Uttar Pradesh Rajarshi Tandon Open University

M.Sc.

Environmental Science

PGEVS-106 (N)

Numerical and Statistical Computing

COURSE INTRODUCTION

Numerical and statistical computing is an important branch of computational science that combines mathematics, computer science, and statistics to handle hard problems with big datasets and elaborate mathematical models. Statistical computing focuses on the use of computational techniques to analyze and interpret data. It involves implementing statistical methods and models to understand patterns, make predictions, and inform decision-making. A variety of software tools and programming languages are required for numerical and statistical computation. MATLAB, R software and SPSS are well-known for their powerful built-in functions and comprehensive libraries for numerical and statistical work. Statistical methods for analyzing and interpreting environmental data. This includes researching patterns in climate change, pollution, biodiversity, and natural resource management. Environmental data is quantified by statisticians, which aids in understanding ecological patterns, predicting environmental changes, and guiding effective environmental management techniques. This course covers measures of central tendency, dispersion, correlation, probability, hypothesis, and a basic understanding of computers and data interpretation tools. The course is organized in the following blocks:

Block 1 covers the descriptive statistics

Block 2 deals the probability and testing of hypothesis

Block 3 describes in brief of introduction to computer

Block 4 deals the computational analyses

Block-1 **PGEVS-106N**

Rajarshi Tandon Open

Universitv. Pravaarai

Numerical and Statistical Computing

Block- 1

Descriptive statistics

Unit-1

Data and Sampling

Unit-2

Descriptive Statistics

Unit-3

Correlation and Regression Analysis

Rajarshi Tandon Open

University, Prayagraj

Numerical and Statistical Computing

Course Design Committee

Prof. Ashutosh Gupta **Chairman**
School of Science, UPRTOU, Prayagraj

Dr. Uma Rani Agarwal **Member**
Rtd. Professor, Department of Botany
CMP Degree College, Prayagraj

Dr. Ayodhaya Prasad Verma **Member**
Red. Professor, Department of Botany
B.S.N.V. P.G. College, Lucknow

Dr. Sudhir Kumar Singh **Member**
Assistant Professor
K. Banerjee Centre for Atmospheric and Ocean Studies
University of Allahabad, Prayagraj

Dr. Ravindra Pratap Singh **Member**
Assistant Professor (Biochemistry)
School of Science, UPRTOU, Prayagraj

Dr. Dharmveer Singh **Course Coordinator**
Assistant Professor (Biochemistry)
School of Science, UPRTOU, Prayagraj

Course Preparation Committee

Dr. Anuj Kumar Singh **Author** **Block-1** (Unit: 1)
Assistant Prof. (Statistics)
School of Sciences, UPRTOU, Prayagraj

Dr. Upasana Singh **Author** **Block-1&2** (Unit: 2-6)
Assistant Professor-Zoology
Prof. Rajendra Singh Rajju Bhaiya
University, Prayagraj

Dr. Jaspal Singh **Author** **Block-1&3,4** (Unit: 1,7,8,9,10,11)
Assistant Professor
Department of Environmental Science,
Bareilly College, Bareilly

Dr. Nishtha Seth **Author** **(All blocks and units)**
Associate Professor
Department of Environmental Science,
Bareilly College, Bareilly

Dr. Dharmveer Singh
(Course and SLM Coordinator)
School of Sciences, UPRTOU, Prayagraj

Introduction

This first block of numerical and statistical computing, this consists of following three units:

Unit-1: This section covers data analysis methods and sampling techniques, as well as sampling and distribution of biological and environmental samples. It also covers frequency distributions, graphical and diagrammatic representations, and graph interrelationships.

Unit-2: This unit includes measurements of central tendency (mean, median, and mode) as well as measures of dispersion (range, mean deviation, standard deviation, variance, and so on). The coefficient of variation, skewness, and kurtosis are also covered in this unit.

Unit-3: This unit covers the scatter diagram and Karl Pearson's coefficient of correlation. This unit discusses correlation properties as well as correlation coefficient limits. Pearson's coefficient, Spearman's coefficient, regressions and linear regression models, the notion of least squares, regression lines, regression coefficients, and their properties are all covered.

Unit-1: Data and Sampling

Contents

1.1.Introduction

The term "statistics" denotes both status in Latin and "an organized political state" in German, and it could have sprung from either. Initially, statistics was known as the science of statecraft, and it was utilized by the government to collect the varied data required to run the state. The great philosopher Chanakya noted in his "Arthshaastra" that for effective state management, the ruler should be knowledgeable about the composition of the state population in terms of literacy, public health, income, and cost of living, among other things. In the absence of these data (later referred to as statistics), management may become like fumbling in the dark. Data provides the foundation for research and analysis in a variety of sectors, including science, industry, and social sciences. It reflects information gathered through observations, measurements, or experiments. Data can be quantitative (numerical) or qualitative (descriptive), and it is critical for making informed decisions, testing ideas, and developing models.

Quantitative refers to numerical values that can be measured and quantified. Examples include height, weight, temperature, and sales numbers. Quantitative data can be classed as discrete or continuous. Discrete data consists of distinct values (for example, the number of students in a class).

Continuous data can take any value within a range (for example, weight or height). Qualitative Data is descriptive and categorical. It includes attributes, labels, and non-numerical entries (e.g., colors, types of cuisine, or customer feedback). Sampling is the process of picking a subset of individuals, items, or observations from a larger population to draw conclusions about that population. It is critical since examining an entire population can be inefficient, time-consuming, and costly. Probability sampling is every member of the population has a known, non-zero probability of getting selected. Probability sampling includes simple random sampling, systematic sampling, stratified sampling, and cluster sampling. Another is non-probability sampling, which is Not every member has a known possibility of getting chosen, which may introduce prejudice. It is classified into three types: convenience sampling, judgmental or purposeful sampling, and snowball sampling.

Objectives

After going through this unit you should be able to

- Know origin of Statistics, its meaning, definitions and applications
- Define universe/Population
- Define statistical problems and Limitations
- Know measurements and scales

- Distinguish between discrete and continuous data
- Know methods of data collection, primary and secondary data

1.2. Data and Statistical Data:

Statistics is a science that collects, interprets, and validates data. Statistical data analysis is a method that involves conducting numerous statistical operations. It is a type of quantitative study that attempts to quantify the data. Quantitative data is primarily descriptive data, such as survey results and observational data. Descriptive statistics is a discipline of statistics concerned with the description of gathered data. These descriptions are used to designate a certain population category based on the traits they represent. The majority of observations in this universe are variable, particularly those relating to human conduct. It is commonly understood that attitude, intelligence, and personality vary from person to person. To develop a logical definition of the group or to identify the group based on their observations/scores, they must be expressed precisely. For this purpose remarks must be expressed. Data are measurements or observations gathered for informational purposes. There are many different forms of data and ways to display it. A data unit is a single entity (such as a person or organization) in the population under study for which data are gathered. A data unit is sometimes known as a unit record or simply a record. A data item is a property (or attribute) of a data unit that can be measured or quantified, such as height, birthplace, or earnings. A data item is also known as a variable since its characteristics can fluctuate across data units and over time. An observation is when a specific data item occurs and is recorded about a data unit. It can also be known as datum, which is the single form of data. An observation can be numerical or non-numerical (categorical). For example, 173 is a numerical observation of the data item 'height (cm)', but 'Australia' is a categorical (non-numerical) observation of the data item 'country of birth'. Descriptive statistics is a branch of statistics that describes obtained data through classification, tabulation, diagrammatic and graphical presentation, and measures of central tendency and variability. These measures help researchers understand the tendency of data or scores, making it easier to describe phenomena. Descriptive statistics provide a summary of a series of data points. Descriptive statistics involves two operations:

(1) Organization of Data.

(2) Summarization of Data

1.2.1 ORGANISATION OF DATA

There are four major statistical techniques for organizing the data. These are:

(i) Classification

(ii) Tabulation

(iii) Graphical Presentation, and

(iv) Diagrammatical Presentation

1.2.2 Classification

Classification refers to the organization of data into groups based on similarities. A classification is a summary of the frequency of specific scores or ranges of scores for a variable. In the most basic form of a distribution, we shall have the value of the variable as well as the number of people who have had that value. Data should be collected and organized in a way that allows them to draw conclusions. Thus, by classifying data, the investigators get one step closer to making a choice. When raw data are structured as a frequency distribution, a significantly clearer image of score information emerges. The frequency distribution displays the number of cases that fall into a specific class interval or range of scores. A frequency distribution is a table that shows each score earned by a group of people and how frequently each score occurred.

1.2.3 Frequency Distribution can be with Ungrouped Data and Grouped Data

i. An ungrouped frequency distribution can be created by listing all score values, either from highest to lowest or lowest to highest, and inserting a tally mark (/) beneath each score whenever it appears. The frequency of recurrence of each score is represented by 'f'.

ii. Grouped frequency distribution: If the data has a large range of score values, it is difficult to obtain a clear image of such a series. In this scenario, a clustered frequency distribution should be used to get a clear view of the data. A group frequency distribution is a table that divides data into categories. It shows the number of observations from the data set that fall into each of the class.

1.2.4 Construction of frequency distribution

To prepare a frequency distribution it is essential to determine the following:

(1) The range of the given data =, the difference between the highest and lowest scores.

(2) The number of class intervals = There are no hard and fast guidelines for the number of classes into which data should be classified. If there are few scores, having a high number of class intervals is ineffective. Typically, the number of classes should be between 5 and 30.

(3) Limits of each class interval = The size, width, or range of the class—known as the "class interval" and represented by the letter "i"—is another aspect taken into account when calculating the total number of classes. The frequency distribution should produce classes of the same size when the class interval has a consistent width. The class width should be a whole number that can be easily divided by two, three, five, ten, or twenty. The class limits for distribution can be described using one of three techniques:

i. Exclusive method,

ii. Inclusive method

iii. True or actual class method.

(i) Exclusive method

The classes are created using this manner so that the upper bound of one class becomes the lower bound of the following class. It is assumed in this classification that a score equal to the top bound of the class is exclusive; for example, a 40 will fall into the 40 to 50 class rather than the 30 to 40 class (30-40, 40-50, 50-60)

(ii) Inclusive method

The inclusive approach is the technique of counting or measuring data using both endpoints of a class interval. For example, the interval 10-20 includes both 10 and 20. This method ensures that all data points within and at the interval's borders are evaluated, resulting in a full dataset representation. It differs from the exclusive method, which excludes one or both ends. The inclusive approach is commonly used in descriptive statistics and frequency distribution tables to prevent data loss and provide accurate data representation.

(iii) True or Actual class method

The True or Actual class method specifies class intervals with precise limits, including fractional or exact values. For example, if the data contains continuous values, the genuine class intervals could be 9.5-20.5 rather than rounded or approximate intervals such as 10-20. This strategy is critical for correct data representation, especially when working with continuous variables, because it removes ambiguity and overlap between intervals. The genuine class technique improves statistical analysis precision and reliability by ensuring that each data point is assigned to a separate and appropriate class.

1.2.5 Types of Frequency Distribution

Frequency distributions are techniques of organizing and presenting data that illustrate how frequently a certain value or range of values occurs. The primary types are:

i. Ungrouped Frequency Distribution: Shows each individual value and how many times it appears. Ideal for tiny datasets with different values.

ii. **Grouped Frequency Distribution:** Divides data into intervals (or classes) and displays the frequencies of values inside each interval. Useful for larger datasets.

iii. **Cumulative Frequency Distribution**: Displays the sum of frequencies up to the upper limit of each period. Helps to comprehend the distribution's evolution.

iv. **Relative Frequency Distribution:** Shows the proportion or percentage of total observations that fall into each class or interval.

v. **Percentage Frequency Distribution:** Similar to relative frequency, but calculated as a percentage of the total number of observations.

Tabulation

Tabulation is a systematic grouping of data in rows and columns that allows for easier comparison, analysis, and interpretation. It arranges raw data in an organized way, typically as tables, to emphasize linkages and patterns. Frequency distributions can take the shape of tables or graphs. Tabulation is the presentation of classified data in the form of a table. A tabular arrangement of data is more understandable and suitable for subsequent statistical analysis. A table is a structured arrangement of classified data in rows and columns, with suitable headers and subheadings. The main components of a table are:

(a) Table number: When there are multiple tables in an analysis, each table should be assigned a number for easy reference and identification. The number should be centered at the top of the table.

(b) Title of the table: Every table should have a title that accurately conveys its contents. The title should be straightforward, concise, and self-explanatory. The table's title should be put directly on top of the table, or just below or after the table number.

(c) Caption: Captions are concise, self-explanatory column headings. Captions may include headers and subheadings. The captions should be in the center of the columns. For example, we can divide pupils in a class into males and females, rural and urban, high and low socioeconomic status, and so on.

(d) Stub: Stubs stand for brief and self-explanatory headings for rows.

(e) Body of the table: This is the real table and contains numerical information or data in different cells. This arrangement of data remains according to the description of captions and stubs.

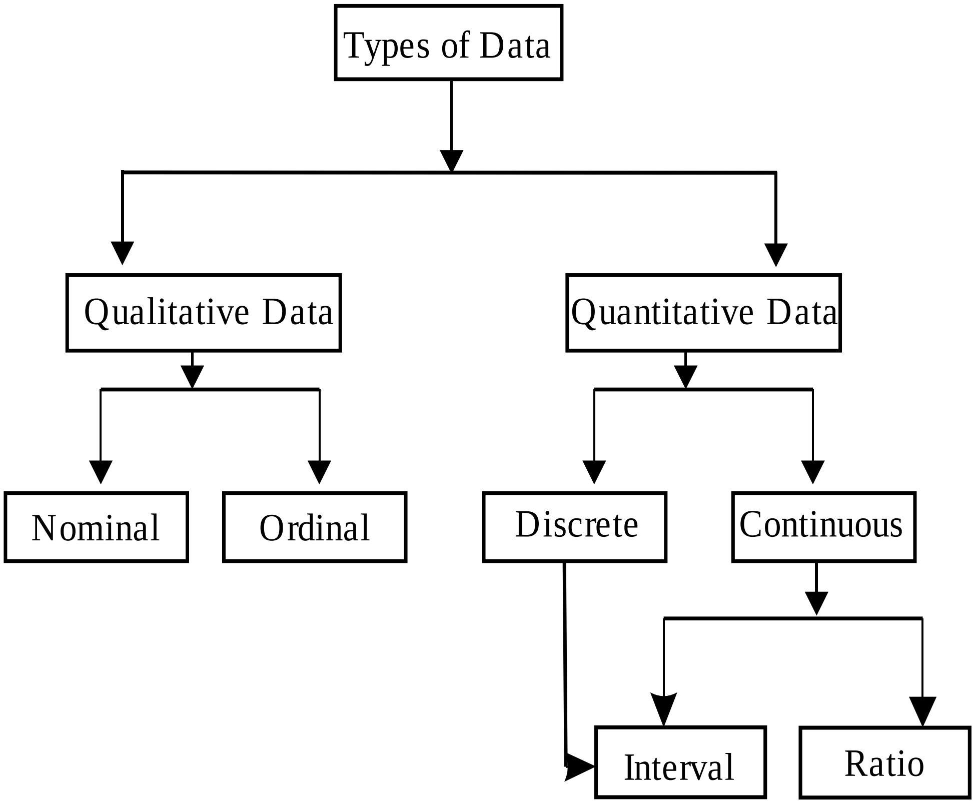
(f) Head note: This is written at the extreme right hand below the title and explains the unit of the measurements used in the body of the tables.

(g) Footnote: This is a qualifying statement which is to be written below the table explaining certain points related to the data which have not been covered in title, caption, and stubs.

(h) Source of data: The source from which data have been taken is to be mentioned at the end of the table.

1.3. Types of Statistical Data

Statistical data can be categorized into various types based on their nature and characteristics. The main types of statistical data are mentioned as



1.3.1. Qualitative Data

Qualitative data is descriptive information that characterizes and sheds light on a subject's features and properties without using numerical measurements. Unlike quantitative data, which deals with numbers and measurable forms, qualitative data focuses on the quality and features of the phenomenon under investigation. This type of data is critical in subjects such as social sciences, anthropology, psychology, market research, and any other domain where understanding the depth, context, and complexity of human behavior, experiences, or societal trends is required.

