CAPS LOCK is ON so that you get uppercase letters in name. Please notice that roll numbers are in a series starting from 1001 to 1010. There is a very efficient way of entering series in worksheets. First enter ***1001*** in Cell ***A2*** and ***1002*** in cell ***A3.*** Now, select both the cells A2 and A3 by first bringing the mouse pointer on cell A2. At this point press left mouse button at this cell and keep it pressed while moving your mouse downward such that the cell A3 also gets selected (as shown below). At this point release the left mouse button.



Please note that both the cells A2 and A3 are enclosed in a box and on the right bottom corner of this box there is a small cross, when you bring your mouse on this cross it will change to a plus sign at this point press and hold the mouse button and drag the pointer down till cell A11 and release the mouse button. The values of all other cell will be filled with the series 1001 to 1010. Fig. 3.6 shows the data along with additional headings and Maximum Marks data entered in cells.

A2 | 1001

	A	B	C	D	E
1	Roll No	Student Name	Marks Obtained	%	Grade
2	1001	KAMAL JAIN	57		
3	1002	PREM LATA SINGH	91		
4	1003	ADRIAN	155		
5	1004	SAPNA JHA	150		
6	1005	ARIF KHAN	152		
7	1006	ARJUNAN R	136		
8	1007	SWAPN PATNAIK	173		
9	1008	DEVENDRA MISHRA	154		
10	1009	UMESH CHANDRA	83		
11	1010	NAVEEN KUMAR	101		
12					
13		MAXIMUM MARKS	*200*		

Fig. 3.6: Data of Activity 4

Creating Formulas in cell: Once you have entered data in the worksheet, you now need to calculate the percentage obtained by a student and on the basis of this percentage you can calculate the grade of the student. For this you need to create a formula for calculating percentage, it is Marks Obtained/Maximum marks *100. For the first student the marks obtained are in cell C2 and Maximum marks are shown in cell C13 and the calculated percentage should be shown in cell D2. Therefore, in cell D2, we need to enter the formula C2/C13*100. However, this can be considered as text input by excel, therefore, to distinguish between text and formula input ALL FORMULA ARE STARTED WITH = SIGN. Therefore, you will enter the formula ***=C2/C13*100*** in the cell D2. Since, percentage is to be calculated for the

cells C2 to C11 in the cells D2 to D11 respectively. In other words, the following formulas are to be entered in cells:

Cell	Formula
D2	=***C2***/C13*100
D3	=***C3***/C13*100
D4	=***C4***/C13*100
....	...
D11	=***C11***/C13*100

Notice that the formula in cells D2 to D11 are highlighted in above table. These formula are following a simple sequence. Are you required to enter these formulas one by one in cells? If so it will be really time consuming. Rather excel allows us to copy formula from one cell to another changing the cells addresses automatically as you go down. So after entering the formula in cell D2, you just need to copy the formula to cell D3 to D11. So first enter the formula in cell D2, then select cell D2 (as shown in Fig. 3.7) and bring the mouse pointer to right bottom corner of this cell where it will change to a plus sign; at this time press and hold Left mouse button and drag the mouse upto Cell D11 and then relase the Left mouse button. The screen display will change to as shown in Fig. 3.8.



Fig.3.7

D2 =C2/C13*100

	A	B	C	D
1	Roll No	Student Name	Marks Obtained	%
2	1001	KAMAL JAIN	57	28.5
3	1002	PREM LATA SINGH	91	#DIV/0!
4	1003	ADRIAN	155	#DIV/0!
5	1004	SAPNA JHA	150	#DIV/0!
6	1005	ARIF KHAN	152	#DIV/0!
7	1006	ARJUNAN R	136	#DIV/0!
8	1007	SWAPN PATNAIK	173	#DIV/0!
9	1008	DEVENDRA MISHRA	154	#DIV/0!
10	1009	UMESH CHANDRA	83	#DIV/0!
11	1010	NAVEEN KUMAR	101	#DIV/0!
12				
13		MAXIMUM MARKS	200	

Fig. 3.8: Copying of formula

Please note that after copying the formula in cell D3 to D11 all of them show division by zero error rather than percentage. Why? Due to some problem in copying of formula? To find it check the formulas in cells D3 to D11, you will notice the following fomula in the cells:

Cell	Formula Desired	Formula that was Copied
D3	=C3/C13*100	=C3/*C14*100
D4	=C4/C13*100	=C4/*C15*100
D5	=C5/C13*100	=C5/*C16*100
….	…	…
D11	=C11/C13*100	=C11/*C22*100

Which means that in the formula cell C13 also got incremented like C2. Since cells C14 to C22 contains no value and we are dividing by these cells, therefore, you will get divide by zero. How to solve this problem? You need to use a concept called Absolute Addressing which tells excel that the given cell in fomula is fixed (C13 in this case) and should not be changed while copying. For absolute addressing is you need to put a $ sign before the column (example $C13 (this is called mixed addressing)) if you are copying in column and do not want change in column address; or row (example C$13(this is also called mixed addressing)) if you are copying in row; or both (for example $C$13(this is called absolute addressing)), if you do not want the address of the cell changed while copying either in row or the column both. For the purpose of this example, we will use the last example.

So instead of putting the Formula ***=C2/C13*100*** in the cell D2 you need to put the formula ***=C2/$C$13*100*** in the cell D2 and copy the formula to cells D3 to D11. You can notice that results will be proper now. Fig. 3.9 shows the one of the copied formula and the complete percentage calculation.

D8 =C8/$C$13*100

	A	B	C	D
1	Roll No	Student Name	Marks Obtained	%
2	1001	KAMAL JAIN	57	28.5
3	1002	PREM LATA SINGH	91	45.5
4	1003	ADRIAN	155	77.5
5	1004	SAPNA JHA	150	75
6	1005	ARIF KHAN	152	76
7	1006	ARJUNAN R	136	68
8	1007	SWAPN PATNAIK	173	86.5
9	1008	DEVENDRA MISHRA	154	77
10	1009	UMESH CHANDRA	83	41.5
11	1010	NAVEEN KUMAR	101	50.5
12				
13		MAXIMUM MARKS	200	

Fig. 3.9: Copying with absolute addressing

Please notice the formula in cell D8 shown in the above figure.

Grade Calculation

Grade calculation requires the Grade calculation rule as given at the start of Activity 4. You can perform this activity either by using IF function or by using VLOOKUP function.

Grade Calculation using IF function:

IF function of Excel has the following syntax:

if(logical condition to be checked, what to return/do if condition is True, what to return/do if condition is false).

The following table describes the formula for the Grade calculation in cell E2. Please note this table assumes that the cell D2 contains the percentage marks.