Characteristics of Qualitative Data

Qualitative data is frequently acquired using methods such as interviews, focus groups, observations, and content analyses. It encompasses the full breadth of people's experiences, thoughts, emotions, and interactions. Some essential qualities are:

- **Descriptive Nature:** Qualitative data provides extensive descriptions, allowing for a better comprehension of the topic matter. It frequently appears in the form of words, phrases, narratives, or visual content.
- **Contextual:** Contextual data is information that is particular to a given setting. The meaning generated from qualitative data is strongly dependent on the environment in which it is collected, making it distinct and frequently non-generalizable.
- **Subjective:** Qualitative data is inherently subjective because it is frequently based on individual viewpoints, opinions, and experiences. This subjectivity can provide useful information on complex human behaviors and social issues.
- **Non-Numerical:** Unlike quantitative data, qualitative data is expressed in textual or visual formats such as interview transcripts, observation notes, or movies.
- **Exploratory:** Qualitative research is frequently exploratory, aiming to get a better grasp of underlying causes, opinions, and motivations. It can assist identify trends in thoughts and attitudes and serve as a foundation for quantitative research.

Analysis of Qualitative Data

Analyzing qualitative data entails recognizing patterns, themes, and classifications. Common ways include:

- **Thematic Analysis**: This method identifies and analyzes patterns or themes in qualitative data. It is commonly used to analyze interview and focus group data.
- **Content analysis**: Content analysis entails coding and categorizing content to detect patterns, trends, and linkages. This applies to text, media, and documents.
- **Grounded Theory:** A systematic process for developing hypotheses through the meticulous collection and analysis of facts. It is especially beneficial for investigating novel or developing phenomena.
- **Narrative analysis**: Narrative analysis focuses on comprehending people's tales and personal experiences. It is valuable for investigating how people interpret their experiences.
- **Discourse analysis**: Discourse analysis investigates the use of language in texts and conversations to better understand social and cultural situations. It is used to investigate how language shapes social realities.

Qualitative data is a significant tool in many fields, providing insights that go beyond numbers to reveal the complexities of human experiences and social phenomena. Despite its problems, the depth and context it provides make it an important supplement to quantitative study, allowing for a more complete knowledge of complicated situations. Researchers can get a thorough grasp of the nuances and complexities that define human behavior and societal trends by gathering and evaluating qualitative data with care.

The observation may also include measures such as heights and weights, which are considered continuous variables. The kind of variable determines the type of data as well. There is a natural numeric scale for measuring discrete or continuous variables such as age, height, and weight, which can be expressed numerically. It can be quantified using one of two scales: interval or ratio.

Interval Scale

The interval scale is more complex than the nominal or ordinal scale. This scale may be arranged, and the distance between two measurements can be calculated. The distance between these ordered category values is equal since there is an accepted physical unit of measurement. However, the zero point is chosen arbitrarily. It can accept continuous or discrete values. **For examples** Fahrenheit (or Celsius) temperature scale in which 00F does not indicate no hears. In fact, some heat remains at temperatures ranging from 100F to -200F. When zero is designated as the measurement, there is still some heat in the (variable being measured); therefore, zero is not absolute zero.

Ratio Scale:

The ratio scale represents the highest level of measurement. This scale has a real zero point and exhibits "equal ratio" features. It is made up of meaningful, ordered features separated by equal intervals. The presence of zero point is not arbitrary; it is absolute. It is possible to multiply and divide on a ratio scale. The ratio of two values on the scale is a useful indicator of the relative magnitude of the two measures. The name ratio makes sense when you consider that a 2.5 cm line is half the length of a 5 cm line. Similarly, it makes reasonable to suggest that 20 seconds is twice as long as 10 seconds.

1.3.2. Quantitative Data

Quantitative data refers to information that can be quantified, measured, and expressed numerically. This type of data is crucial in various fields such as natural sciences, economics, engineering, and social sciences for making precise and objective assessments. Quantitative data enables researchers to test hypotheses, measure variables, and use statistical techniques to analyze relationships and trends.

Characteristics of Quantitative Data

- Numerical: Quantitative data is always expressed in numbers, allowing for clear, unambiguous measurement.
- Objective: It is less susceptible to personal bias, providing a more objective basis for decision-making.
- Measurable: Quantitative data can be counted or measured, enabling comparison and statistical analysis.
- Structured: It is typically collected in a structured manner, using tools like surveys, experiments, and observational checklists.

Types of Quantitative Data

Quantitative data is separated into two major categories:

- **Discrete Data:** Made up of unique, individual values. Examples include the number of students in a classroom, the number of cars in a parking lot, and the frequency with which an event occurs. Discrete data frequently requires counting and has integer values.
- **Continuous data:** Continuous data can have any value within a specific range and is frequently measured. Examples include height, weight, temperature, and time. Continuous data allows for more exact measurements and frequently includes fractional values.

Levels of Measurement

Quantitative data can be further classified based on the level of measurement:

- **Nominal**: The simplest level, representing categories without any quantitative value or order. For example, gender, race, or type of car.
- **Ordinal**: Represents categories with a meaningful order but no consistent difference between them. Examples include rankings (e.g., first, second, third) or levels of satisfaction (e.g., satisfied, neutral, dissatisfied).
- **Interval**: Numeric data with meaningful differences between values, but no true zero point. Examples include temperature in Celsius or Fahrenheit and IQ scores.
- **Ratio**: Similar to interval data, but with a true zero point, allowing for the calculation of ratios. Examples include height, weight, and age.

Methods of Collecting Quantitative Data

- **Surveys and Questionnaires**: Standardized tools used to gather data from a large number of respondents. Questions can be closed-ended with predefined options.
- **Experiments**: Controlled studies where variables are manipulated to observe their effect on other variables. This method is common in natural and social sciences.
- **Observational Studies**: Systematic recording of observable phenomena or behaviors. This can be structured (using predefined criteria) or unstructured.
- **Secondary Data**: Analysis of existing data collected by others, such as census data, administrative records, or previously conducted surveys.

Applications of Quantitative Data

- **Natural Sciences**: Used to measure physical properties, test scientific hypotheses, and develop models and simulations.
- **Economics and Business**: Helps in market analysis, financial forecasting, and assessing economic performance.
- **Social Sciences**: Quantitative data is used to study social phenomena, understand population trends, and evaluate policies.
- **Healthcare**: Involves measuring health indicators, analyzing treatment effectiveness, and conducting epidemiological studies.
- **Engineering and Technology**: Used for performance testing, quality control, and optimizing processes and systems.

Analysis of Quantitative Data

Quantitative data analysis involves statistical techniques to summarize, describe, and infer patterns. Common methods include:

- **Descriptive Statistics:** Summarizes data using measures such as mean, median, mode, standard deviation, and variance.
- **Inferential Statistics:** Makes inferences about a population based on sample data, using techniques such as hypothesis testing, confidence intervals, and regression analysis.
- **Correlation and Regression Analysis:** Examines relationships between variables. Correlation measures the strength and direction of a relationship, while regression predicts values of one variable based on another.
- **Analysis of Variance (ANOVA):** Tests differences between group means to determine if they are statistically significant.
- **Time Series Analysis:** Analyzes data points collected or recorded at specific time intervals to identify trends, seasonal patterns, and cyclic behaviors.

Quantitative data is crucial for scientific study, policymaking, commercial strategy, and a variety of other applications. Its ability to give exact, objective, and repeatable measurements makes it an effective instrument for comprehending and studying complicated processes. Despite its limits, quantitative data, when supplemented with qualitative insights, can provide a full picture of the world, enabling informed decision-making and supporting improvements in a variety of fields. Researchers and practitioners can use rigorous data collecting and analysis approaches to identify important patterns, test ideas, and construct strong models that improve our knowledge and capabilities.

The examples of quantitative data are as follows

- **Height measurements:** Numeric data collected by measuring the heights of individuals, expressed in centimeters or inches.
- **Test scores:** Numerical values representing performance on exams, quizzes, or standardized tests.

- **Sales revenue:** Monetary values representing the income generated from sales transactions.
- **Stock prices**: Numeric data representing the prices of stocks in financial markets at different points in time.
- **Census data:** Numeric information collected during a population census, such as population size, age distribution, or income levels.

Discrete and Continuous Data

Statistical data can be defined as a collection of numerical facts, observations, or information on variables under study in relation to the population/universe or a sample from the universe in order to meet the study or research objectives. That variable that is capable of adopting every possible fractional value within its potential bounds (called domain), when measured on different units, is called continuous; e.g., individual weight, height, age, rod length, etc. As a result, continuous data have an uninterrupted range of values and can take on either integral or fractional values.

A discrete or discontinuous variable is one that takes on particular or integral values only when measured. Examples include the number of members in a family, the number of petals in a flower, the number of fruits in a basket, and so on. Discrete data are distinct, separate, and invariable whole numbers. Statistical data are referred to be discrete or continuous data depending on the variable with which they are related. The statistical methods are only useful when some data is available. The data can be quantitative or qualitative. If the data is qualitative, it is quantified using methods such as rating, scoring, scaling, or coding. Data are acquired through experiments or surveys (directly or indirectly), tallied, and statistically analyzed. Whatever the resulting value derived from analysis, accurate and correct inferences must be drawn from these numerical data. These inferences lead to the final judgment.

1.4. Methods of Data Collection

After the sort of study to be conducted has been determined, data collection regarding the study in question is required. For this reason, it is necessary to directly or indirectly gather information from specific people. This kind of strategy is called the survey method. These are frequently applied to problems in the social sciences, such as psychology, political science, sociology, and other economic disciplines. When conducting surveys, the necessary data is either provided by the subject of the study or is derived from measurements in certain units. Data can be classified into two categories: primary and secondary, based on how they were collected.

Primary data and its collection

The data which are collected from the units or individual respondents directly for the purpose of certain study or information are known as primary data.

Secondary Data and its collection

It is the information that has been gathered and statistically processed by specific individuals or organizations. Its contents are now utilized once more, this time from records that have been processed and statistically examined to extract data for other uses. Secondary data is typically gathered via yearbooks, official records, survey reports, census reports, and documented experimental findings. Large-scale data collection is not feasible due to a lack of resources, including time, money, and personnel. Therefore, it is inevitable that some studies may incorporate secondary data. The following considerations should always be made when using secondary data:

(a) Determine the suitability of data for study.

(b) (b) Evaluate the reliability of data sources. If there is any dispute regarding the validity of the data, it should not be used.

(c) The data is not obsolete.

(d) When analyzing data from a sample, ensure that it accurately represents the population.

Preparation of Tables

Tabulation should not be mistaken with classification because they differ in many ways. The primary goal of classification is to split data into homogeneous groups or classes, whereas tabulation presents data in rows and columns. As a result, classification comes first, followed by tabulation. The steps for preparing the table are as follows.

1. Ensure the table has the necessary number of rows and columns, as well as stubs and captions, and that all data fits within the designated cells.
2. When a quantity in a table is zero, it should be recorded as zero. Leaving a blank space or using a dash in place of zero is confusing and undesired.
3. When two or more figures are identical, use ditto marks instead of original numerals in tables.
4. Specify the unit of measurement in parenthesis, either below the column captions or beside the row stubs.
5. Mark any figures in a table with an asterisk or dagger to indicate their purpose. The specification of the marked figure should be described at the foot of the table, using the same mark.

1.5. Processing Classification of Data

Before tabulation of primary data, it should be edited for (i) consistency (ii) accuracy and (iii) homogeneity

Consistency:

Some information given by the respondent may not be compatible in the sense that information furnished by the individual either does not justify some other information or is contradictory to earlier one. For example, the total expenditure exceeds the total income reported by the respondent, the number of children mentioned is less than the total number of sons and daughters, and then the respondent should be contacted again to correct the mistake so that there may be consistency.

Accuracy:

Accuracy is extremely important. If the data is erroneous, the conclusions formed from it will be irrelevant and unreliable. Only minor adjustments can be made by reviewing the schedules and questionnaires. For example, if the sum of a specific figure is incorrect, it can be amended; but, if the investigation produced a fraudulent report or the respondent purposefully provided incorrect information about his income, age, assets, and so on, editing will be ineffective. In recent years, measures have been used to ensure accuracy, such as sending supervisors to review the work of investigators or reinvestigating a select responders after a specific period of time.

Homogeneity:

To maintain homogeneity, the information sheets are verified to ensure that the unit of information or measurement is same throughout all schedules. For example, some people may have reported both monthly and annual incomes. In such a case, it must be converted to the same unit during editing. It should also be examined to see if all of the information sheets include the same information for a certain issue. The ambiguity originates from different readings of the same issue and should be deleted. Once the primary data have undergone the above process it is fit for further analysis.

Statistical methods or as a tool of analysis:

When used in a singular sense, the term refers to the science of theory and the procedures used to collect, represent, analyze, and draw conclusions about data. A. L. Bowley described statistics as "the science of measurements of social organisms, regarded as a whole in all their manifestations." In truth, a lot of definitions of statistics showing singularity exist, but possibly the best one known so far is given by Croxton and Cowden as

“Statistics may be defined as the science of collection, presentation, analysis and interpretation of numerical data.”

On the basis of these ideas, we can broadly **summarize that statistics is a science of**

- Collecting numerical information (data)
- Classification, summarization, organization and analysis of data
- Evaluation of the numerical information (data)

Methods of data collection

There are different ways of data collection. Some of them are as follows:

1. Physical observations and measurements:

The surveyor individually contacts the respondent during the meeting. He observes the sample unit and notes the results. The surveyor can always use his prior experience to collect data more effectively. For example, a young man claiming to be 60 years old can be easily identified and rectified by the surveyor.

2. Personal interview:

The surveyor is supplied with a well-prepared questionnaire. The surveyor goes to the respondents and asks the same questions mentioned in the questionnaire. The data in the questionnaire is then filled up accordingly based on the responses from the respondents.

3. Mail enquiry:

The well-prepared questionnaire is sent to the respondents through postal mail, e-mail, etc. The respondents are requested to fill up the questionnaires and send it back. In case of postal mail, many times the questionnaires are accompanied by a self-addressed envelope with postage stamps to avoid any non-response due to the cost of postage.

4. Web-based enquiry:

The survey is conducted online through internet-based web pages. There are various websites which provide such facility. The questionnaires are to be in their formats, and the link is sent to the respondents through e-mail. By clicking on the link, the respondent is brought to the concerned website, and the answers are to be given online. These answers are recorded, and responses, as well as their statistics, are sent to the surveyor. The respondents should have an internet connection to support the data collection with this procedure.

5. Registration:

The respondent is required to register the data at some designated place. For example, the number of births and deaths along with the details provided by the family members are recorded at the city municipal offices which are provided by the family members.

6. Transcription from records:

The sample of data is collected from the already recorded information. For example, the details of the number of persons in different families or number of births/deaths in a city can be obtained from the city municipal office directly.

The methods in (1) to (5) provide primary data which means collecting the data directly from the source. The method in (6) provides secondary data, which means getting the data from the primary sources

1.6. Frequency Distribution

Frequency distribution is an effective tool for presenting and interpreting data since the graph's structure makes it simple to answer a number of crucial questions. Typically, a frequency distribution's tabular form is less effective than a graphic presentation at emphasizing the key elements of the data. The graph's form provides a precise representation of the variances, skewness, peakedness, modes, extremes, outliers, spread, and other characteristics of the data distribution. Consequently, frequency distribution graphs are useful tools for efficiently analyzing and comparing two or more distributions. When the graphs of two frequency distributions are superimposed, it becomes clear where the differences and patterns occur.

Histogram

The most basic and often used graphical depiction of a frequency distribution is a histogram. It depicts a count of the number of occurrences (frequency) in each group and groups the values of a variable into evenly spaced intervals. Another way to indicate the count is as a percentage. A histogram is used to visually represent a univariate distribution's frequency distribution. Class frequency is measured along the vertical Y-axis, which starts at zero at the origin every time. Subsequently, rectangular bars with an equal width base on the X-axis are elevated across subsequent class intervals. Every bar's height, as determined by the Y-axis, is maintained at par with the matching class frequency. Each class interval's area of the bar is determined by multiplying the class frequency "r" by the class interval's width (C). Since class intervals for frequency distributions might be equal or unequal, the process for creating a histogram is explained separately for each scenario below.

(i) Histogram for Equal Intervals

The histogram of the frequency distribution with equal interval provided in Unit 2 is shown in Figure 3.1. Marking dots are used to divide the horizontal X-axis into equal sections, two or three more in number than the number of class intervals that make up the distribution. Each dot is labeled with the lower class limit of the next class, starting from the left and not always ending with zero. On either extreme, there is a space equal to the size of one class interval. On occasion, the midpoints of the succeeding class intervals are also displayed using the horizontal scale.

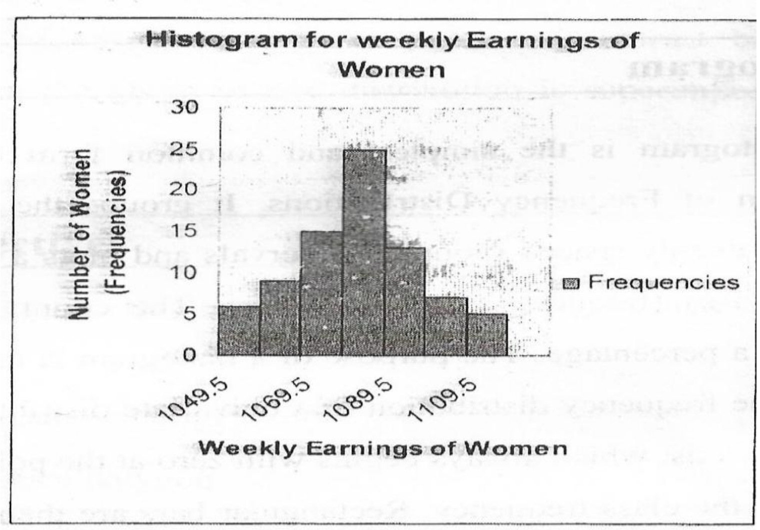


Fig. 1.1: Showing Histogram for data

(ii) Histogram for Unequal Class Intervals

Distributions with different class intervals do not differ much. Only small changes to the spacing dots indicated on the X-axis are necessary. Instead of plotting frequencies, frequency densities are shown here against class intervals.

$$Frequecny\ desity\ of\ any\ class = \frac{Class\ frequency}{width\ of\ the\ class}$$

Frequency densities are computed using a straightforward procedure that involves dividing each class's frequencies by the corresponding class widths. With the exception of not taking the open end class into account, a histogram for an open ended distribution is generated essentially in the same manner. To make the breadth of the class interval equal to the previous (or subsequent) class, limits are freely selected.

Frequencies Curve

By smoothing the frequency polygon and creating a freehand smooth curve through each point that, when joined, produces a frequency polygon, frequency curves are generated. A significant drawback of a free-hand drawn smooth curve is that no two people will ever smooth the polygon precisely the same manner. It is impossible to overstate the importance of smoothing a polygon, even with this restriction. Because of the inherent inconsistencies in the data, a frequency polygon does not become too uneven. The decision of class width instead causes the class frequencies to vary suddenly, making it more unpredictable; the main benefit of smoothing, therefore, is in removing the polygon's abrupt behavior and improving its representation of the actual fluctuations in the data. When the number of observations in the sample grows, it can be observed that a frequency distribution based on more sample data observations will have a smoother frequency polygon. As a result, it will approximate a polygon based on the total population.

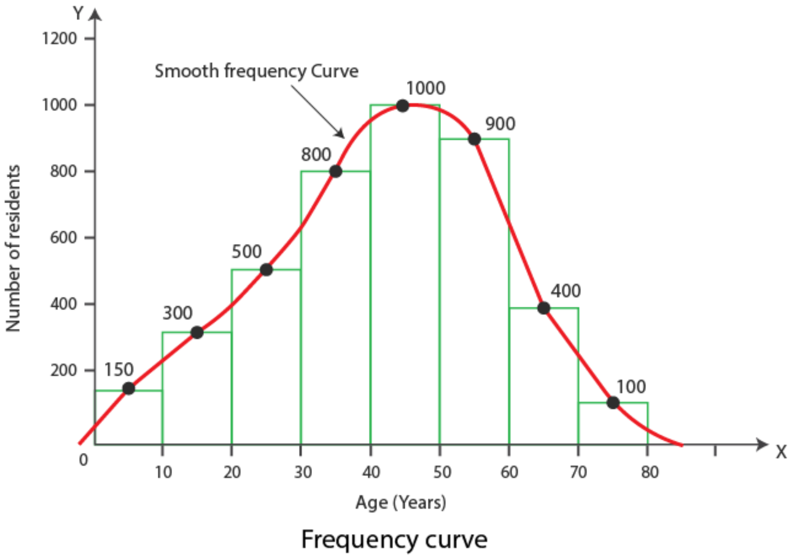


Fig.1.2: Showing frequency curve

Although a polygon can take on many different forms, symmetrical and skewed polygons are the two most common forms. The J-shaped, V-shaped, S-shaped, bimodal, and other shapes are less typical. The vertical Y-axis, which always starts at zero at the origin, is used to measure class frequency. Next, for each class interval, rectangular bars are raised with an identical width base on the X-axis. Every bar's height, as determined by the Y-axis, is maintained at par with the matching class frequency. The product of the class frequency and the class interval widths yields the area of the bar corresponding to each class interval.

1.7. Graphical and Diagrammatical Representation

Making a frequency distribution gives you a methodical approach to "looking at" and comprehending data. The data found in a frequency distribution is frequently shown in diagrammatic and/or visual formats to deepen this knowledge. When frequency distribution is presented graphically, it is plotted on a graph, which is a visual platform made up of vertical and horizontal lines.

Two mutually perpendicular lines, referred to as the X and Y-axes, are used to produce a graph on which the relevant scales are represented. The vertical line is referred to as the ordinate, and the horizontal as the abscissa. Similar to various frequency distributions, there are numerous graph types as well, which improve the reader's comprehension of science. The frequency polygon, cumulative frequency curve, frequency curve, and histogram are the most often used graph types. We'll talk about a few of the most significant categories of graphical patterns in statistics here. Using a few appropriate problems, we will demonstrate how to create line graphs. Visual tools such as diagrammatical and graphic representations are used to show relationships, data, and information visually. They aid in the succinct and straightforward communication of intricate patterns, trends, and concepts. The diagrammatical and graphical representations that are frequently employed in a variety of industries, including project management, research, data analysis, and presentations. The type of data you have and the insights or message you wish to successfully communicate will determine which format is best. Typical graphical and diagrammatical representations include the following:

Bar Chart/Bar Graph: A bar chart uses rectangular bars to represent data categories or variables. The length of each bar corresponds to the magnitude or frequency of the data. Bar charts are useful for comparing discrete categories or showing changes over time.

- **Line Graph:** A line graph displays data points connected by lines, showing the relationship between variables or the trend over time. It is particularly suitable for depicting continuous data and highlighting patterns or trends.
- **Pie Chart:** A pie chart uses a circular shape divided into slices to represent different categories or proportions of a whole. The size of each slice corresponds to the relative frequency or proportion of the data category. Pie charts are commonly used to show proportions or percentages.
- **Histogram:** A histogram is a graphical representation of the distribution of a continuous variable. It uses bars to represent the frequency or count of data falling within specific intervals or bins. Histograms are helpful in understanding the shape, central tendency, and variability of data.
- **Scatter Plot**: A scatter plot displays the relationship between two variables by plotting individual data points on a graph. Each point represents the values of the two variables, allowing us to observe patterns or correlations between them.
- **Flowchart**: A flowchart is a diagrammatic representation of a process, system, or workflow. It uses various shapes and arrows to illustrate the sequence of steps, decisions, and outcomes in a visual manner.
- **Venn Diagram:** A Venn diagram uses overlapping circles or shapes to show the relationships or commonalities between different sets or categories of data. It is useful for illustrating logical relationships, overlaps, or differences.
- **Gantt chart:** A Gantt chart is a horizontal bar chart that visualizes project schedules, tasks, and timelines. It shows the start and end dates of tasks, their durations, and the overall project timeline.

1.8. Inter relationship of graphs

Graphs are effective tools in statistics and data analysis for visually representing and interpreting data. Different types of graphs can be interconnected and complement one another, providing a full view of data from multiple angles. Understanding the interrelationships improves data analysis and interpretation. Graphs are important for viewing and analyzing data. They aid in summarizing and presenting information, identifying patterns, and drawing conclusions. Here are several interrelationships between graphs with statistical examples:

- **Histograms:**

Histograms are graphs that show the distribution of a dataset. They are made up of a sequence of bars, with the height of each bar representing the frequency or relative frequency of observations occurring during a given interval. Histograms help you visualize the shape, central tendency, and variability of data. They are often used for exploratory data analysis and determining the normality of a distribution.

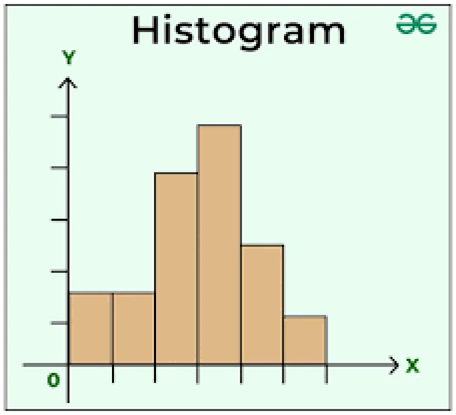


Fig.1.3: Representation of Histogram

- **Box plots:**

Box plots, also known as box-and-whisker plots, use five essential metrics to visually represent data distribution: minimum, first quartile (Q1), median, third quartile (Q3), and maximum. The box depicts the interquartile range (IQR) from Q1 to Q3, with a line marking the median. Whiskers extend to the minimum and maximum values within 1.5 times the IQR of the quartiles, emphasizing the data's range. Outliers are points that fall outside the norm. Box plots are excellent for comparing distributions, detecting skewness, and finding outliers, as they provide a quick summary of data variability and central tendency.

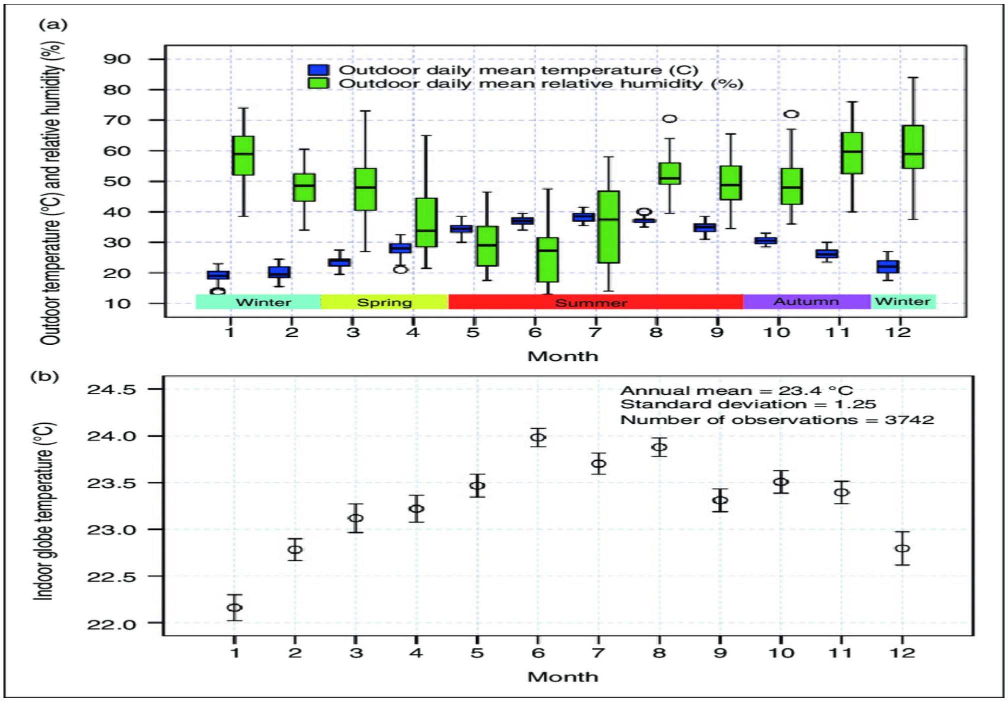


Fig.1.4: Box plot showing the outdoor temperature and relative humidity

- **Scatter plots**:

Scatter plots graphically depict the relationship between two quantitative variables in Cartesian coordinates. Each point on the plot represents an observation from the dataset, with the x-axis indicating one variable and the y-axis representing another. Scatter plots are used to determine relationships, trends, and outliers. A positive correlation is defined as an upward trend, a negative correlation as a downward trend, and no correlation as a random scatter of points. They are useful tools for displaying the strength and direction of correlations between variables, which helps with preliminary data analysis and hypothesis testing.

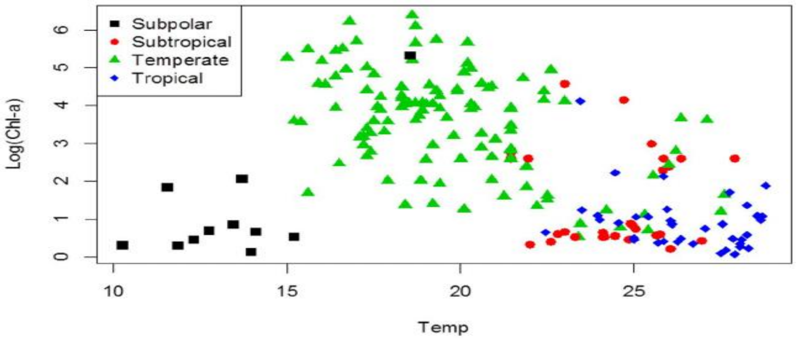


Fig.1.5: Bivariate scatter plot of environmental variables temperature

- **Bar charts:**

Bar charts use rectangular bars to illustrate categorical data, with each bar's height or length representing the frequency or value of a category. Categories are normally placed along the x-axis, with measured values displayed on the y-axis. Bar charts can be both vertical and horizontal. They are useful for comparing groups, identifying patterns, and displaying distributions. Grouped bar charts compare various categories, while stacked bar charts depict pieces of a whole. Bar charts are extensively used because they are simple and clear in providing discrete data comparisons and trends.

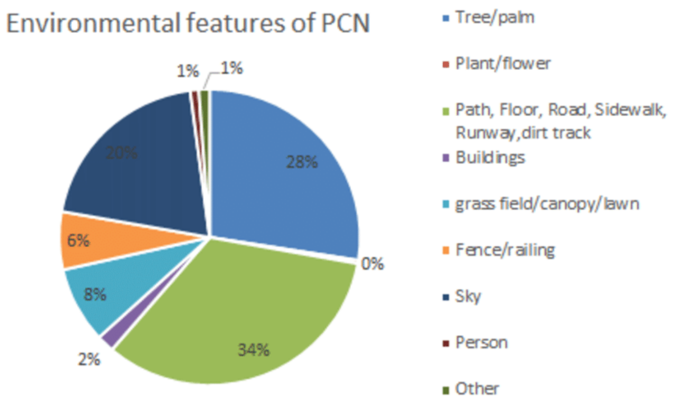


Fig.1.6: Bar chart showing the PCN's broader environmental features and its coverage

- **Line graphs:**

Line graphs, often known as line charts, exhibit data points connected by straight lines, allowing them to effectively represent trends over time or continuous variables. The x-axis often depicts time intervals or categories, and the y-axis displays measured values. Each point represents a data value, and the lines linking them indicate changes or trends. Line graphs are very effective for visualizing data trends, comparing numerous datasets, and detecting patterns or changes across time. They are commonly used in sectors including as economics, science, and business to track changes, anticipate future values, and study correlations between variables.

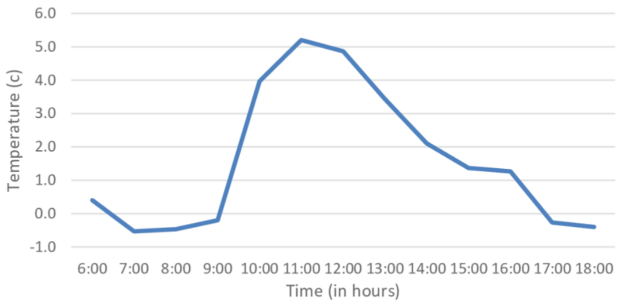


Fig.1.7: A line graph representing the average data of the three days.

- **Probability distribution plots:**

Probability distribution plots visually represent the likelihood of different outcomes in a random variable. These plots include histograms, probability density functions (PDFs), and cumulative distribution functions (CDFs).

- **Histograms:** Histograms show the frequency of outcomes in discrete intervals, which approximate a dataset's distribution.
- **Probability Density Functions (PDF):** PDFs are used for continuous data to show the probability of a variable falling within a specific range. The area under the curve reflects the entire probability, which is 1.
- **Cumulative Distribution Functions (CDFs):** Show the cumulative probability up to a specific value, with a running total of probabilities.

These graphs help you comprehend the shape, central tendency, and variability of data distributions. They are critical in statistical analysis because they allow the detection of patterns, anomalies, and the likelihood of certain events. Probability distribution charts are commonly used in finance, engineering, and natural sciences to represent and analyse random processes and events.