Logical Condition	IF (condition,…,…)	When Condition is TRUE	When Condition is FALSE
>=75%	***=IF (D2>=75,…, …)***	=IF (D2>=75***,"A"***, …) Percentage Marks of student is equal to or above 75% which means s/she should be awarded A grade. Please note A is put in "" as it is a character.	Marks are less than 75% =IF (D2>=75,"A", *Create another IF condition*)
>=60% but <75%	=IF (D2>=75,"A", ***IF(D2>=60, …, …)*** Please note that the second IF condition has been created in the FALSE portion of the first IF condition, means second. IF will be calculated only when marks are below 75% (outer IF is False)	=IF (D2>=75,"A", IF(D2>=60, ***"B",*** …) Please note the case of B grade- marks % < 75% that is why you are checking IF = D2>=60, thus marks are >-60 but < 75%	Marks are less than 60% =IF (D2>=75,"A", IF(D2>=60, "B", *Create another IF condition*)
>= 40% but <60%	=IF (D2>=75,"A", IF(D2>=60, "B", ***IF (D2>=40, …, …)*** Please note that the second IF condition has been created in the FALSE portion of the second. IF condition, means this third IF will be calculated only when marks are below 60% (both the outer IF's are False)	=IF (D2>=75,"A", IF(D2>=60, "B", IF (D2>=40, ***"C",*** …) Please note that the third IF condition has been created in the FALSE portion of the second IF condition, means third IF will be calculated only when marks are below 60% (both the outer IF's are False)	Marks are surely less than 40% =IF (D2>=75,"A", IF(D2>=60, "B", IF (D2>=40, "C", *The grade will be D*)
< 40%	=IF (D2>=75,"A", IF(D2>=60, "B", IF (D2>=40, "C", ***"D")))*** Final expression after including all the closing parentheses:		

Now, you can copy the formula from cell E2 to cells E3 to E11. Fig. 3.10 shows the grades of the students

E7 =IF(D7>=75,"A",IF(D7>=60,"B",IF(D7>=40,"C","D")))

	A	B	C	D	E	F	G	H
1	Roll No	Student Name	Marks Obtained	%	Grade			
2	1001	KAMAL JAIN	57	28.5	D			
3	1002	PREM LATA SINGH	91	45.5	C			
4	1003	ADRIAN	155	77.5	A			
5	1004	SAPNA JHA	150	75	A			
6	1005	ARIF KHAN	152	76	A			
7	1006	ARJUNAN R	136	68	B			
8	1007	SWAPN PATNAIK	173	86.5	A			
9	1008	DEVENDRA MISHRA	154	77	A			
10	1009	UMESH CHANDRA	83	41.5	C			
11	1010	NAVEEN KUMAR	101	50.5	C			
12								
13		MAXIMUM MARKS	200					

Fig. 3.10: Grade Calculation using IF function

You must start using Excel functions, as shown in Activity 4. You can nest functions as per your needs. Another important function that can be used for looking into a table of values and selecting a grade is VLOOKUP function. Let us solve this Activity 4 with the help of VLOOKUP function.

Grade Calculation using VLOOKUP function:

The VLOOKUP function of excel has the following syntax:

=VLOOKUP(Value to be searched, Range which contains value to be looked up as well as value that is to be returned, The column number of the specified range that contains the return values, Specifying
Approximate match (TRUE) or exact match (FALSE))

Let us perform the Activity 4 using VLOOKUP function. For VLOOKUP function you need to create a sorted table on Cutoff Marks. The following table has been created for this purpose.

Lookup Table for Grade		
	Percentage	**Grade**
< 40%	0.0	D
>= 40% but <60%	40.0	C
>=60% but <75%	60.0	B
>=75%	75.0	A

Please note that we will be using only the shaded part of the table. You can create this table in the worksheet. You can now enter formula for cell **E2** as:

=VLOOKUP(D2,*$H$4:$I$7*,2,TRUE)

Explanation on formula: Cell **D2** contains the percentage that is to be looked in the percentage column of the Range H4 to I7 which contains the shaded portion of the table. Since this formula is to be copied for the same range in all cells, therefore, absolute addressing has been used that is why it

is: ***$H$4:$I$7*** . The value **2** indicates that the 2nd column of the lookup table that is Grade will be used to return the result after matching the value of D2 with Percentage values. As match here cannot be exact, therefore, **TRUE** has been selected as the fourth parameter. After entering the formula in cell E2, you can copy the formula to cells C3 to C11. Fig. 3.11 shows the output. Please note the formula of cell E8.

E8 =VLOOKUP(D8,$H$4:$I$7,2,TRUE)

	A	B	C	D	E	F	G	H	I
1	Roll No	Student Name	Marks Obtained	%	Grade				
2	1001	KAMAL JAIN	57	28.5	D		Lookup Table for Grade		
3	1002	PREM LATA SINGH	91	45.5	C			Percentage	Grade
4	1003	ADRIAN	155	77.5	A		< 40%	0.0	D
5	1004	SAPNA JHA	150	75	A		>= 40% but <60%	40.0	C
6	1005	ARIF KHAN	152	76	A		>=60% but <75%	60.0	B
7	1006	ARJUNAN R	136	68	B		>=75%	75.0	A
8	1007	SWAPN PATNAIK	173	86.5	A				
9	1008	DEVENDRA MISHRA	154	77	A				
10	1009	UMESH CHANDRA	83	41.5	C				
11	1010	NAVEEN KUMAR	101	50.5	C				
12									
13		MAXIMUM MARKS	200						
14									

Fig. 3.11: Grade Calculation using VLOOKUP function

You can match the Grades of Fig. 3.10 and 3.11 to verify that both the formulas have produced same results.

Activity 5: For the data of task 1 calculate the Maximum, Minimum and Average Marks also find the frequency distribution for grades.

How to do the activity?

You just need to enter formula for the Maximum, Minimum and Average. These formulas are:

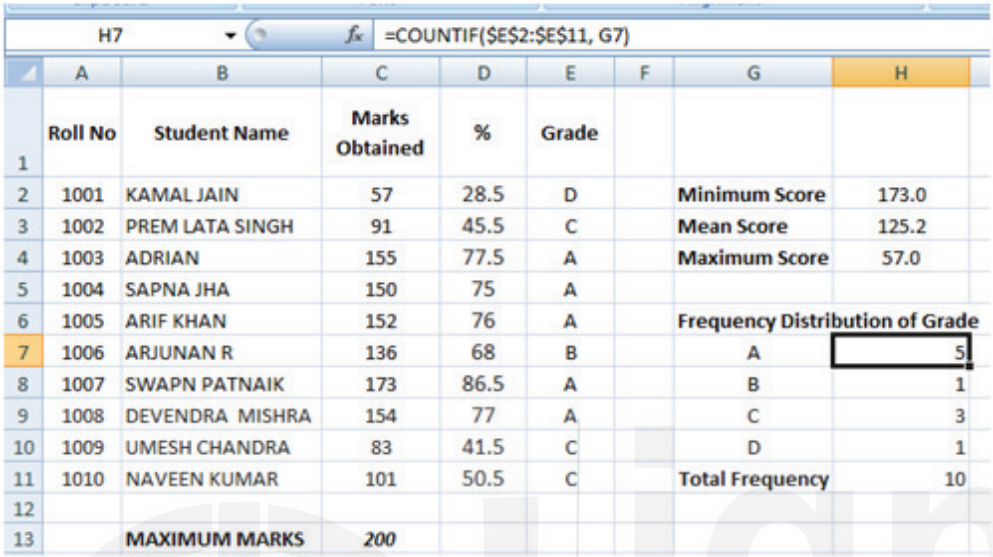
Maximum Score	=MAX(C2:C11)	For finding the Maximum
Mean Score	=AVERAGE(C2:C11)	For finding the Average
Minimum Score	=MIN(C2:C11)	For finding the Minimum

How to find which formula to be used? You must take the help of the software that you are using. Every statistical software package provide rich help documentation.