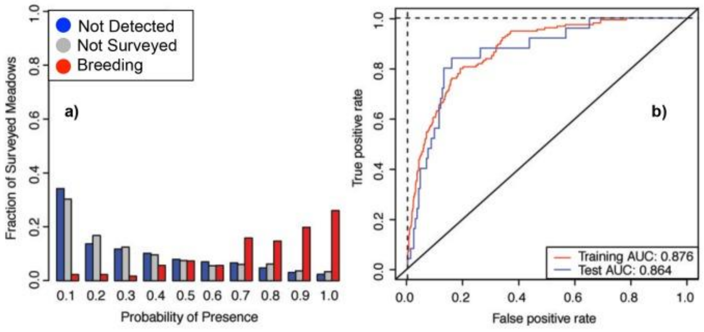


Fig.1.8: The distribution of probabilities for the three classes of data

1.9. Sampling

Sampling is an integral component of any research projects. The correct sampling strategy can make or break the validity of your research, so it is critical to select the appropriate method for your unique subject. In this post, we'll look at some of the most common sampling methods and present real-world examples of how to utilize them to collect accurate and dependable data. Sampling is the process of picking individual members or a subset of a population in order to draw statistical conclusions from them and estimate the characteristics of the entire population. Researchers in market research utilize a variety of sampling approaches to avoid having to examine the complete population in order to obtain relevant insights. It is also a time-saving and cost-effective strategy, and hence serves as the foundation of any research design. Sampling strategies can be utilized in research survey software to achieve optimal results.

For example, assume a drug maker want to investigate the detrimental effects of a medicine on the country's population. In that circumstance, it is nearly impossible to perform a research study involving all participants. In this situation, the researcher selects a sample of persons from each demographic and conducts study on them, providing indicative feedback on the drug's behavior.

.

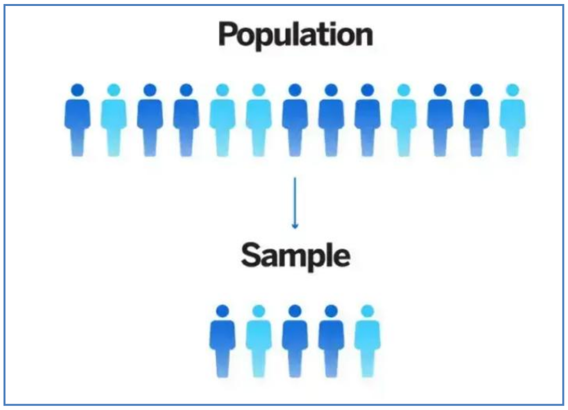


Fig. 1.9: Representation of sample

Key point of sampling

(a) Sampling allows researchers to draw observations and conclusions based on a tiny subset of a larger population.

(b) Sampling methods include random sampling, block sampling, judgment sampling, and systematic sampling.

(c) Researchers should be cognizant of sample errors, which could be due to random sampling or bias.

(d) Sampling is a marketing tactic that businesses use to determine their target market's needs and desires.

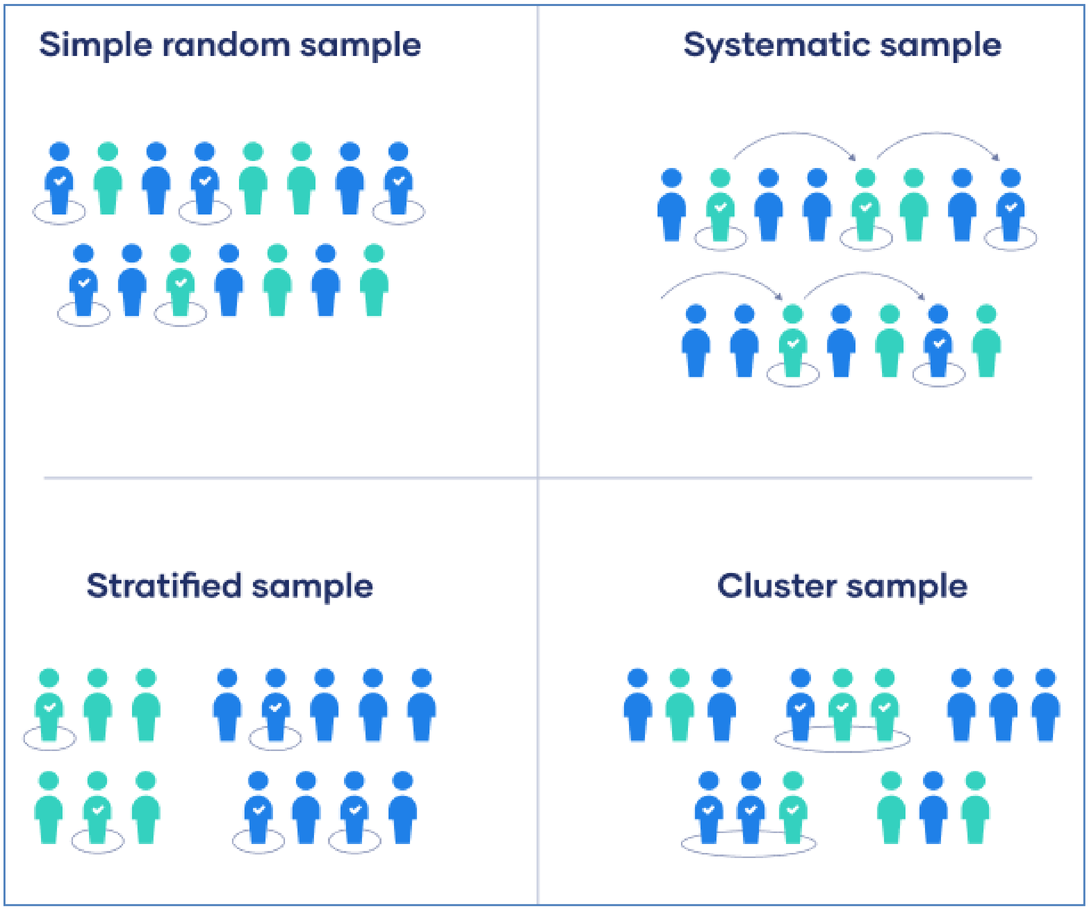
(e) Certified public accountants employ sampling during audits to ensure that account balances are accurate and complete.

1.9.1. Types of samplings:

Sampling in market action research is of two types – probability sampling and non-probability sampling. Let’s take a closer look at these two methods of sampling.

A. Probability sampling:

Probability sampling is a sampling approach in which a researcher selects a few criteria and chooses members of a population at random. This selection option ensures that all members have an equal opportunity to participate in the sample. There are four types of probability sampling methods:



1. **Simple random sampling:**

The Simple Random Sampling method is one of the most efficient probability sampling approaches for saving time and money. It is a dependable way of gathering information in which every single member of a population is chosen at random, entirely by chance. Each person has an equal chance of being picked to be part of a sample. You wish to take a basic random sample of 1000 employees from a social media marketing company. Every employee in the corporate database is assigned a number ranging from 1 to 1000, and 100 numbers are chosen at random.

Procedure of selection of a random sample:

The procedure of selection of a random sample follows the following steps:

1. Identify the units in the population with the numbers 1 to N . N

2. Choose any random number arbitrarily in the random number table and start reading numbers.
3. Choose the sampling unit whose serial number corresponds to the random number drawn from the table of random numbers.
4. In the case of SRSWR, all the random numbers are accepted even if repeated more than once.

In the case of SRSWOR, if any random number is repeated, then it is ignored, and more numbers are drawn.

- Some sampling procedures of random Sampling are Simple Random Sampling, Stratified Sampling, Systematic Sampling, Cluster Sampling, and Multistage Sampling etc..

- Some sampling procedures of Non-random Sampling are Judgmental Sampling, Convenience Sampling, Quota Sampling, and Snowball Sampling etc..

- In the study of sampling theory, there are some possibilities to the occurrence of two types of errors, say *sampling errors* and *non sampling errors*.

In this unit we study only flowing random sampling procedures.

A sampling in which, every unit of the population (here population mustbe finite and homogeneous) has equal independent chance of selection in sample. There are two types of simple random sampling, when

- The sample units are selected without replacement (no element can be selected more than once in the same sample), is known as *Simple random sampling without replacement (SRSWOR).*

- The sample units are selected with replacement (an element may appear multiple times in the same sample), is known as *Simple random sampling with replacement (SRSWR).* (Practically it is not used for further analysis).

Sampling For Proportions and Percentages

In many situations, the characteristic under study on which the observations are collected is qualitative in nature. For example, the responses of customers in many marketing surveys are based on replies like ‘yes’ or ‘no’, ‘agree’ or ‘disagree’ etc. Sometimes the respondents are asked to arrange several options in the order like the first choice, second choice etc. Sometimes the objective of the survey is to estimate the population proportion or the percentage of browneyed persons, unemployed persons, graduate persons or persons favoring a proposal, etc. In such situations, the first question arises how to do the sampling and secondly how to estimate the population parameters like population mean, population variance, etc.

2. **Cluster sampling:** Cluster sampling is a strategy in which researchers divide the total population into parts or clusters that represent the population. Clusters are discovered and included in a sample based on demographic characteristics such as age, gender, and geography. This makes it very easy for a survey author to draw useful conclusions from the feedback. The organization has offices in ten cities across the country (all with roughly the same number of employees performing similar functions). You do not have the resources to visit every office to collect data, so you use random sampling to select three offices - these are your clusters. In cluster sampling
    i. Sort the population into clusters based on a stated rule.
    ii. Treat clusters as sample units.
    iii. Select a sample of clusters based on a procedure.
    iv. Enumerate all sampling units in selected clusters.
3. **Systematic sampling**

Researchers utilize the systematic sampling approach to select sample members from a population at regular periods. It is necessary to choose a beginning point for the sample and determine the sample size at regular intervals. This sampling method has a specified range and hence requires the least amount of time. All employees of the company are listed alphabetically. From the first ten digits, you choose a starting point at random: number 6. Starting with number 6, every tenth person on the list is chosen (6, 16, 26, 36, and so on), resulting in a sample of 100 persons.

4. **Stratified random sampling**:

Stratified random sampling is a method in which the researcher separates the population into smaller groups that do not overlap while yet representing the total population. During sampling, these groups can be organized and a sample drawn from each group independently. For example, suppose you wish to determine the height of students at a university where 80% are female and 20% are male. We know that gender is strongly connected with height, and if we took a simple random sample of 200 students (from the 2,000 who attend the institution), we might obtain 200 females and not one male. This would bias our data, causing us to underestimate the overall height of the students. Instead, we might divide our sample by gender, ensuring that 20% (40 students) are male and 80% (160 students) are female.

Suppose the N units in the population are numbered 1 to N in some order. Suppose further that N is expressible as a product of two integers n and , so that $N= nk$.

To draw a sample of size n ,

- select a random number between 1 and k .
- Suppose it is i .
- Select the first unit, whose serial number is i .
- Select every *kth* unit after *ith* unit.
- The sample will contain $i, i + k, 1+ 2k, \ldots, i +(n -1)k$ serial number units.

So the first unit is selected at random and other units are selected systematically. This systematic sample is called ***k*th systematic sample** and k is termed as a **sampling interval.** This is also known as **linear systematic sampling.**

B. Non-probability sampling:

Non-probability sampling involves selecting members at random for research purposes. This sample approach does not follow a fixed or predefined selection process. This makes it difficult for all demographic elements to be represented equally in a sample. The non-probability approach is a sampling method that collects feedback based on the sample selection capabilities of a researcher or statistician rather than a predetermined selection process. Most surveys with non-probable samples produce skewed results that may not represent the desired target group. However, in other cases, such as in the early stages of research or when costs are limited, non-probability sampling is far more effective than the other type. Four categories of non-probability sampling describe the goal of this sampling approach more clearly:

1. **Convenience sampling:** This strategy is based on the ease of access to topics, such as surveying mall customers or passing by on a crowded street. It is commonly referred to as convenience sampling due to the researcher's ease of carrying it out and communicating with the subjects. Researchers have little authority in selecting sample elements, which are chosen solely on the basis of proximity rather than representativeness. This non-probability sampling method is utilized when obtaining input is limited by time and cost. In circumstances where resources are limited, such as in the early stages of research, convenience sampling is used. For example, startups and NGOs commonly do convenience sampling at a mall to distribute brochures about future events or promotion of a cause. For example, You are conducting research on student support services at your university, therefore after each class, you invite your classmates to submit a survey on the subject. This is a convenient method of data collection, but because you only questioned students in the same classes as you at the same level, the sample is not typical of all students at your university.
2. **Judgmental or purposive sampling**: The researcher chooses whether to use judgmental or purposive samples. Researchers focus solely on the study's purpose and a grasp of the target audience. For example, you want to learn more about the attitudes and experiences of handicapped students at your university, so you purposely choose a group of students with varying support needs to collect a diverse set of data on their interactions with student services.
3. **Snowball sampling:** Snowball sampling is a procedure used by researchers when subjects are difficult to trace. For example, you're researching homelessness in your city. Probability sampling is not practicable because there is no comprehensive list of the city's homeless population. You meet one individual who agrees to participate in the study, and she connects you with additional homeless persons she knows in the region.

4. **Quota sampling:** Quota sampling involves selecting members based on a predetermined standard. In this situation, because a sample is established based on specified criteria, the resulting sample will share the characteristics of the entire population. It is a quick way for collecting samples. For example, you want to measure consumer interest in a new vegetable delivery business in Boston that caters to specific dietary preferences. You divide the population into three groups: meat eaters, vegetarians, and vegans, and draw a sample of 1,000 people. Because the corporation wishes to cater to all customers, you establish a quota of 200 for each dietary group. This ensures that all dietary choices are evenly represented in your research, allowing for easy comparisons between groups. You continue to recruit until you reach the quota of 200 participants for each subgroup.

1.9.2. Sampling techniques

Sampling theory provides the tools and procedures for data collecting, taking into account the goals to be achieved and the nature of the population. There are two methods for acquiring the information.

1. **Sample surveys**

Sample surveys are frequently used in educational research to explore student academic performance, educator workforce characteristics, societal attitudes on education, and correlations between student learning and family or community factors, adult literacy, and a wide range of other topics. This article examines survey design principles that enable investigators to generalize sample results to a population of interest, as well as stratification, balancing, and cluster sampling procedures that are widely employed to improve sampling precision or lower sampling costs. Probability samples, which employ random selection methods, are contrasted with other types of data collecting, and factors for selecting a sampling design are addressed.

2. **Complete enumeration or census**

A complete enumeration, commonly known as a census, is a method used to collect data from every member of a population. Here are key aspects and purposes of a census:

a. **Coverage**: A census aims to include every individual within the defined population. For national censuses, this means every person residing in the country at the time of the census.

b. **Frequency**: National censuses are typically conducted at regular intervals, often every 10 years, although this can vary by country.

c. **Data Collection**: Detailed data is collected on a range of demographic, social, and economic characteristics such as age, gender, occupation, education, housing conditions, and more.

d. **Legal Mandate**: Many countries have laws requiring participation in the census to ensure comprehensive data collection.

e. **Accuracy**: A complete enumeration aims to minimize sampling errors by attempting to count every individual, though non-sampling errors such as non-response and data entry errors can still occur.

Purposes of a Census

- **Government Planning and Policy Making**: Census data is crucial for planning public services such as healthcare, education, transportation, and infrastructure. It helps governments allocate resources effectively and design policies that address the needs of the population.
- **Political Representation**: Census data is often used to determine the number of seats each region is allocated in legislative bodies. This ensures fair representation based on population size.
- **Economic Planning**: Businesses and economists use census data to understand market trends, labor force characteristics, and economic conditions. This information supports investment decisions and economic development strategies.
- **Research and Analysis**: Researchers and social scientists use census data to study population dynamics, social trends, and economic patterns. It provides a rich dataset for analyzing changes over time.
- Resource Allocation: Many funding formulas for distributing government aid and grants are based on population data from the census. Accurate counts ensure equitable distribution of resources.

Challenges in Conducting a Census

- **Cost and Resources:** Conducting a census is expensive and resource-intensive, requiring extensive planning, personnel, and technology.
- **Logistical Complexity**: Ensuring coverage of remote or hard-to-reach areas, addressing language barriers, and managing large-scale data collection are significant challenges.
- **Privacy and Data Security**: Protecting the confidentiality of respondents' information is crucial to maintaining trust and compliance. Ensuring data security is a major concern.
- **Non-Response and Underreporting**: Some populations, such as transient individuals, undocumented immigrants, or those distrustful of government, may be underrepresented, leading to data gaps.
- **Updating and Maintenance**: Keeping the data current between censuses requires supplementary surveys and administrative data sources to track changes in the population.

Examples of National Censuses

- **United States:** Conducted every 10 years, the U.S. Census is mandated by the Constitution and is used for reapportioning the House of Representatives.
- **United Kingdom**: The UK Census is also conducted every 10 years and collects detailed information about the population and households.
- **India:** The Indian Census is conducted every 10 years and is one of the largest administrative exercises in the world, covering a population of over a billion people.

A full enumeration or census is an important instrument for acquiring comprehensive data about a population, which helps with successful governance, planning, and research. Despite the limitations, precise and detailed census statistics provide enormous benefits to society.

Advantages of sampling over complete enumeration:

Sampling, as an alternative to complete enumeration or a census, offers several advantages:

1. **Cost Efficiency:** Sampling needs fewer resources than a full census. Conducting surveys on a smaller segment of the population dramatically decreases costs for staff, data collection, processing, and analysis.
2. **Time Savings:** Because fewer people are polled, the data collection procedure is more efficient. This enables firms to collect and evaluate data more quickly, which is critical for making timely decisions.
3. **Manageability**: Managing the logistics of a survey is simpler when dealing with a smaller sample. It is easier to train staff, ensure data quality, and handle data processing and analysis for a sample compared to an entire population.
4. **Detailed and Accurate Data**: With a smaller sample size, more resources may be dedicated to assuring data quality. Interviewers can receive better training, and more extensive checks can be performed, resulting in more accurate and dependable data. Sampling enables more precise data collecting on specific concerns. Surveys can include more questions and in-depth interviews without overwhelming respondents, resulting in more valuable data.
5. **Flexibility:** Sampling strategies can be tailored to target certain subgroups or regions of interest, allowing researchers to focus on select portions of the population without conducting a full census. Sampling enables continuous data gathering at regular intervals, allowing for the tracking of trends and changes over time without the need for a full census each time.
6. **Lower Respondent Burden:** A sample survey imposes a lower burden on the population. Fewer people are required to participate, which can reduce survey fatigue and improve response rates.
7. **Ethical Considerations:** Sampling can enhance privacy protection as fewer individuals are asked to provide personal information. This can be particularly important in sensitive surveys.
8. **Reduced Non-Sampling Errors:** With fewer respondents to survey, there is more possibility to reduce data collecting mistakes such as interviewer bias, measurement errors, and non-response concerns. Examples and contexts. Where sampling is preferable. Companies frequently utilize sampling to better understand consumer preferences and market trends without polling their whole client base. Sampling is

used to investigate illness prevalence, health practices, and the efficacy of public health interventions in specific populations. Pollsters use sampling to evaluate public opinion and anticipate election results without surveying every voter.

While a complete enumeration generates extensive data, sampling has various practical advantages that make it the favored option in many cases. These benefits include cost savings, faster data collecting, management, flexibility, and the possibility of improved data quality. Organizations can gain reliable insights by proactively selecting a representative sample while preserving resources and reducing respondent stress.

1.9.3. Type of surveys:

Surveys are a fundamental method for collecting data across various fields, and they come in several types, each suited to different research needs and contexts. Here are the main types of surveys:

1. Cross-Sectional Surveys

Cross-sectional surveys gather information from a group at a single point in time, offering an overview of current attributes, behaviors, or beliefs. They are commonly used in a variety of sectors, including public health, market research, and social sciences, to determine the frequency of specific characteristics or conditions. This strategy is both cost-effective and efficient, allowing researchers to examine and compare various groups within the population. However, cross-sectional surveys cannot determine causality or track changes over time. Despite these limitations, they are useful for finding trends, developing hypotheses, and making policy decisions using current data. **Example**: A survey measuring public opinion on a political issue conducted just before an election.

2. Longitudinal Surveys

Longitudinal surveys collect data from the same subjects across time, allowing researchers to monitor changes and advancements. This approach is critical for identifying long-term effects and patterns, making it useful in sectors such as public health, education, and social sciences. Longitudinal surveys can discover causal linkages by tracking how variables change over time for the same people or groups. Examples include cohort studies, which follow a specific group of people who share a common trait, and panel studies, which return to the same participants on a regular basis. While longitudinal surveys yield detailed insights, they are frequently more expensive and difficult to perform than cross-sectional surveys.Example: A survey tracking the career progression of a group of graduates over ten years.

3. Descriptive surveys

Descriptive surveys are designed to collect thorough information about a certain population's features, behaviors, and attitudes. They provide a broad overview of the issue under study without digging into causal linkages. Researchers utilize descriptive surveys to document the prevalence of specific characteristics, uncover patterns, and comprehend the distribution of diverse components within a community. **Example**: A survey detailing the demographic profile of internet users in a country.

4. Analytical Surveys

Analytical surveys go beyond descriptive statistics to find correlations between variables. They use complex statistical approaches like regression analysis and hypothesis testing to investigate correlations, relationships, and possible causality. Analytical surveys are widely utilized in fields such as public health, education, and market research to gain a better understanding of complicated phenomena. They help decision-makers, policymakers, and strategists by discovering patterns and linkages. However, conducting analytical surveys necessitates proper research design and statistical competence to assure the validity and reliability of the results. Despite their complexities, these surveys provide vital insights into understanding and tackling a wide range of societal and corporate concerns. **Example**: A survey examining the relationship between exercise frequency and health outcomes.

5. Census Surveys

Census surveys aim to collect data from every member of a population, providing a comprehensive snapshot of demographic, social, and economic characteristics. Mandated by governments, census surveys are conducted at regular intervals to inform policy-making, resource allocation, and demographic analysis. They offer invaluable insights into population size, composition, and distribution, facilitating equitable distribution of resources and representation. Census data is used across sectors, from healthcare and education to urban planning and market research. Despite their extensive scope and logistical challenges, census surveys are essential for ensuring accurate and inclusive data collection, enabling informed decision-making and addressing societal needs effectively. Example: The national population census conducted every ten years.

6. Sample Surveys

Sample surveys collect data from a subset of a larger population in order to draw conclusions about the entire group. These surveys use a variety of sampling procedures, including random, stratified, and cluster sampling, to guarantee that the selected sample accurately represents the population. Sample surveys are adaptable instruments used in a variety of sectors, including market research, public opinion polls, and the social sciences. They provide cost-effective and efficient alternatives to census surveys, allowing researchers to collect accurate data while conserving time and resources. However, careful evaluation of sampling strategy, sample size, and potential biases is required to ensure that the results are reliable and generalizable. Despite their limits, sample surveys offer useful insights for decision-making, policy formation, and understanding population dynamics in complex cultures. Example: A public opinion survey conducted with a sample of 1,000 voters.

7. Online Surveys

Online surveys collect data from respondents through digital platforms, which provide ease, accessibility, and cost-effectiveness. These surveys, which are distributed by email invites, websites, or survey platforms, reach a large number of people fast, enabling for rapid data collecting and analytics. They are commonly used for market research, customer feedback, and academic investigations. Online surveys provide greater flexibility in survey design and distribution, allowing for personalization and targeting specific groups. However, they may be prone to sampling bias and poor response rates, necessitating techniques for increasing participation and ensuring data accuracy. Despite the challenges, online surveys can give significant data for digital decision-making, product development, and research. **Example**: A customer satisfaction survey sent via email.

8. Telephone Surveys

Telephone surveys collect data through phone interviews, making it a low-cost and efficient way to acquire information from a wide range of groups. They are widely employed in market research, public opinion polls, and social science investigations. Telephone surveys enable speedy data gathering and real-time engagement with respondents, as well as the ability to ask follow-up questions and clarify answers. However, decreased landline usage and increased mobile phone screening pose problems to reaching respondents, resulting in lower response rates and potential biases. Despite its limitations, telephone surveys are nevertheless a useful tool for gaining insights about attitudes, behaviors, and opinions across a variety of topics, especially when paired with other survey methodologies. Example: A political poll conducted by calling a sample of registered voters.

9. Face-to-Face Surveys

Face-to-face surveys allow interviewers to interact directly with respondents, providing a more personal and in-depth approach to data collection. They can be conducted in a variety of contexts, from homes to public spaces, and allow for sophisticated questioning, rapport building, and nonverbal cue observation. Face-to-face surveys are useful for cultural contexts, complex themes, and difficult-to-reach groups. However, they necessitate tremendous resources, effort, and qualified personnel. Furthermore, privacy concerns and the social desirability bias may influence replies. Despite their limits, face-to-face surveys generate rich qualitative data, foster trust, and provide insights into attitudes, behaviors, and perceptions that are essential for informed decision-making in areas such as healthcare, social sciences, and market research. For example, field interviewers may perform a household survey on living conditions.

10. Self-Administered Surveys

Self-administered surveys are filled out by respondents without the presence of an interviewer. They provide flexibility, anonymity, and cost-effectiveness, and are commonly used in online forms, mail, and mobile applications. While decreasing interviewer bias, they may have poor response rates and restricted clarifying chances. However, self-administered surveys continue to be useful for gathering data on sensitive themes and reaching out to varied communities. Consider an online feedback form for a recent service or product.

11. Interviewer-Administered Surveys

Interviewer-administered surveys entail direct interaction between interviewers and respondents, which can take place either in person or over the phone. These surveys allow for question clarification, rapport development, and higher response rates than self-administered approaches. However, they are resource-intensive, necessitate skilled interviewers, and may result in interviewer bias. Despite these obstacles, interviewer-administered surveys produce rich qualitative data and are critical for addressing various people and complicated issues, making them useful in sectors such as public health, social sciences, and market research. Example: An in-depth interview regarding personal health routines.

12. Exploratory Surveys

Exploratory surveys are designed to collect preliminary data and insights on a specific topic or phenomenon where limited information is available. They are frequently used in the early stages of research to characterize problems, create hypotheses, and inform the design of more specific studies. Exploratory surveys utilize open-ended questions and qualitative approaches like interviews or focus groups to investigate a wide range of opinions and experiences. While they may lack generalizability and statistical rigor, exploratory surveys are critical for discovering new patterns, understanding complicated topics, and guiding future research in a variety of sectors, including social sciences, market research, and healthcare. For example, a poll to better understands new social media usage trends.

13. Diagnostic Surveys

Diagnostic surveys are used to identify faults or problems in a given setting, such as an organization, community, or system. These surveys are designed to diagnose the underlying causes of issues, estimate their severity, and identify viable solutions. Diagnostic surveys frequently combine quantitative and qualitative data collection approaches, such as questionnaires, interviews, observations, and document analysis. They are widely used in areas such as organizational development, healthcare quality improvement, and community needs assessment. Diagnostic surveys provide a full understanding of underlying concerns, which informs decision-making and guides initiatives geared at effectively addressing identified challenges. For example, a survey could be used to assess employee happiness and its impact on productivity.

14. Evaluation Surveys

Evaluation surveys measure the efficacy, efficiency, and impact of programs, interventions, or policies. They seek to establish whether objectives are met, identify areas for improvement, and assess outcomes using predetermined criteria. Data from stakeholders, beneficiaries, and other relevant parties is collected through evaluation surveys using a combination of quantitative and qualitative methodologies. Surveys, interviews, focus groups, and document reviews are among the most used procedures. Evaluation surveys provide evidence-based insights into program performance and outcomes, which aid in decision-making, resource allocation, and future planning. They are critical for assuring accountability, increasing program efficacy, and encouraging continual improvement in a variety of industries, including healthcare, education, and social services. Example: A survey evaluating the outcomes of a public health campaign.

1.9.4 Sampling Distribution

A sampling distribution is the probability distribution of a statistic based on a large number of samples gathered from a certain population. The sampling distribution is the distribution of a statistic (such as the mean or proportion) computed from many samples of equal size drawn from a population. It allows us to better grasp the statistic's variability and features.

1. **Statistic**: A statistic is a numerical feature calculated from a sample, such as the mean, proportion, or variance.
2. **Population Parameter**: A population feature, such as mean, percentage, or variance.

Characteristics of Sampling Distributions

- **Center**: The mean of a statistic's sampling distribution (for example, the sample mean) is frequently the same as the population parameter it estimates. This quality is referred to as the estimator's unbiasedness.
- **Spread**: The sampling distribution's spread (variance) is determined by the sample size as well as population variance. Larger samples typically have lower variances in their sampling distributions.

- **Shape**: Regardless of the form of the population distribution, the sampling distribution tends to approach a normal distribution as sample size grows. This is due to the Central Limit Theorem.

Central Limit Theorem (CLT)

The Central Limit Theorem says that, for a sufficiently large sample size, the sampling distribution of the sample mean (or sum) will be approximately normally distributed, independent of the population's distribution, if the population has a finite variance. This theorem is important because it allows us to draw conclusions about the population using the normal distribution.

Types of Sampling Distributions

- **Sampling Distribution of the Mean:**

Let's say we want to determine how tall an adult population is on average. We compute the mean height of each sample by taking numerous identical random samples from the population. The central limit theorem predicts that the distribution of these sample means will resemble a normal distribution with the genuine population mean at its center. This is referred to as the mean's sampling distribution

- **Sampling Distribution of the Proportion**:

Let us examine a scenario in which we wish to calculate the percentage of a city's population that supports a specific policy. Several identically sized random samples are taken from the population, and the percentage of supporters in each sample is determined. These sample proportions will have a distribution that is roughly normal and centered on the actual population percentage.

- **Sampling Distribution of the Difference in Means:**

Let's say we wish to compare the mean incomes of two distinct occupations. We compute the mean wage in each of the few random samples we choose from each profession, and we also compute the difference in means between the two samples. We can draw conclusions about the population mean difference by analyzing the sampling distribution of the difference in means, which is formed by the distribution of these differences.

- **Sampling Distribution of the Difference in Proportions**:

Suppose we want to compare the percentage of men and women who have finished a college degree. We take random samples from each group, determine their proportions, and then compute the difference in proportions between the two groups. The distribution of these differences will provide us the sampling distribution of the difference in proportions, allowing us to draw conclusions about the population difference in proportions. Sampling distributions are important in statistical inference because they allow us to understand the behavior and features of sample statistics. They enable us to estimate parameters, test hypotheses, create confidence ranges, and make population forecasts using sample data. Sampling distributions are central to the field of inferential statistics. They provide the foundation for making predictions and inferences about a population based on sample data. Understanding the properties and behavior of sampling distributions allows statisticians to assess the reliability and accuracy of their estimates and to make well-informed decisions based on sample data.

1.10. Application of Statistics

Statistics is a versatile field with applications across a wide range of disciplines. The some important applications of statistics are:

1. Healthcare and Medicine

- ***Clinical Trials***: Statistics is used to design, conduct, and analyze clinical trials to test the efficacy and safety of new drugs and treatments.
- ***Epidemiology:*** Statistical methods help in studying the distribution and determinants of health and diseases in populations.
- ***Medical Research*****:** Biostatistics applies statistical techniques to biological research, helping in the understanding of complex biological processes and phenomena.

2. Business and Economics

- ***Market Research****:* Businesses use statistics to understand consumer behavior, preferences, and trends through surveys and data analysis.
- ***Quality Control*****:** Statistical process control techniques are used in manufacturing to monitor and improve product quality.
- ***Economic Forecasting***: Economists use statistical models to predict future economic conditions, such as inflation rates, unemployment, and GDP growth.

3**. Government and Public Policy**

- ***Census and Demographic Studies***: Governments conduct censuses and surveys to collect data on population size, composition, and distribution, which inform policy decisions.
- ***Public Health:*** Statistics help in planning and evaluating public health interventions and policies.
- ***Crime Statistics***: Law enforcement agencies use statistical analysis to understand crime patterns and develop strategies for crime prevention.

4. **Education**

- ***Educational Assessment*****:** Statistics is used to design tests, analyze educational performance, and evaluate the effectiveness of educational programs.
- ***Research in Education*****:** Educational researchers use statistical methods to study learning processes and educational outcomes.

5. **Environmental Science**

- ***Climate Studies*****:** Statistics is crucial in analyzing climate data, studying environmental changes, and predicting future climate scenarios.
- ***Environmental Impact Assessment***: Statistical methods are used to assess the impact of human activities on the environment and to guide conservation efforts.

6. **Sports and Entertainment**

- ***Sports Analytics***: Teams and coaches use statistics to analyze player performance, develop strategies, and improve team performance.
- ***Audience Analysis****:* Entertainment companies use statistical analysis to understand audience preferences and optimize content.

7. Engineering and Technology

Reliability Engineering: Statistics helps in assessing and improving the reliability and lifespan of engineering products and systems.