For finding frequency distribution, you may use the following formulas (please refer to Fig. 3.12 for cell references):

Frequency Distribution of Grade		Comment
A	=COUNTIF($E$2:$E$11, G7)	Absolute addressing for Range containing Grades. Formula counts the number of cells if a cell in the range
B	=COUNTIF($E$2:$E$11, G8)	
C	=COUNTIF($E$2:$E$11, G9)	

D	=COUNTIF($E$2:$E$11, G10)	is equal to specified grade in G7 or G8 or G9 or G10 cells.
Total Frequency	=SUM(H7:H10)	Formula for summing all the frequencies should be equal to total number of students.



H7 =COUNTIF($E$2:$E$11, G7)

	A	B	C	D	E	F	G	H
1	Roll No	Student Name	Marks Obtained	%	Grade			
2	1001	KAMAL JAIN	57	28.5	D		Minimum Score	173.0
3	1002	PREM LATA SINGH	91	45.5	C		Mean Score	125.2
4	1003	ADRIAN	155	77.5	A		Maximum Score	57.0
5	1004	SAPNA JHA	150	75	A			
6	1005	ARIF KHAN	152	76	A		Frequency Distribution of Grade	
7	1006	ARJUNAN R	136	68	B		A	5
8	1007	SWAPN PATNAIK	173	86.5	A		B	1
9	1008	DEVENDRA MISHRA	154	77	A		C	3
10	1009	UMESH CHANDRA	83	41.5	C		D	1
11	1010	NAVEEN KUMAR	101	50.5	C		Total Frequency	10
12								
13		MAXIMUM MARKS	200					

Fig. 3.12: Results of Activity 5

Check Your Progress 2

1) List the purpose and syntax of all the functions used in the Section 3.4, giving one example of each.

……

……

……

2) What are the advantages of using absolute addressing, formulas and functions?

……

……

……

3) Create a spreadsheet containing marks of 10 students in 5 different subjects. Assume marks in every subject is our of 50. Find the percentage of marks in each subject for every student. Also find the total Marks and total percentage for every student. A student is declared PASS in a subject if s/he gets 40% marks in that subject. To be declared as SUCCESSFUL student must PASS all the subjects. Find the students who are SUCCESSFUL. Make suitable assumptions, if any.

……

……

……

3.5 MAKING CHARTS

Charts provide visual representation to data. Every statistical software provide extensive set of charts. Details on the charts in the context of statistical analysis will be discussed in Unit 6 of this course. This section provides a brief introduction to the process of making a simple chart.

Activity 6: Draw a column chart for the frequency distribution of grades as shown in Fig. 3.12. The graph should have proper Axis titles.

How to do the activity?

In order to make graph with axis titles, we modified the grade distribution data to have proper column titles. The modified data is visible in the following table:

Frequency Distribution of Grade	
Grade	Frequency
A	5
B	1
C	3
D	1
Total Frequency	10

Select the shaded portion of the table in the worksheet and to create graph, select the ***Insert*** tab, and Select ***Column*** from the ***Charts*** ribbon, you will get a list of column chart types, select ***Clustered Column*** under ***@-D Column*** option as shown in Fig. 3.13.

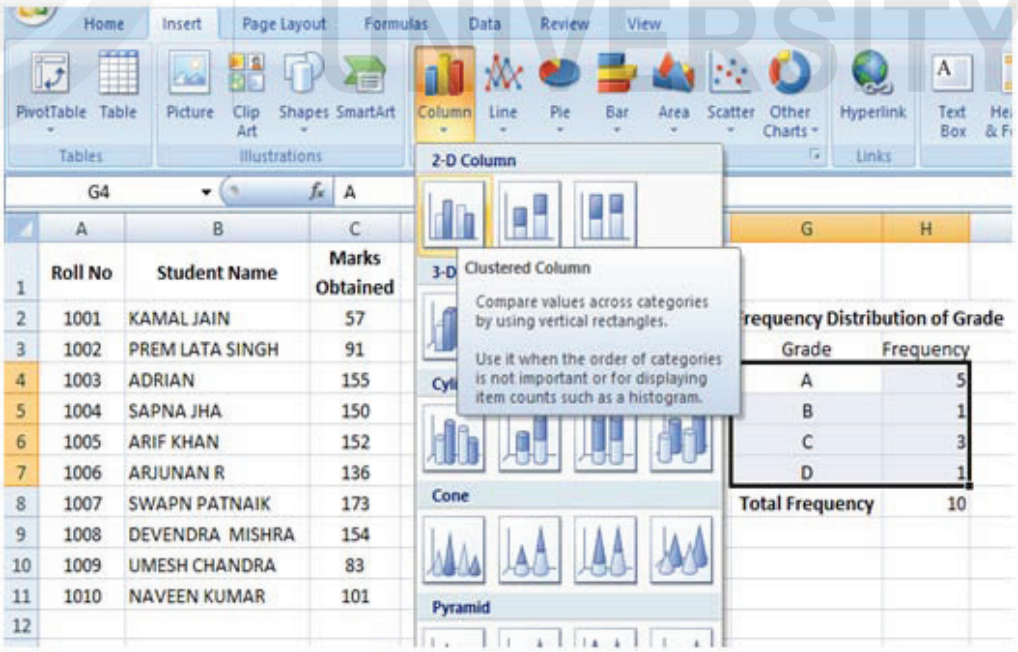


Fig. 3.13: Inserting a Column Chart

You will find that a chart will be inserted in the worksheet, but this chart has no titles (Fig. 3.14).

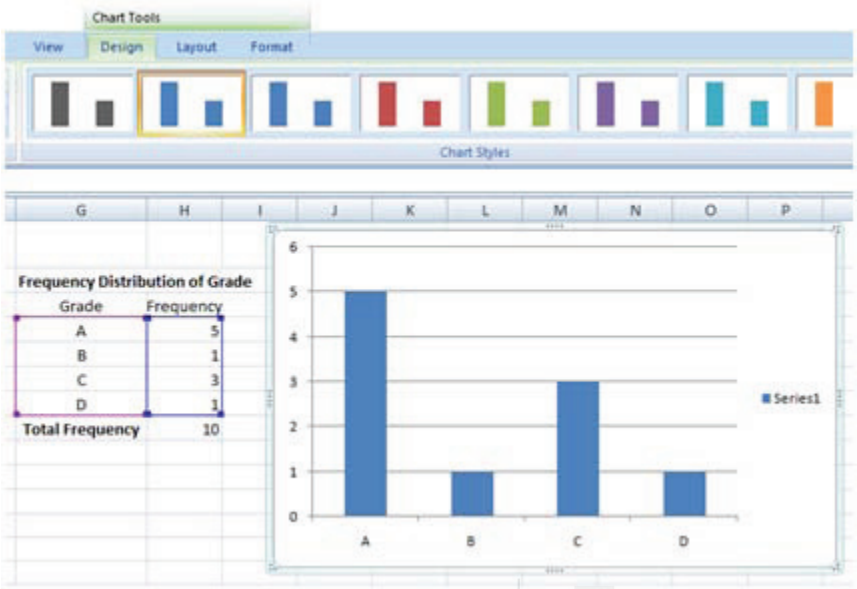
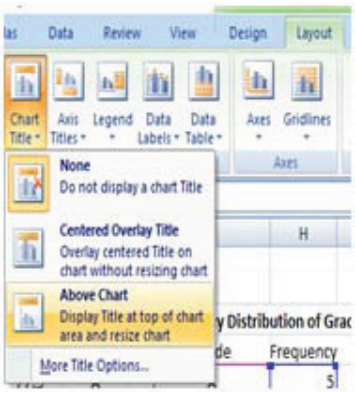