Data Science and Machine Learning: Statistical methods are fundamental in developing algorithms for data analysis, prediction, and machine learning applications.

8. **Social Sciences**

- ***Sociology and Psychology****:* Researchers use statistics to study social behavior, mental health, and human interactions.
- ***Political Science*****:** Statistics is used to analyze election data, public opinion, and policy impacts.

Statistics is integral to decision-making and problem-solving in various fields. Its applications are diverse, ranging from improving public health and business operations to enhancing our understanding of social behaviors and environmental changes. By providing a framework for collecting, analyzing, and interpreting data, statistics enables informed decisions and drives progress across disciplines.

1.11. Summary

Data collection is the process of obtaining information from a variety of sources in order to assess and make informed decisions. Sampling, a fundamental statistical approach, entails picking a subset of a population to represent the entire. This method saves time and resources compared to thorough enumeration (census) and are critical for practicality and efficiency. Random, stratified, and cluster sampling approaches all help to ensure that the sample is representative. Proper data collecting and sampling are critical in disciplines such as healthcare, business, and social sciences, allowing for reliable analysis, predictions, and policy making. They ensure the consistency and validity of statistical conclusions. Because sampling employs a smaller number of individuals in the community with representative traits to stand in for the entire population, large-scale research can be conducted at a more realistic cost and timeframe. When you decide to sample, you embark on a new task. You must decide who should be on your sample list and how to select individuals who will best reflect the entire population. The practice of sampling is concerned with how you accomplish this. Amplifying tactics in research range greatly across disciplines and research topics, as well as between studies. Probability sampling, commonly known as random sampling, is a type of sample selection that uses randomization rather than conscious decision. Every member of the population has a known, non-zero probability of getting chosen. Non-probability sampling strategies involve the researcher selecting things or individuals for the sample based on non-random criteria such as convenience, geographic availability, or cost.

1.12. Terminal question

Q.1. What is the data and statistical data? Describe the types of statistical data

Answer: ---

Q.2. Discuss the qualitative data and quantitative data in briefly**.**

Answer: ---

Q.3. What are the methods of data collection?

Answer: ---

Q.4. What is the processing classification of data?

Answer: ---

Q.5. What do you mean about frequency distribution?

Answer: ---

Q.6. Discuss the Inter-Relationships of Graphs with suitable diagram.

Answer: --

1.13. Further suggested readings

1. Introduction to Statistics, David Lane, Rice University

2. Basic Statistics, B.L. Agrawal, New Age International Private Limited

3. Basic Statistics, Thomas Higher Education Textbooks

4. Computer Fundamentals : Concepts, Systems & Application, Priti Sinha, Pradeep K., Sinha , BPB Publications

5. The Book of R: Tilman M. Davies, San Francisco

Unit-2: Descriptive Statistics

2.1. Introduction

Statistical Methodology is a comprehensive term which includes almost all the methods involves in the collection, processing condensing and analyzing of data. The data collected from the field for a number of items vary greatly in their qualitative as well as quantitative nature. For example, the rainfall at a particular region is erratic in nature and shows variation from year to year, month to month and even day to day. The condensation of data in terms of maps, charts, diagrams, etc. is a first and necessary step in rendering a long series of observation comprehensible. But for practical purpose it is not enough, particularly when we want to compare two or more different series of data, e.g. we may wish to compare the distribution of status in two races of man, or the birth rates in India in two successive decades or rainfall in two different regions, or the number of wealthy people in two different countries. For such problems, there are certain statistical techniques one of which is a measure of central tendency.

It is found that the observations have a tendency to cluster round a central value, and this characteristic of observations is called the "**Central tendency".** Any statistical measure which gives the point round which the observation has a tendency to cluster is known as '**measure of central tendency'.** The central value of the variable in any series of observations is useful in finding the location of the distribution and so it is also called an average. Thus an '**average**' of a series of measurements is a single value of the variable which is stationary representative of the distribution.

In this unit have been highlighted different measures of central tendency are covered. Various situations where they find calculation of these measures for ungrouped and grouped data are described.

Objectives

After studying this unit you will be able to

- ➢ Understand the meaning of central tendency of data
- ➢ Compute common measures of central tendency, i.e. mean, median and mode.
- ➢ Compute the various measures of partitions of data such as quartiles, deciles and percentiles.
- ➢ Understand how to choose proper measure of central tendency

2.2. Measure of Central Tendency

If you take a close look at any data set, you would notice that though the manifestation of the variable is different for different observational units, the values tend to cluster around a central value. This property is referred to as **central tendency.**

A representative value around which a given set of observations tends to cluster (or equivalently be located) is a measure of central tendency or location or is simply an average. **Arithmetic mean** (a.m.), **median** and **mode** are the three commonly used averages. Other averages are **geometric mean** and **harmonic mean.**

The measures of central tendency or averages are of different types, but the most common in use are of three types:

1. Mean
2. Median
3. Mode

The mean is further classified as:

(i) Arithmetic mean

(ii) Geometric mean

(iii) Harmonic mean

Since each one of the above measures of central tendency has its own individual characteristics and properties, a decision must always be make as to which would be the most appropriate and useful in view of the nature of the statistical data and purpose of the inquiry. The qualities desired in a measure should be (a) rigidly defined, (b) easily computed, (c) capable of a simple interpretation (d) not unduly influenced by one or two extremely large or small values and (e) likely to fluctuate relatively little from one random sample to another (of the same size and from the same population). However the decision about which of the three measures of central tendency to use will be clear after learning the computation of each one. A few general considerations in choosing a measure of central tendency are: (i) the purpose of research- what characteristic of the data are of interest; (ii) the level of measurement of the data- nominal, ordinal, interval or ratio level; (iii) the shape of the frequency distribution as indicated by a graph- symmetric or skewed; (iv) level of expertise of the researcher and the audience- what can you accomplish and what your audience is able to understand.

2.2.1. Arithmetic Mean (Ungrouped Data)

Case (i) Simple Case:

The arithmetic mean of a series of n observations X_1, X_2, X_3, ……X_n is obtained by summing up the values of all the observations and dividing the total by the number of observations. Thus,

$$X = \frac{Sum\ of\ observations\ (or\ values)}{Number\ of\ obsevations}$$

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

Where ∑ (sigma) stands for summations and x_i is the i^{th} value of the observation (variable).

Case (ii) when Frequency is given:

If the value X_1 occurs f_1 times, the value X_2 occurs f_2 times,…..the value X_n occurs f_n times, then arithmetic mean of a series of n observations X_1, X_2, X_3, ……X_n is obtained by summing up the values of all the observations and dividing the total by the number of observations. Thus,

$$\bar{X} = \frac{x_1\mathrm{f1} + x_2\mathrm{f2} + x_3\mathrm{f3} \ldots \ldots + x_n\mathrm{fn}}{n}$$

$$= \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

$$= \frac{\sum_{i=1}^{n} f_i x_i}{N}$$

Where $\sum_{i=1}^{n} f_i = N$ and ∑ (sigma) stands for summations and x_i is the i^{th} value of the observation (variable) and f_i is the i^{th} value of the observation (variable).

2.2.1.1. Short-Cut Method

(I) Individual series:

This method is applied to avoid lengthy calculations. When the individual set of reading are large in size, an arbitrary value is selected as a working mean (known as assumed mean) and differences between the working mean and the individual readings (known as deviations from the assumed mean) are worked out. By summing these differences dividing by the number of readings we can get the mean of deviations from assumed mean. Let $x_1 + x_2 + x_3 \ldots \ldots + x_n$ be n individual reading on the variable and let A be the working mean. Let d_1, d_2, d_3…..d_n denote the differences between the working mean and individual values $x_1 + x_2 + x_3 \ldots \ldots + x_n$ respectively. The mean $\bar{x}$ of X, in terms of the mean $\bar{d}$ of differences, is calculated as:

$$d = \frac{d_1 + d_2 + d_3 \ldots \ldots + d_n}{n}$$

$$= \frac{(x_1 - A) + (x_2 - A) + (x_3 - A) + \cdots \ldots + (x_n - A)}{n}$$

$$= \frac{\sum_{i=1}^{n} x_i - nA}{n} = \frac{\sum_{i=1}^{n} X_i}{n} - A$$

$$= \bar{X} - A$$

Or,

$$\bar{X} = A + \bar{d} = A + \frac{1}{n}\sum_{i=1}^{n} d_i$$

True mean guessed mean + (sum of deviations from guessed mean/number of cases)

This is useful, if the size of frequencies is large. An illustration is given below.

Example 1.2

Calculation the mean for the following scores: 60, 65, 74, 85, 95.

Solution:

Table 2.1: Distribution of Scores

x_i (Scores)	**x_i-74**
60	-14
65	-9
74	0
85	+11
95	+21
	+9

Then, mean score is

$$\bar{x} = 74 + (+9/5) = 74 + 1.8 = 75.8$$

(II) Discrete Frequency Distribution:

Let discrete frequency distribution be

$$x : x_1 + x_2 + x_3 \dots \dots + x_n$$

$$f : f_1 + f_2 + f_3 \dots \dots + f_n$$

Let *A* be the working mean and d is the deviation of x , i.e d_i= $(x_i - A)$ and M be the A.M. of x, then

$$\frac{\sum_{i=1}^{n} f_i d_i}{N}$$

$$= \frac{\sum_{i=1}^{n} f_i (x_i - A)}{N}$$

$$= \frac{f_1(x_1 - A) + f_2(x_2 - A) + f_3(x_3 - A) + \cdots \dots + f_n(x_n - A)}{N}$$

$$= \frac{\sum_{i=1}^{n} f_i x_i - f_i A}{N} = \frac{\sum_{i=1}^{n} f_i X_i}{N} - A$$

$$= \bar{X} - A$$

Or,

$$M = A + \bar{d} = A + \frac{1}{N} \sum_{i=1}^{n} f_i d_i$$

Where $\sum_{i=1}^{n} f_i = N$

22.1.2. Theorems on Arithmetic Mean

1. First Property of Mean:

The sum of the deviations about the arithmetic mean equals zero.

Mathematically

$$\sum [f_i(x_i - \bar{x})] = 0$$

Proof:

$$\sum f_i(x_i - \bar{x}) \quad = \sum_i f_i x_i - \sum_i f_i \, \bar{x}$$

$$= \sum_i f_i, x_i - \bar{x} \sum_i f_i = (Since\ \bar{x}\ be\ independent\ of\ i)$$

$$= N\bar{X} - \bar{X}N = 0$$

Since,

$$\left(\bar{x} = \frac{1}{N} \sum_i f_i, x_i \ where\ N = \sum_i f_i \qquad then \sum_i f_i, x_i \ = N\,\bar{x} \right)$$

Hence proved

This property says that if the mean is subtracted from each score, the sum of the differences will equal zero. The property results from the fact that the mean is the balance point of the distribution. The mean can be thought of as the fulcrum of a seesaw. When the score are distribution along the seesaw according to their values, the mean of the distribution occupies the position where the scores are in balance. This is known as first property of mean.

2. Second Property of Mean:

The sum to the squared deviations of all the scores about their arithmetic mean is minimum.

That is,

$$\sum [f_i(x_i - \bar{x})]^2 = minimum$$

This is an important characteristic and is used in many areas of statistics, particularly in regression analysis.

Proof: Let us suppose that the sum of square of deviations from point (a).

$$\sum f_i(x_i - a)^2 = k$$

According to the principle of maxima and minima, *k* will be minimum if,

$$\frac{\partial k}{\partial a} = 0 \qquad and \qquad \frac{\partial^2 k}{\partial a^2} > 0$$

$$Now\ \frac{\partial k}{\partial a} = (-2)\sum f_i(x_i - a) = 0$$

$$\rightarrow \sum f_i(x_i - a) = 0 \rightarrow \sum f_i x_i - Na = 0$$

$$a = \frac{1}{N}\sum f_i x_i = \bar{x}$$

Again

$$\frac{\partial^2 k}{\partial a^2} = (-2)\sum f_i(-1) = 2\sum f_i = 2N > 0$$

Hence K is minimum at a = X and

$$\sum [f_i(x_i - \bar{x})]^2 = minimum$$

Remarks we shall see later on that,

$$\sigma^2 = \frac{1}{N}\sum f_i(X_i - \bar{X})^2$$

is a measure of dispersion.

3. Third Property of Arithmetic Mean:

Arithmetic Mean is not independent of change of origin and scale.

Proof : Let

x	x_1	x_2	x_3	.	.	x_n
f	f_1	f_2	f_3	.	.	f_n

be a discrete frequency distribution. Then

$$\bar{X} = \frac{x_1\text{f1} + x_2\text{f2} + x_3\text{f3} \ldots \ldots + x_n\text{fn}}{n}$$

$$\frac{\sum_{i=1}^{n} f_i X_i}{N\sum_{i=1}^{n} f_i}$$

Let $u = \frac{(x-a)}{h}$ changing origin and scale. Then

$$\bar{u} = \frac{\sum_{i=1}^{n} f_i u_i}{\sum_{i=1}^{n} f_i}$$

$$= \frac{\sum_{i=1}^{n} f_i \frac{(x-a)}{h}}{\sum_{i=1}^{n} f_i}$$

$$= \frac{(\bar{X}-a)}{h}$$

Example: Compute the mean of the following by both direct and short-cut method.

Height in c.m	219	216	213	210	207	204	201	198	195
Men	2	4	6	10	11	7	5	4	1

Solution:

Ht in c.m	**f**	**fx**	**D=(x-A)**	**fd**
219	2	438	12	24
216	4	864	9	36
213	6	1278	6	36
210	10	2100	3	30
207	11	2277	0	0
204	7	1428	-3	-21
201	5	1005	-6	-30
198	4	792	-9	-36
195	1	195	-12	-12
Total	50	10377		27

(I) By direct method :

$$M = \frac{\sum_{i=1}^{n} f_i x_i}{N}$$

$$= \frac{10377}{50} = 207.540\ cm$$

(II) By short-cut method :

$$M = A + \frac{1}{N}\sum_{i=1}^{n} f_i d_i$$

$$= 207 + \frac{27}{50} = 207.540\ cm$$

2.2.1.3. Combined (Additive) Property of Mean:

If $\bar{X}_1$ and $\bar{X}_2$ be the means of two series of sizes n_1 and n_2 respectively, then the mean of the combined series can be computed as:

$$\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$$

Proof:

If $\bar{X}_1$ be the mean of series X_{11}, X_{12}…..X_{1n1}

And

$\bar{X}_2$ be the mean series X_{21}, X_{22}……..X_{2n2}

Then by definition,

$$\bar{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_i} X_i = \frac{1}{n_1}(\mathrm{X}_{11}, \mathrm{X}_{12} \ldots\ldots \mathrm{X}_{1n1})$$

$$\bar{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_i} X_i = \frac{1}{n_2}(\mathrm{X}_{21}, \mathrm{X}_{22} \ldots\ldots \mathrm{X}_{2n2})$$

The combined series is $\mathrm{X}_{11}, \mathrm{X}_{12} \ldots\ldots \mathrm{X}_{1n1}\mathrm{X}_{21}, \mathrm{X}_{22} \ldots\ldots \mathrm{X}_{2n2}$

The mean is

$$\bar{X} = \frac{1}{n_1 + n_2}[(\mathrm{X}_{11}, \mathrm{X}_{12} \ldots\ldots \mathrm{X}_{1n1}) + (\mathrm{X}_{21}, \mathrm{X}_{22} \ldots\ldots \mathrm{X}_{2n2})]$$

$$\bar{X} = \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1} X_i + \sum_{j=1}^{n_2} X_j\right]$$

$$= \frac{1}{n_1 + n_2} [n_1 \bar{X}_1 + n_2 \bar{X}_2]$$

Similarly, if $\bar{X}_1, \bar{X}_2, \bar{X}_3 \dots \dots . . \bar{X}_k$ be the means of k series of sizes $n_1, n_2, n_3 \dots \dots . n_k$ respectively then the mean $\bar{X}$ of combined series is

$$\bar{X} = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2 + \cdots \dots \dots . n_n \bar{X}_n}{n_1 + \ n_2 + \ \dots \dots . n_n}$$

Example 1.9

The average ages of 250 males and 210 females in a village are 41.6 and 38.5 years respectively. Find the average age combining both males and females together.

Solution:

Here (Combined average) is N$_1$= 250, $\bar{X}_1$= 41.6 years and N$_2$= 210, $\bar{X}_2$= 38.5 years. Therefore,

$$\bar{X} = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2}{n_1 + n_2}$$

$$= \frac{250 \times 41.6 + 210 \times 38.5}{250 + 210} = \frac{10400 + 8085}{460}$$

$$= 40.18 \, years$$

Example: Compute the weighted mean of the first natural numbers whose weights are equal to the squares of the corresponding number.

x	**1**	**2**	**3**	**.**	**.**	**.**			**n**
Men	$(1)^2$	$(2)^2$	$(3)^2$	.		.	.	.	$(n)^2$

Solution:

$$\sum_{j=1}^{n} W_j = \sum_{j=1}^{n_2} (n)^2 = (1)^2 + (2)^2 + \cdots \ldots .. (n)^2$$

$$= \frac{n(n-1)+(2n+1)}{6}$$

$$\sum_{j=1}^{n} W_j\, x_j = \sum_{j=1}^{n_2} (n)^3 = (1)^3 + (2)^3 + \cdots \ldots .. (n)^3$$

$$= \frac{\big(n(n+1)\big)^2}{(n)^2}$$

$$\bar{X}_1 = \frac{\sum_{j=1}^{n} W_j\, x_j}{\sum_{j=1}^{n} W_j}$$

$$= \frac{3n(n+1)}{2(2n+1)}$$

2.2.1.4. Advantages of Mean

i. The mean is sensitive to the exact values of all the scores in the distribution. Since you have to add the scores to calculate the mean, a change in any of the scores will cause a change in the mean.

ii. Mean is very sensitive to extreme scores. If we add an extreme score (one that is very far from the mean), it would greatly disrupt the balance. The mean would have to shift a considerable distance to reestablish balance. The mean is more sensitive to extreme than is the median or the mode. This known as 2nd property of mean.

iii. Of the measures used for central tendency, the mean is least subject to sampling variation under most circumstances. If repeated samples are drawn from a population, the mean would vary from sample to sample. The same is true for the median and the mode. However the mean varies less than these other measures of central tendency. This is very important in inferential statistics and is a major reason why the mean is use in inferential statistics whenever possible.

iv. It takes into account all the scores in a distribution so; mean offers a good representation of the central tendency by making use of the most information.

v. Mean is used in many statistical formulas, making it a more widely used measure.

2.2.1.5. Limitations of Mean

Mean can be misleading if there are extreme values in the distribution, for example, if the distribution is skewed (asymmetrical) of the level of measurement is less than interval. Sometimes people are interested in misleading others by making use of ‘illegitimate’ statistics. The following example illustrates this point.

Example 1.10

The amounts of money that a sample of people contributed to political campaigns in the last election were, in thousands rupees; 1,2,5,25,10,0,2,0,5,10, 500. Calculate the mean.

Solution:

Make a table that contains columns: x_i, $f_i and\ f_i x_i$ Sum all the entries in columns $X_i, f_i and\ f_i x_i$ to get:

$$f_i x_i = 560, \quad f_i = n = 11$$

$$\text{Mean} \quad \bar{X} = \frac{f_i x_i}{n} = \frac{(1+2+5+25+\cdots+\cdots+10+500)}{10}$$

$$= \frac{560}{11} = 5091\ or\ Rs.50{,}910$$

A mean of Rs. 50.91 thousands suggests that the typical contribution was Rs. 50,910. We notice that the mean is this example is not at all a measure of central tendency. All but one person contributed less than the mean. This is due to the presence of an extreme contributor who contributed Rs. 500, 000. This high value inflates the mean making it a misleading statistics in each situation.

2.2.2. Median

Median is another important and useful measure of central tendency. It has connotation of the middle most or most central value of a set of measurements. When the observations are arranged in ascending or descending order of magnitude, then the middle value is called the median of this observation. It is usually defined as the value which divides a distribution in such a manner that the number of items below it is equal to the number of items above it. The median is thus a positional average. It is better indication of central tendency when one or two of the peripheral readings are too large or too small because they give the wrong idea of the average when mean is computed. Median is that variate value of the data or frequency distribution which divides it in two equal halves.

2.2.2.1. Calculation of Median (Ungrouped data)

Case 1 : Formula in case of individual series: let X_1, X_2, X_3, ……X_n be a n observations written in ascending order of magnitude then median denoted by obtained by M_e or M or Md is given by

Median = value of the middle items

(i) when n is odd: In case of ungrouped data when the number of observations are odd, then median is the middle value after the measurements have been arranged in ascending or descending order of magnitude, i.e. if there are n number of measurements and measurements are arranged in ascending or descending order of magnitude, the median of the measurements is $\left(\frac{n+1}{2}\right)^{th}$ measurement where n is an odd number.

(ii) when n is even*:* If the numbers of observations are even, median is defined as the mean of the two middle observations, when observations are arranged in ascending or descending order to magnitude i.e.

$$median = \frac{\left(\frac{n}{2}\right)^{th} value + \left(\frac{n+1}{2}\right)^{th} value}{2}$$

Example 1.16 Calculate median for the following data:

(a) 68, 62, 75, 82, 68, 71, 68, 71, 62, 68, 74, 59, 74, 68, 60, 71, 59, 73, 73, 58.

(b) 200, 150, 260, 285, 380, 305, 4989, 307, 1280, 233, 403

Solution (a)

To compute the median first we arrange the values in ascending order of magnitude as:

58, 59, 59, 60, 62, 62, 68, 68, 68, 68, 68, 71, 71, 71, 73, 73, 74, 74, 75, 82.

The number of observation n is even in this case, i.e., n=20

So

$$median = \frac{\left(\frac{n}{2}\right)^{th} value + \left(\frac{n+1}{2}\right)^{th}\ value}{2} = \frac{10th\ value + 11th\ value}{2}$$

$$= \frac{68+68}{2} = 68$$

(b) Let us first arrange the values in ascending order of magnitude as:

150, 200, 233, 260, 285, 305, 307, 380, 403, 1280, 4989

The number of observations n in this case is odd, i.e., n=11 so the median is the $\left(\frac{n+1}{2}\right)^{th}$ value i.e., $\left(\frac{11+1}{2}\right)^{th}$ or 6th value of the observation and thus underlined value. i.e., 305 is the median.

2.2.2.2. Calculation of Median (Grouped Data)

In case of discrete frequency distribution median can be obtained with the help of cumulative frequencies as follows:

(i) First find N/2 where N=$\sum f$

(ii) Find the cumulative frequency just greater then N/2.

(iii)Corresponding value of X (i.e., of variable) is median.

Example 1.17

Calculating the median height from the data given in example 1.4.

Calculation

(a)

Height (in metre) (X)	**Number of units (f)**	**Cumulative frequency (f)**
200	142	142
600	265	407
1000	560	967
1400	271	1238
1800	89	1327
2200	16	1343
Total	1343	

Here f=N= 1343; $\frac{N}{2} = 671.5$

The cumulative frequency just greater than 671.5 is 967 and corresponding to this cumulative frequency, the value of X is 1000 and thus the median height is 1000 meters.

Median (Continuous Grouped Data): Median for such distribution is computed by the following formula

$$Median\ (Md) = l_m + \frac{\left(\frac{n}{2} - f_e\right)}{f_m}.h$$

Where l_m is the lower limit f_m is the frequency of the median class, f_e is the cumulative frequency of the class, preceding the median class and h is the class width of the median class and N =$\sum f_i$

Example 1.18

Calculate median for the following grouped data.

Interval	5-45	5-55	5-65	5-75	75-85
Frequency	2	3	5	1	1
Cum. Freq.	2	5	10	11	12

The cum. Freq. is computed from the given freq. dist.

The median position = (n+1)/2= (12+1)/2 = 6.5

Median lies between observation 55 and 65. Both of these observations fall in category 3, i.e., in class (35-65) with cumulative frequency of 10. Therefore,

$$Median\ (Md) = l_m + \frac{\left(\frac{n}{2} - f_e\right)}{f_m} . h$$

Where l_m =55, f_m =5, h=10, fe =5, N=12

$$Median\ (Md) = 55 + \frac{[(12/2) - 5]}{5} \times 10 = 55 + 2 = 57$$

Example 1.19

The following table gives the size of land holding of families in a village. Find out the median holding size.

Area of land (in acres)	5-9	10-14	15-19	20-24	25-29	30-34
No. of families	20	35	150	70	44	38

Solution:

Since the class groups are given in the discrete from hence we first have to convert it into continuous form by adding 0.5 to the upper limits and subtracting 0.5 from the lower limits as given in column 2 below:

Area of land (in acres) (X)	**Area of land (in acres) Continuous from**	**No. of families (f)**	**Cumulative frequency (f)**
5-9	4.5-9.5	20	20
10-14	9.5-14.5	35	55
15-19	14.5-19.5	150	205
20-24	19.5-24.5	70	275
25-29	24.5-29.5	44	319
30-34	29.5-34.5	38	357
Total		357	

Here N/2= 178.5, the cumulative frequency just greater than 178.5 is 275 and the corresponding class group is the median class. For this median class, we have

Where l_m =14.5, f_c =55, h=5, f_e =150

$$Median\ (Md) = 14.5 + \frac{[178.5 - 55]}{150} \times 5 = 14.5 + 4.12 = 18.62\ acres$$

2.2.2.3. Calculation of Median by (Graphical Method)

One of the methods to compute median is the graphical method. In this case take the class intervals (or the individual readings) on the axis of X and plot the corresponding cumulative frequencies on the axis of Y against the upper limit of the class interval (or against the variants value in case of discrete frequency distribution). The Curve obtained by joining the points by means of free hand drawing is the cumulative frequency curve or give. For the calculation of median take a point on the axis of Y that is equivalent to N/2 and form this point draw a line parallel to X-axis. This line will cut the curve and form the cutting point draw a line perpendicular on X axis. This distance from origin to the point at which the perpendicular line cuts the X axis is the value of median.

Example 1.20

Find out the median rainfall from the distribution given in example 1.13 by the graphical methods:

Figure 3.1 shows the cumulative frequency curve formed between the upper limits of classes and corresponding cumulative frequencies. The point *N/2* is shown on Y axis and the dotted line parallel to X-axis cuts the cumulative curve at C. Perpendicular line cuts the X-axis at M. The distance *OM* is the median value.

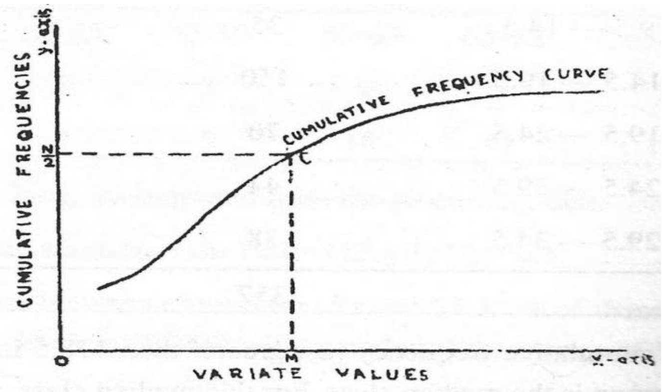


Fig. 3.1 Cumulative Frequency Curve

2.2.2.4. Advantages of Median

An important advantage of median is that it is less sensitive than the mean to extreme scores. For skewed data, the median is a better choice because it is usually not affected by a few outliers. Median is also a desirable measure when the distribution has to be truncated for some reasons. If the purpose is to describe the central tendency of a set of scores, the median is preferable to other measures. It gives an undistorted picture of central tendency whether the data are skewed or not.

2.2.2.5. Disadvantages of Median

Under usual circumstances the median is more vulnerable to sampling variable than the arithmetic mean. This makes median less stable than the mean from sample to sample and therefore it is not very useful in inferential statistics. For ordinal data median also ignores the actual values of observations and simply takes into account their positions.

In some circumstances the mean is a better measure than the median, and in others the Converse is true. The following factors contribute to the determination of whether the mean or the median should be used.

Uses of the Mean and the Median

Both the mean and the median are important and useful measures of central tendency.

Sensitivity to Extreme Observations: The median is often preferred over the mean when the latter can be influenced strongly by extreme observations. Consider for instance, the computation of an average income of the families in an apartment building containing 14 families, 3 of which earn Rs. 10,000 per month, 5 of which earn Rs.12.000 per month, 5 earn Rs.15,000 per month and 1 of which earns Rs. 1 lakh per month. Then the mean income per month in rupees of the 14 families equals Rs. 18,929 approx.

However, this figure is not a very good description of the monthly income level of the majority of the families in the building. A better measure might be the median, which in this case is Rs. 12,000 per month. The median is much less affected by the one extreme value.

Open Ended Class Intervals: It may happen that in a frequency distribution some intervals do not have finite upper or lower limits. For example, in a frequency distribution of the monthly income of families, two class intervals might be "less than Rs.

15,000" and "Rs.30,000 and more". Each of these class intervals is open-ended. With such class intervals, there may be no alternative but to use the median, since calculation of the mean requires knowledge of the sum of the measurements in the open-ended classes.

Mathematical Convenience: The mean rather than the median is often the preferred measure of central tendency because it possesses convenient mathematical properties that the median lacks. For instance, the mean of two combined populations or samples is a weighted mean of the means of the individual populations or samples. On the other hand, given the medians of two populations or samples, there is no way to determine what the median of the two populations combined or two samples combined would be.

Extent of Sampling Variation: Sample statistics such as the sample mean or the sample median are often used to estimate the population mean. A major reason for preferring the mean to the median is that the sample mean tends to be more reliable than the sample median in estimating the population mean. In other words, the sample mean is less likely than the sample median to depart considerably from the population mean.

Partition Value: These are the values of the variants which divide the total frequency into a number of equal parts. Some of them are quartiles, deciles and percentiles.

Quartiles: Quartiles are three values that split sorted data into four parts, each with an equal number of observations. Quartiles are a type of quartile.

First quartile: Also known as Q1, or the lower quartile. This is the number halfway between the lowest number and the middle number.

$$(Q_1) = l + \frac{\left(\frac{N}{4} - F\right)}{f} . h$$

Where l = lower limit of median class

f= the frequency of the median class.

h= width of the median class

F= the cumulative frequency of the class preceding the median class, that is total of all frequencies before the median class.

N= total frequency

Second quartile: Also known as Q2, or the median. This is the middle number halfway between the lowest number and the highest number.

$$(Q_2) = l + \frac{\left(\frac{N}{2} - F\right)}{f} . h$$

Where l = lower limit of median class

f= the frequency of the median class.

h= width of the median class

F= the cumulative frequency of the class preceding the median class, that is total of all frequencies before the median class.

N= total frequency

Third quartile: Also known as Q3, or the upper quartile. This is the number halfway between the middle number and the highest number.

$$(Q_3) = l + \frac{\left(\frac{3N}{4} - F\right)}{f} . h$$

Where l = lower limit of median class

f= the frequency of the median class.

h= width of the median class

F= the cumulative frequency of the class preceding the median class, that is total of all frequencies before the median class.

N= total frequency

Merits and limitations: QD is a simple measure of dispersion. While the measure of central tendency is taken as median, QD is most relevant to find out the dispersion of the distribution. In comparison to range, QD is more useful because range speaks about the highest and lowest scores while QD speaks about the 50% of the scores of a distribution. As middle 50% of scores are used in QD there is no effect of extreme scores on computation, giving more reliable results.

In case of open-end distribution QD is more reliable in comparison to other measures of dispersion. It is not recommended to use QD in further mathematical computations. It is not a complete reliable measure of distribution as it doesn't include all the scores. As QD is based on 50% scores, it is not useful to study in each and every statistical situation

Deciles: Deciles are those values of the variants which divide the total frequency into 10 equal parts.

Thus the first deciles(D_1) is given by,

$$(D_1) = l + \frac{\left(\frac{N}{10} - F\right)}{f} . h$$

Where l = lower limit of median class

f= the frequency of the median class.

h= width of the median class

F= the cumulative frequency of the class preceding the median class, that is total of all frequencies before the median class.

N= total frequency

The second deciles(D_2) is given

$$(D_2) = l + \frac{\left(\frac{2N}{10} - F\right)}{f} . h$$

Similarly

The ninth deciles(D_9) is given

$$(D_9) = l + \frac{\left(\frac{9N}{10} - F\right)}{f} . h$$

Percentiles: Percentiles are those values of the variants which divide the total frequency into 100 equal parts.

Thus the first Percentiles (P_1) is given by

$$(P_1) = l + \frac{\left(\frac{N}{100} - F\right)}{f}.h$$

Where l = lower limit of median class

f= the frequency of the median class.

h= width of the median class

F= the cumulative frequency of the class preceding the median class, that is total of all frequencies before the median class.