Fig. 3.14: Column chart for Grades Data

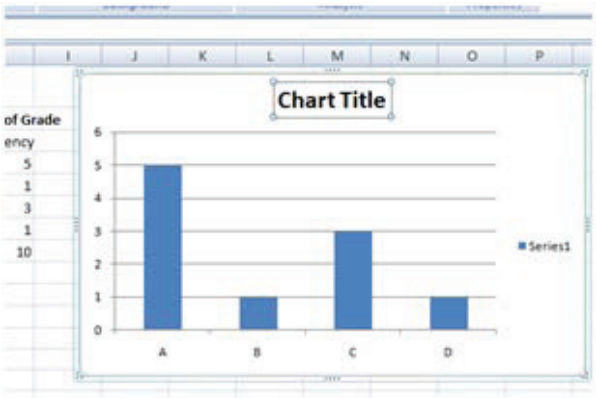
Please note that under the Chart Tools tab at the top, three more Tabs are visible: ***Design, Layout*** and ***Format***. You can select any of the Design of the columns from the displayed options. For this activity, you can focus on creating Title of Chart as well as Axis Titles of the chart. Also the legend displayed in the chart (Series 1) should be deleted. All these activities can be performed using ***Layout*** Tab. Select the ***Layout*** tab and perform the following activities:

Select ***Chart Title*** in the ***Labels*** ribbon and select ***Above Chart*** option as shown in Fig. 3.15(a). As you select the option, the chart will display **Chart Title** at the top of the chart (Fig. 3.15(b)). To add a proper chart title which is *Frequency Distribution of Grades* which is in cell **G2** enter **=** and select the Cell G2. It will display ***=Charts!$G$2*** in the formula bar (Fig. 3.15(c)).

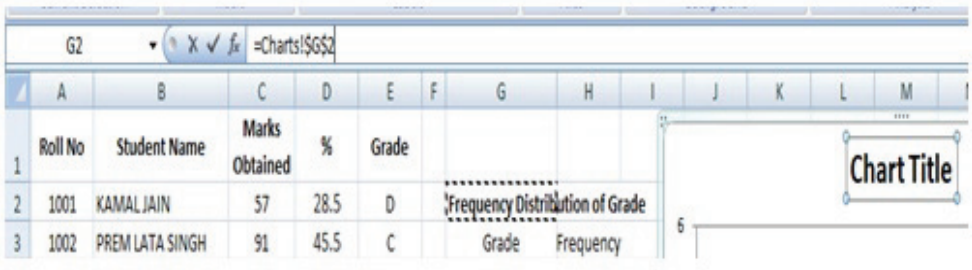
Here Charts is the name of the worksheet, in case you have not named it then it may be like Sheet 1 or Sheet 2 etc. ***between*** in *Data:* list, ***0*** in *Minimum:* text box and ***=$C$1*** in the *Maximum:* text box as shown in Fig. 3.15(a).



(a) Chart Title Option selection



(b) Result of the action in (a)



(c) Selection of Chart Title from the worksheet cell

Fig. 3.15: Selection of Chart Title

As you press Enter the Chart Title will change to Frequency Distribution of Grade (Fig. 3.16)

Now, you can select axis title in a similar way using options.

Axis Title → Primary Horizontal Axis Title → Title Below Axis → (put **=** in formula bar and select cell ***G3*** and press Enter key)

Axis Title → Primary Vertical Axis Title → Rotated Title → (put **=** in formula bar and select cell ***H3*** and press Enter key)

Now, Select ***Series1*** legend on the graph, click right mouse button and select Delete option.

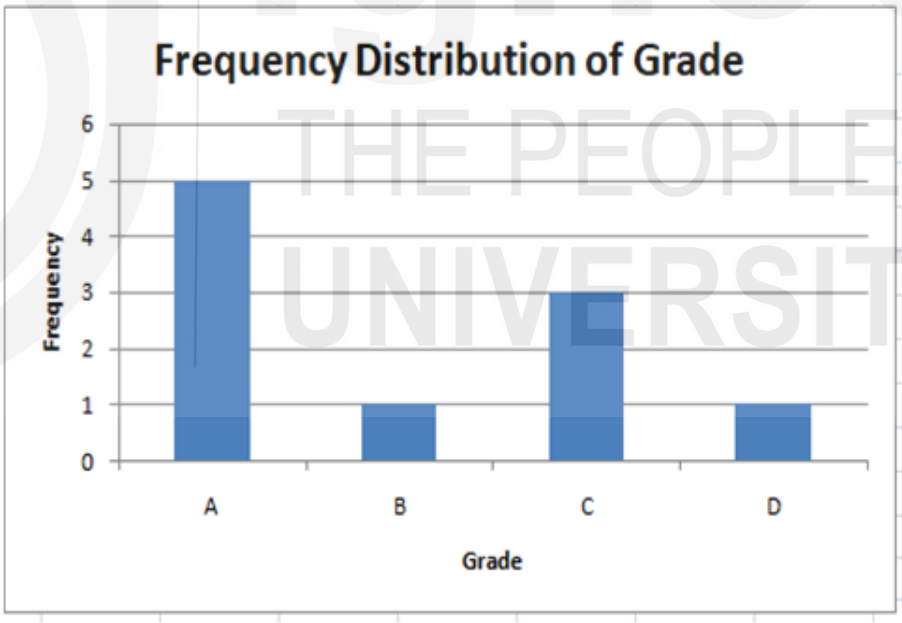


Fig. 3.16: Column chart of Activity 6

You can make a Pie Chart of the same data by simply selecting ***Pie*** from the ***Charts*** ribbon and selecting any type of pie chart. We have selected 2D Pie chart. To display the Data Labels, select ***Layout → Data Labels → Center.*** Fig. 3.17 shows the final Pie chart.

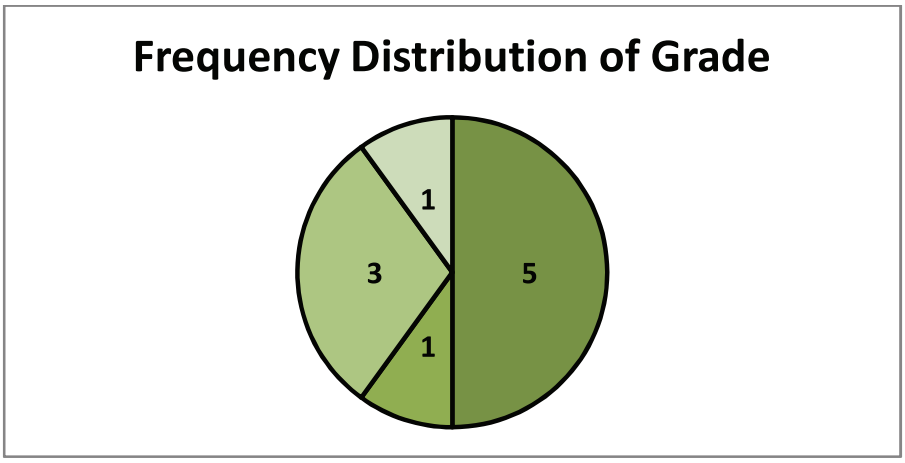


Fig. 3.17: Pie Chart for Activity 6

Similarly, you can make other charts using statistical software.

3.6 ACTIVATING DATA ANALYSIS TAB

One of the important activities that you can do through statistical software is to perform various statistical tests on the data. These tests can be performed using **Data Analysis** Option available in the **Data** tab. However, in many MS-Excel versions, this option is not immediately available and needs to be added. This section explains the process of Analysis Tool pack of Excel, so that you can easily perform many statistical tests on data.

But, first check if the Data Analysis option is there in Excel version you are using by selecting the **Data** tab. Check the right most part of the data ribbon, if it does not show **Data Analysis** icon (Fig. 3.18), then perform the steps that are detailed in this section.

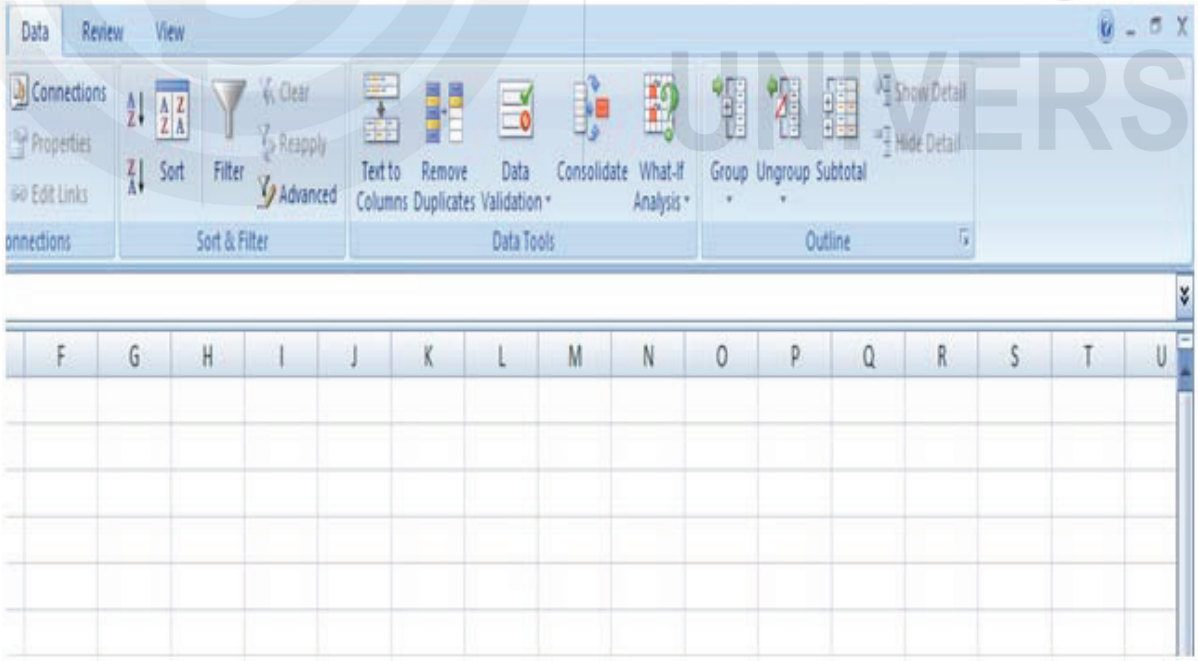


Fig. 3.18: Data tab without option of Data Analysis

Adding data analysis pack:

Select the Office button on the MS-Excel. it will show following options:

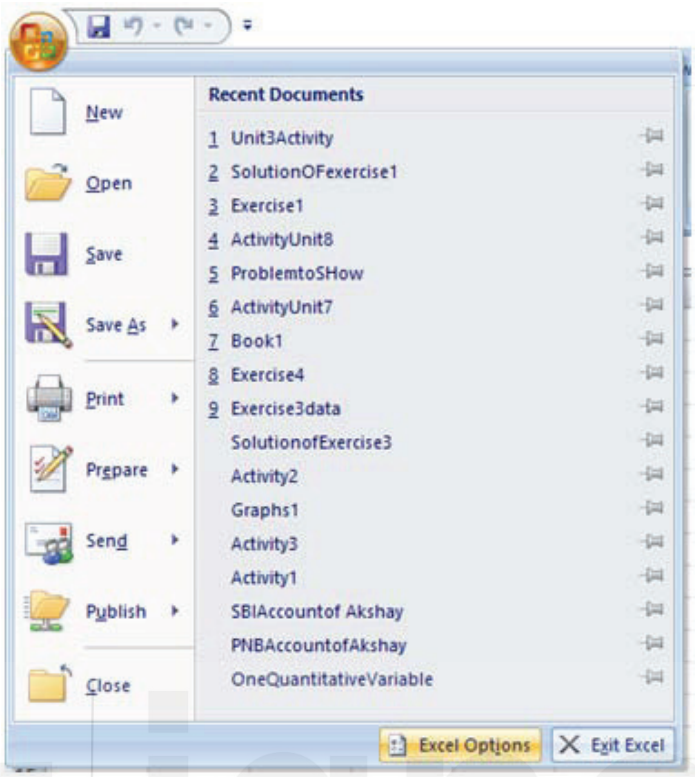


Fig. 3.19: Options list on Pressing Office button (on the top-left corner)

Please Press the ***Excel Options*** button at the bottom. A dialog box of Excel options will be displayed. Select ***Add ins*** shown in the left pane (Fig. 3.20) and press OK button which is one of the button (not displayed in Figure) on the bottom right corner.

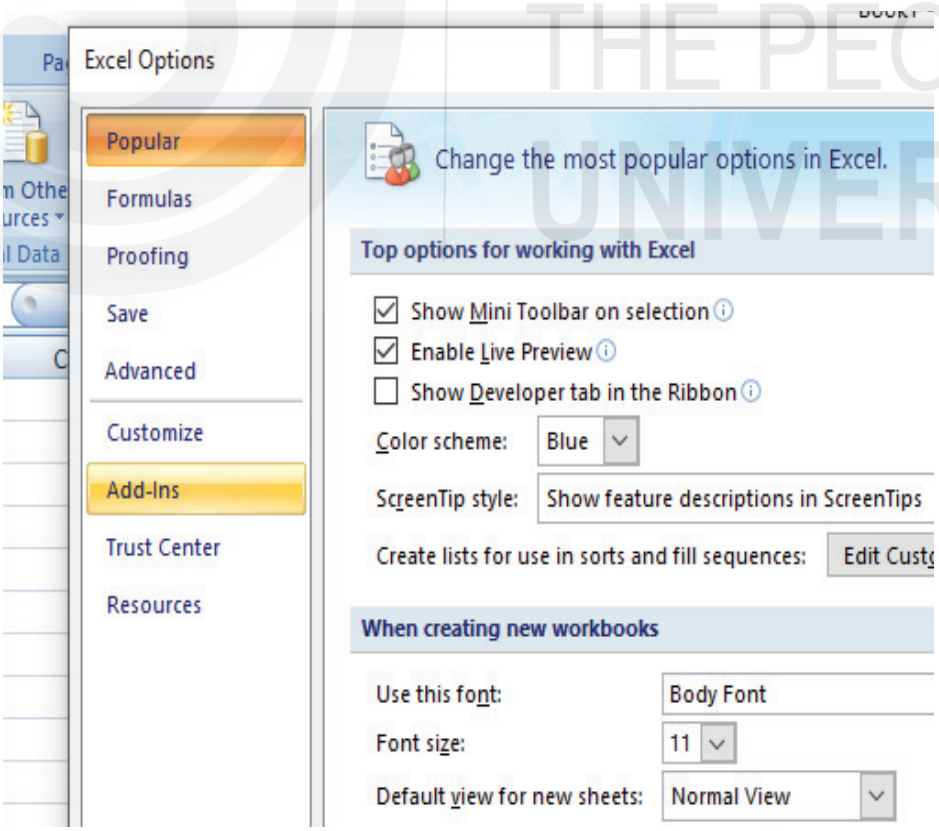
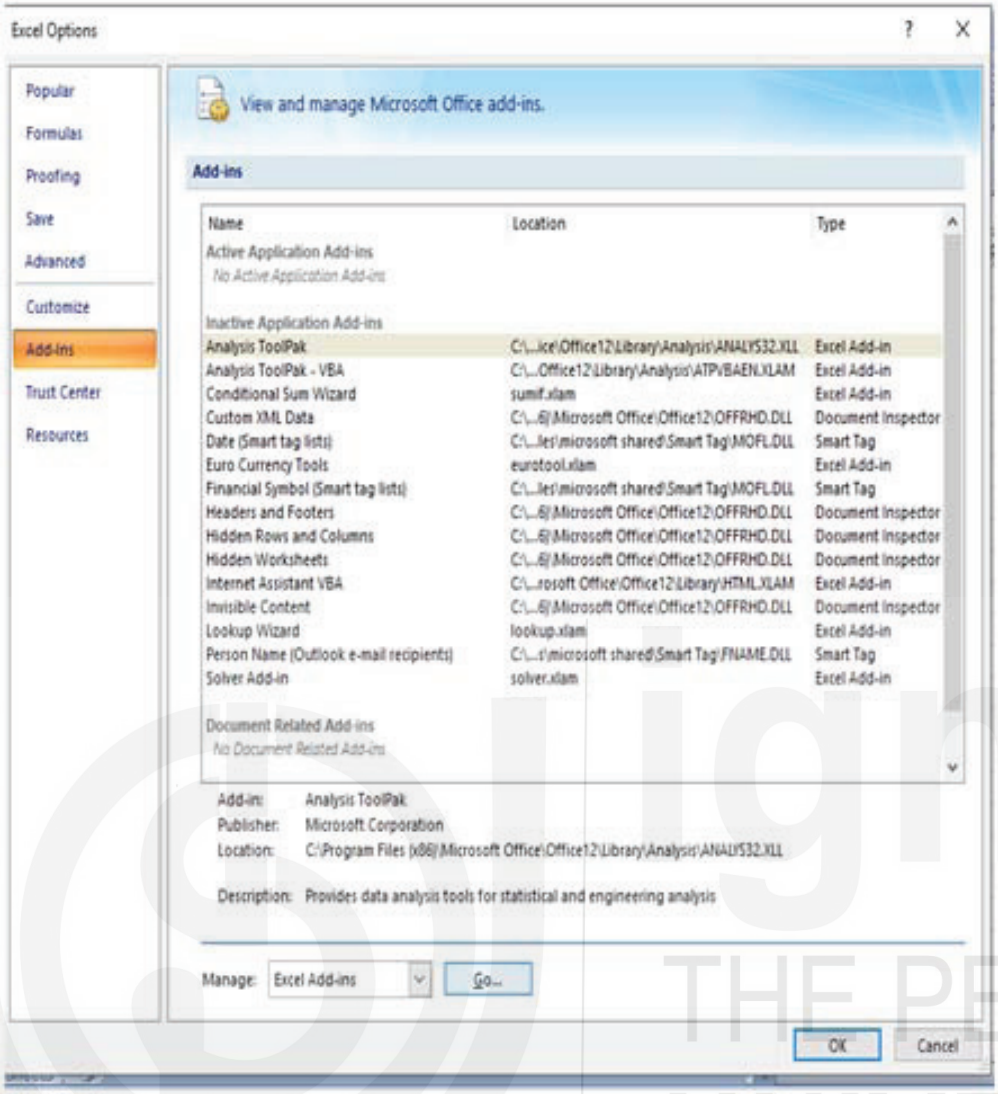
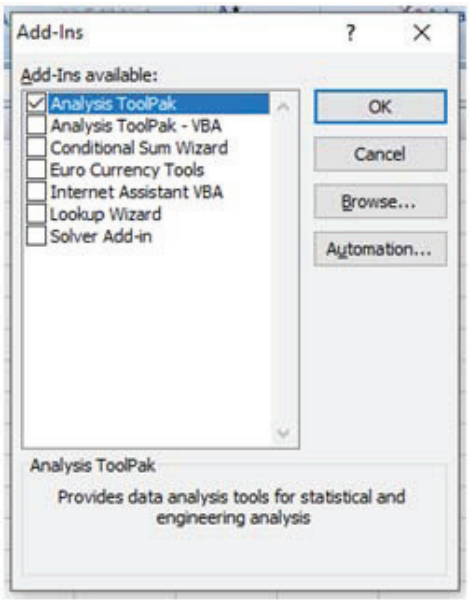


Fig. 3.20: Add-Ins option

The dialog box will change to the following:



Please notice that Analysis Tool Pak is highlighted and in the ***Manage:*** list box ***Excel Add-Ins*** is selected. Please press the ***Go…*** button immediately after this list box. The following dialog box will be displayed:



Check the check box before Analysis Tool Pak and press OK button. You will find Data Analysis option appears on the **Data** tab as shown below in Fig. 3.21. You will use this option in the next blocks of this course.

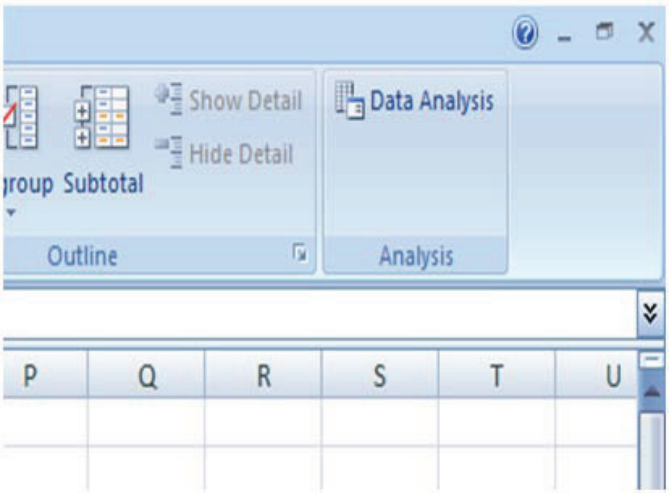


Fig. 3.21: Data tab with option of Data analysis

3.7 LET US SUM UP

This Unit introduces you to some of the important features that are part of some of the statistical packages. The process of dealing with various options in various statistical tools may differ. The Unit explains some of the basic steps required to create and validate data. For the purpose of this Unit, MS Excel has been used to demonstrate some important aspects of data handling. The Unit also explains the formula and functions which are part of statistical packages. Functions are used to process data more efficiently. A number of simple Excel functions were demonstrated in the Unit.

Charts are important tools used for statistical analysis. The Unit introduces you to the process of creating simple charts. More details on many of these aspects will be discussed in the later blocks.

3.8 EXERCISES

1) Find the top 10 most popular statistical software. Find their important functions. Also find the cost of each software.

2) List 5 different types of functions. Explain any four functions of each type along with parameters and example.

3) Add the Analysis tool pack in Excel in your computer.

4) List different types of charts supported by most statistical software. Give example of their use.

5) Following are the marks obtained by 25 students in an examination (Maximum marks 75). :

39 37 59 61 56 63 38 30 56 50

70	65	59	50	32	58	51	42	25	43
38	30	54	61	44					

i) Create a worksheet containing marks and find the percentage of marks of every student.

ii) Find the number of students who have got more than 50% marks.

iii) Calculate the Grade of each student on the basis: 80% and above A+, 60% and above but less than 80% B+, less than 60% C+

iv) Draw the Pie chart of grade distribution.

3.9 REFERENCES

Statistical software are still evolving with better and newer features being made available by the developer of the software. With gradual introduction of Artificial Intelligence and Machine Learning tools in the mainstream, more intelligent tools for statistical analysis will be available. One of the best ways to enhance your learning is to use the ***help*** of the software that you are using. The help of these software are generally supported by YouTube videos. A number of MOOCs on Statistical Analysis are available.

3.10 ANSWERS OR HINTS FOR CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

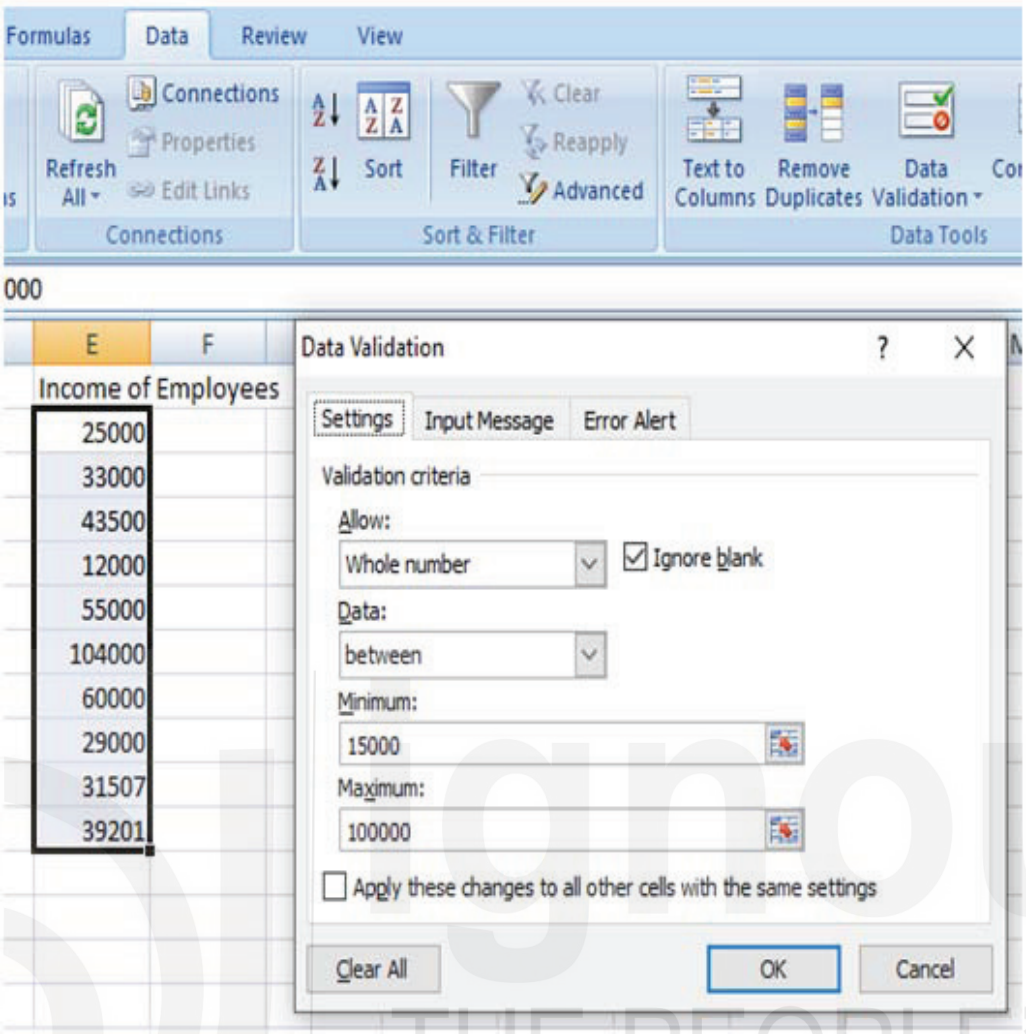
1) There are many good Statistical Softwares. Some of these are: SPSS from IBM, R Programming, MATLAB, Octave, SAS, Minitab and many more. Most of them have very easy interfaces for Data Analysis, they provide support for creating charts/graphs, used for creating descriptives, parametric and non-parametric tests and modeling. You may search for their features on the WWW.

2) One of the most critical part of analysis is the accurate data. Data inaccuracy results in improper results and conclusions. It also wastes lot of time.

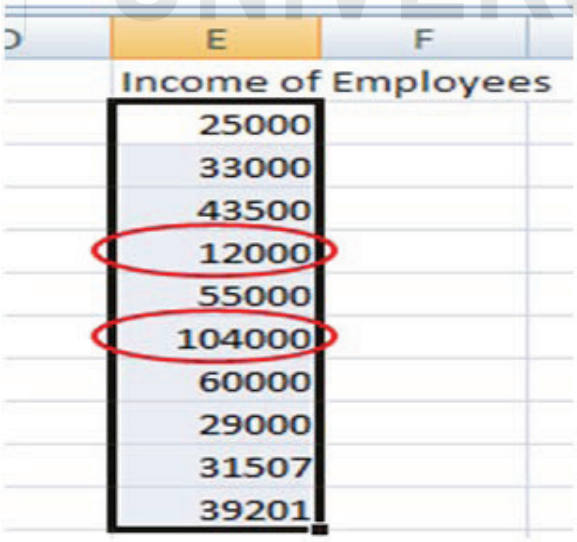
3) Perform the following steps:

- Enter the given income data in a Column.
- Select the income data entered by you and create a data validation using:

  Data → Data Tools → Data Validation → Data Validation…
- In the Data Validation dialog box, select the values as shown in the following figure and Press Ok button:



- Now select Data → Data Tools → Data Validation → Circle Invalid Data
- This will put circle on the cells that violates the specified rule (see figure)



Check Your Progress 2

1) You must use the help of the package that you are using and complete the following table:

Function	Syntax	Purpose
IF	IF(<condition or logical expression>, [value if condition is true], [value if condition is false]	IF function is used for condition checking and setting values as per the condition.
VLOOKUP	…	…
MAX		
MIN		
COUNTIF		
SUM		

2) A worksheet must have Formulas as the formulas make sure that the values are recalculated automatically if some cell containing a value is changed. Formulas also facilitate use of worksheet. Absolute addressing is useful when similar formulas are to be copied in different cell with some cells being common to all the formulas. Functions help you in performing some complex tasks with ease, for example, VLOOKUP, COUNTIF etc.

3) You can make your own solution. One possible solution is given below:

Q21

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		Marks of Student in 5 Subjects						Percentage in Every Subject						TOTAL	TOTAL %	Successful
2	Subject No	1	2	3	4	5		1	2	3	4	5				NOT Successful
3	Maximum Marks	50	50	50	50	50	Min Pass %	40%	40%	40%	40%	40%		250		
4	Student Number															
5	1	30	22	33	23	45		60%	44%	66%	46%	90%		153	61.20%	Successful
6	2	40	43	22	33	43		80%	86%	44%	66%	86%		181	72.40%	Successful
7	3	34	32	47	32	34		68%	64%	94%	64%	68%		179	71.60%	Successful
8	4	45	30	41	44	39		90%	60%	82%	88%	78%		199	79.60%	Successful
9	5	49	21	23	32	38		98%	42%	46%	64%	76%		163	65.20%	Successful
10	6	21	15	50	39	35		42%	30%	100%	78%	70%		160	64.00%	NOT Successful
11	7	24	29	43	41	19		48%	58%	86%	82%	38%		156	62.40%	NOT Successful
12	8	27	45	32	21	24		54%	90%	64%	42%	48%		149	59.60%	Successful
13	9	34	34	25	41	32		68%	68%	50%	82%	64%		166	66.40%	Successful
14	10	23	36	27	27	22		46%	72%	54%	54%	44%		135	54.00%	Successful
15																

The required formulas for the above are:

B3 fx 50

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		Mark:						Percentag						TOTAL	TOTAL %	Successful
2	Subject	1	2	3	4	5		1	2	3	4	5				NOT Successful
3	Maxim	50	50	50	50	50	M	0.4	0.4	0.4	0.4	0.4		=SUM(B3:F3)		
4	Studen															
5	1	30	22	33	23	45		=B5/B$3	=C5/C$3	=D5/D$3	=E5/E$3	=F5/F$3		=SUM(B5:F5)	=N5/N$3	=IF(AND(H5>=$H$3, I5>=$I$3, J5>=$J$3, K5>=$K$3, L5>=$L$3), $P$1, $P$2)
6	2	40	43	22	33	43		=B6/B$3	=C6/C$3	=D6/D$3	=E6/E$3	=F6/F$3		=SUM(B6:F6)	=N6/N$3	=IF(AND(H6>=$H$3, I6>=$I$3, J6>=$J$3, K6>=$K$3, L6>=$L$3), $P$1, $P$2)
7	3	34	32	47	32	34		=B7/B$3	=C7/C$3	=D7/D$3	=E7/E$3	=F7/F$3		=SUM(B7:F7)	=N7/N$3	=IF(AND(H7>=$H$3, I7>=$I$3, J7>=$J$3, K7>=$K$3, L7>=$L$3), $P$1, $P$2)
8	4	45	30	41	44	39		=B8/B$3	=C8/C$3	=D8/D$3	=E8/E$3	=F8/F$3		=SUM(B8:F8)	=N8/N$3	=IF(AND(H8>=$H$3, I8>=$I$3, J8>=$J$3, K8>=$K$3, L8>=$L$3), $P$1, $P$2)
9	5	49	21	23	32	38		=B9/B$3	=C9/C$3	=D9/D$3	=E9/E$3	=F9/F$3		=SUM(B9:F9)	=N9/N$3	=IF(AND(H9>=$H$3, I9>=$I$3, J9>=$J$3, K9>=$K$3, L9>=$L$3), $P$1, $P$2)
10	6	21	15	50	39	35		=B10/B$3	=C10/C$3	=D10/D$3	=E10/E$3	=F10/F$3		=SUM(B10:F10)	=N10/N$3	=IF(AND(H10>=$H$3, I10>=$I$3, J10>=$J$3, K10>=$K$3, L10>=$L$3), $P$1, $P$2)
11	7	24	29	43	41	19		=B11/B$3	=C11/C$3	=D11/D$3	=E11/E$3	=F11/F$3		=SUM(B11:F11)	=N11/N$3	=IF(AND(H11>=$H$3, I11>=$I$3, J11>=$J$3, K11>=$K$3, L11>=$L$3), $P$1, $P$2)
12	8	27	45	32	21	24		=B12/B$3	=C12/C$3	=D12/D$3	=E12/E$3	=F12/F$3		=SUM(B12:F12)	=N12/N$3	=IF(AND(H12>=$H$3, I12>=$I$3, J12>=$J$3, K12>=$K$3, L12>=$L$3), $P$1, $P$2)
13	9	34	34	25	41	32		=B13/B$3	=C13/C$3	=D13/D$3	=E13/E$3	=F13/F$3		=SUM(B13:F13)	=N13/N$3	=IF(AND(H13>=$H$3, I13>=$I$3, J13>=$J$3, K13>=$K$3, L13>=$L$3), $P$1, $P$2)
14	10	23	36	27	27	22		=B14/B$3	=C14/C$3	=D14/D$3	=E14/E$3	=F14/F$3		=SUM(B14:F14)	=N14/N$3	=IF(AND(H14>=$H$3, I14>=$I$3, J14>=$J$3, K14>=$K$3, L14>=$L$3), $P$1, $P$2)
15																

Exercises

1) You need to check this information from the WWW using search engines. The list keeps changing.

2) Please refer to help section of the software that you are using. There are a number of types of function like math, logical, etc…

3) The complete process is defined in the last section of the Unit, you need to follow the steps listed.

4) Please go through the details of the statistical software that you are using.

5) You must attend this exercise yourself.