N= total frequency

Thus the first Percentiles (P_2) is given by

$$(P_2) = l + \frac{\left(\frac{2N}{100} - F\right)}{f}.h$$

Similarly

Thus the first Percentiles (P_{99}) is given by

$$(P_{99}) = l + \frac{\left(\frac{99N}{100} - F\right)}{f}.h$$

Location of Quartiles: location of the point $\frac{N}{4}$ on the Y-axis from this point draw a straight line parallel to the X-axis meeting the polygon at the point A1, say from A1 draw a perpendicular A1Q1 on X-axis, then the distance OQ1 of the point Q1 from the origin is the required first quartile.

Similarly, Locating the points $\frac{2N}{4}$ and $\frac{3N}{4}$ on the Y-axis and drawing a straight line parallel to the X-axis meeting the polygon at the point A2 and A3, say from A2 draw a perpendicular A2Q2 and from A3 draw a perpendicular A2Q3 on X-axis, then the distance OQ2and OQ3 are the required second and third quartiles respectively.

Location of Deciles: locate the point on Y-axis$\frac{N}{10}, \frac{2N}{10}$……..etc. and proceed as in case of Quartiles.

Location of Percentiles: locate the point $\frac{N}{100}, \frac{2N}{100}$, ……..etc. on Y-axis and proceed as in case of Quartiles.

2.2.3. Mode

Mode is defined as the most frequent observed valve of the measurements in the relevant set of data. In a set of observations, if all the observations are distinct so that each of these occur with frequency 1, then it will be meaningless to say each of them is a mode; as such, in such a situation we say that the mode does not exist. However, from the definition, it is clear that a given data set may have more than one mode.

Mode or modal value of a distribution is that the value which occurs most frequency, For example, at any stations the average number of occurrences for thunder storms or days with snowfall wind direction, etc are the most realistically presented by modal value. In case of frequency distribution the mode is that value which has maximum frequency. If two or more observations occur the same number of times then there is more than one mode and the distribution is called multi-model as against uni-model.

2.2.3.1. Calculation of Mode (Ungrouped Data)

Mode is defined as that variants value of the data or the frequency distribution which occurs most frequently. The mode in a series of individual measurements can be located either of two ways.

i. Data should first be placed in an array so that repetition of a value can be identified and quickly counted; the value of that item which occurs most of the times is the modal value.

ii. Data should be converted into a discrete series.

Example 1.21

Find the modal temperature value from the values given in example 1.14

Solution (i) Putting data in array as:

58, 59, 59, 60, 62, 62, 68, 68, 68, 68, 68, 71, 71, 71, 73, 73, 74, 74, 75, 82.

Here mode = 68^0 F.

(ii) Discrete series (converted to frequency distribution form)

Variable (X)	58	59	60	62	68	71	73	74	75	82
Frequency (f)	1	2	1	2	5	3	2	2	1	1

Here the value 68 occurs the maximum number of times, hence it is mode.

2.2.3.2. Discrete Series (Grouped Data)

In case of discrete frequency distribution, mode can be located by inspection of the distribution alone. The size having the maximum frequency will be reckoned as mode.

Example 1.22

Computed the modal size of children born per family in the locality from the data given in example 1.3

Solution:

The highest size of frequency in the given distribution is 154 and corresponding to this frequency the number of children born per family is 2. Hence the modal size of children born per family in the locality is 2.

2.2.3.3. Continuous Series (Grouped)

i. Determine the modal class interval. It is the data class interval with the maximum number of frequencies in it. This can be found out by just observing the series.

ii. Determine the value of mode by applying the following formula:

$$Mode = L + \left(\frac{f - f_P}{2f - f_P - f_s}\right) h$$

Where

L is the lower limit of the modal class; *f* is the frequency of the modal class; f_P is the frequency of the class preceding the modal class ;f_s is the frequency of the class succeeding the modal class and h is the class width of the modal class.

Example 1.23

Compute the modal agricultural holding of the village from the data given in example 1.19,

Solution:

The maximum frequency 'f' in the distribution is 150 which corresponds to class group 15-19, i.e., 14.5-19.5 in continuous case (see column 2, example 17). Hence modal class is 14.5-19.5. Now mode is computed as:

$$Mode = L + \left(\frac{f - f_P}{2f - f_P - f_s}\right) \times h$$

Here $L= 14.5$, $f= 150$, $f_P= 35$, $f_S= 70$ *and* $h=5$.

$$Mode = 14.5 + \left(\frac{150 - 35}{150 \times 2 - 35 - 70}\right) \times 5$$

$$= 14.5 + \frac{115}{195} \times 5 = 17.45\ acres.$$

Sometimes mode is also computed with the help of mean and median. For a symmetrical distribution mean, median and mode coincide and if the distribution is moderately asymmetrical, the mean, median and mode are approximately related by the formula:

Mode ≅ 3 Median - 2 Mean

Example 1.24

If the mean and median of a moderately asymmetrical series are 12.9 and 12.1 respectively, what would be its most probable mode?

Solution:

Mean = 12.9, Median = 12.1, Mode?

Mode=3 Median-2 Mean

$= 3\times12.1-2\times12.9$

$=36.3-25.8$

$=10.5$

2.3. Measures of Dispersion

A measure of central tendency, gives us a general idea about the average value or the magnitude of the observations. However, two distributions through may have the same mean, say, may differ in respect of several other characteristics.

Suppose that the average score obtained by students in 10-marks class test is 5 and suppose it is further known that the scores varied between **3** and 6. With such information in view, one can predict his performance with much more confidence than when the scores are known to have varied between 2 and 8.

This example suggest that the average along with an idea about the scatter or spread of variation or dispersion of the observations about the average give us a more complete picture about the state of affairs than the average alone. The less is the range, the more will be the concentration of observations around the mean. We shall now discuss as to how the dispersion of the observations about the average can be measured

Dispersion is defined as the degree to which scores deviate from the central tendency (usually the mean) of the distribution. The statistical techniques that quantify this dispersion in a distribution are called measures of dispersion. Most commonly used measures of dispersion are range, average deviations, variance and standard deviation.

There are two types of measures of dispersion; distance measures and measures of average deviation

2.3.1. Range

Range is defined as the difference between the highest and the lowest scores in a distribution. Symbolically,

$$R= X_{max} - X_{min}$$

Where R is the range, X_{max} is the highest score, X_{min} is the lowest score.

A large value of range indicates greater dispersion and a small value of range indicates lesser dispersion among the scores. Minimum value that range can achieve is 0 and the maximum is infinity. If all the scores are the same, R will have a value of 0 and hence there is no dispersion.

Example 2:

Find range for value: 87, 92, 47, 58, 87, 62, 73, 73, 61.

Solution:

It is always a good idea to first rank the observation in ascending or descending order. In an ascending order the scores are: 47, 58, 61, 62, 73, 73, 87, 87, 92. A visual examination shows that X_{max}= 92; X_{min}= 47.

Therefore R=92-47=45

2.3.2. Advantages of Range

i. Range gives a quick identification of dispersion. It can be a good measure if there are no outliers in the data that means the distribution is not skewed.

ii. Range is easy to compute and interpret. For variables measured at an ordinal scale, range is the only measure which is technically meaningful.

iii. If the data are to be presented to a relatively unsophisticated audience, the range may be the only measure of dispersion that will be readily understood.

2.3.3. Disadvantages of Range

Calculation of range is based only on two extreme scores, the minimum and the maximum. The rest of the data are ignored.

i. Range tells nothing about the dispersion among intermediate scores.

ii. Range is greatly affected by outliers. Thus for skewed distributions, range is usually very misleading measure.

iii. Since range ignores all the scores except the two extreme scores it cannot be used for making inferences about populations.

iv. Range varies considerably from sample to sample.

2.3.2. Mean Deviation

Average deviation is found by summing the absolute values of the deviations and dividing the sum by number of observations. The formula for average deviation can be written as:

$$MD = \left(\sum |x_i - \bar{x}|\right)/n$$

Where $\bar{x}$ is the arithmetic mean; $x_i - \bar{x}$ is deviation of from $\bar{x}_i$ and $|x_i - \bar{x}|$ is the absolute value of the deviation which is always a positive number. The average deviation tells the distance with which a score will typically deviate from the mean.

Example 2.2

The numbers of terms that five randomly selected Members of Parliament have served are: 3, 10, 12, 7, 8. Find the average deviation of these scores.

Solution:

Make the following table containing the calculation.

Case number	**Terms**	$\boldsymbol{x_i - \bar{x}}$	$\boldsymbol{\lvert x_i - \bar{x}\rvert}$
1	3	3-8=-5	5
2	10	10-8=2	2
3	12	12-8=4	4
4	7	7-8=-1	1
5	8	8-8=0	0
Total	40	0	12

Mean $\bar{x}$ = 40/5=8

Sum of deviation from the mean $(x_i - \bar{x})$ = 0

Sum of absolute deviations $|x_i - \bar{x}|$ = 12. Therefore

MD= 12/5=2.4 terms.

For descriptive purposes, the average deviation can be an adequate and easily interpretable measure for describing the degree of dispersion. But the mathematical properties of average deviation are such that it does not meet the needs of advanced mathematics. Therefore, average deviation is a very infrequently used measure of dispersion.

2.3.3. Variance

Instead of taking absolute values of deviations to remove the negative signs and obtain a nonzero sum, another way to get rid of negative sign of deviations is to square them. Square of a negative number is a positive quantity. A statistic called variance uses this approach. To calculate variance, calculate the deviations, square each deviation, add up the squared deviations to obtain sum of squares and divide this sum of squares by the number of deviations. The resulting quantity is called the mean squared deviation (MSD) or the variance of the distribution of scores.

Variance can be of two types;

(i) Sample variance denoted by S^2, calculated from sample data.

(ii) Population variance denoted by σ^2 calculated from population data.

In practice S^2, a statistic, is always known while σ^2 , a parameter is seldom known. Therefore, S^2 is used as an estimate of σ^2 . While dealing with several variables, It proves to be convenient to attach a subscript to s or σ. The subscript indicates the name of variable for which variance is being calculated. Thus S_x^2 is the sample variance of the variable X, S_y^2 is the sample variance of Y, and so on.

2.3.3.1. Computation of Variance

Variance may be defined as the mean squared deviation of scores around the mean. In the form of a formula, variance is given by:

$$S_x^2 = \sum \frac{(x_i - \bar{x})^2}{\text{n}} \qquad \text{(for sample data)}$$

$$\sigma_x^2 = \sum \frac{(x_i - \mu)^2}{\text{n}} \qquad \text{(for population data)}$$

Where S_x^2 = sample variance of variable X; σ_x^2 is population variance x_i is the value of X variable for ith case; $\bar{x}$ is sample mean; μ is population mean; n is the sample size; and N is population size.

The above formula is used only each score has a frequency of 1 and data are ungrouped. If some scores occur more or less frequently than others, a compact ungrouped frequency table may be constructed in which the entries in the column of frequencies are not all the same and they are not all l's. In such a case the formula for variance is written as:

$$S_x^2 = \sum f_i \frac{(x_i - \bar{x})^2}{\mathrm{n} - 1} \qquad \text{where } \mathrm{f_i} \text{ is the frequency of } \mathrm{x_i}$$

$$\sigma_x^2 = \sum f_i \frac{(x_i - \mu)^2}{\mathrm{N}} \qquad \text{where } \mathrm{f_i} \text{ is the frequency of } \mathrm{x_i}$$

In calculating the sample variance, the reason for dividing by n-1, instead of n, is to get an unbiased estimate of known population variance. The variability of a sample of scores tends to be less than the variability of the population from which the score are taken. In order to use the sample variance as an unbiased estimate of population variance, a correction factor (n-1) is used in the denominator of the formula for the variance of a sample. In other words, sample variance almost always underestimates its corresponding population variance and dividing by (n-1), instead of n, tries to compensate for this underestimate. For larger sample sizes (n>100), it makes little difference whether one divides by n or n-1. Significant error can occur if sample is small (n<25). In situations where the interest is merely in describing the variability in the data at hand, only n should be used as a divisor. As a general rule, one can almost use n-1 for sample data.

As a descriptive statistic for variability, the variance changes in value as a function of the amount of variability in the data. When all scores are identical the value of variance will be zero. As scores become more dispersed around the mean the value of variance increases. Variance is based on squared deviations and therefore it is always than or equal to zero.

2.3.3.2. Coefficient of Variation (CV)

It is sometimes desirable to compare several groups with respect to their relative homogeneity in instances where the groups have very different means. Therefore it might be somewhat misleading to compare the absolute magnitudes of the standard deviations. One might expect that with a very large mean one would find a fairly large standard deviation. One might therefore be primarily interested in the size of the standard deviation relative to that of the mean. This suggests that we can obtain a measure of the relative variability by dividing the standard deviation by the mean. The result has been termed the coefficient of variation, denoted by CV. Thus

$$CV = S/\bar{X}$$

Where s is the SD and $\bar{X}$ is the mean.

The coefficient of variation being a ratio requires that one have a ratio level of measurement and not merely interval measurement. You can also realize that one should always report the mean as well as the SD for the data.

To illustrate the advantages of CV over the SD, suppose a social psychologist is attempting to show that for all practical purpose two groups are equally homogeneous with respect to age. In one group the mean age is 26 with an SD of 3. In the other one, the mean age is 38 with an SD of 5. The coefficients of variation for the two groups are:

$$CV_1=3/26=0.115, \qquad CV_2= 5/38=0.132$$

The different between the two coefficients is smaller than the difference between the two SDs. In view of the fact that exact age usually becomes less important in determining interest, abilities and social status as the average age of group members is increased a comparison of the two coefficient of variation in this instance might very well be much less misleading than if the SDs were used.

As another example suppose one is concerned about the dispersions in traffic flows from one weekday to the next at various times of the day. Dispersions in these flows might be misleading in an absolute sense unless standardized by their means so as to allow for differences in the average volumes of traffic at different times of the day.

2.3.4. Standard Deviation

Standard deviation: Standard deviation is the most stable index of variability.

In the computations of average deviation, the signs of deviation of the observations from the mean were not considered. In order to avoid this discrepancy, instead of the actual values of the deviations we consider the squares of deviations, and the outcome is known as variance. Further, the square root of this variance is known as standard deviation and designated as SD. Thus, standard deviation is the square root of the mean of the squared deviations of the individual observations from the mean. The standard deviation is denoted Greek letter by σ.

Although variance is a very useful measure of variability, its value as a descriptive statistic is limited somewhat by the difficulty most people have in thinking about squared deviations. For instance, if you were calculating the variability for income scores (measured in Rs.), the variance will be expressed in squared units (Rs. Rs or Rs^2) and you might obtain a value of variance say 16 Rs. Rs. In the process of squaring the units also get squared. This is what is done when area is reported and calculated square feet square inches etc.

Computing the square root of the variance expresses this variability in terms of the original score values such as Rs. 4, which is easier to interpret and comprehend. This square root of the variance is called the standard deviation (SD), represented by s. The SD is approximately equal to the mean deviation (MD) of scores around the mean, since variance is the mean squared deviation (MSD), standard deviation is root mean squared deviation (RMSD). Because the standard deviation is more readily interpretable than variance, it is used more often to describe data variability.

That is Standard deviation = Square root of variance or s= $\sqrt{s^2}$

$$s = \sqrt{s^2} = \sqrt{\left[\sum \frac{(x_i - \bar{x})^2}{\text{n} - 1}\ \right]} = \sqrt{\sum f_i \frac{(x_i - \bar{x})^2}{\text{n} - 1}}$$

Similarly, population standard deviation $\sigma = \sqrt{\sigma^2}$

2.3.4.1. Properties of Standard Deviation

(i) Standard deviation gives a measure of dispersion relative to the mean.

(ii) Standard deviation is sensitive to each of the scores in the distribution.

(iii) Like the mean, standard deviation is stable with regard to sampling fluctuations. This property is one of the main reasons why the standard deviation is used so much often than other measures of dispersion.

Properties of SD

If all the score have an identical value in a sample, the SD will be 0 (zero).

In different samples drawn from the same population, SDs differ very less as compared to the other measures of dispersion.

For a symmetrical or normal distribution, the following relationship is true:

Mean ±1 SD covers 68.26 % cases

Mean ± 2 SD covers 95.45 % cases

Mean ± 3 SD covers 99.73 % cases

Merits: It is based on all observations. It is amenable to further mathematical treatments. Of all measures of dispersion, standard deviation is least affected by fluctuation of sampling.

2.3.4.2. Steps in Computing the Standard Deviation

(i) Make a frequency distribution table, if not already made, containing two columns, namely, the columns for score values and their frequencies.

(ii) Calculate the mean score, if not already given:

(iii) $\bar{x} = \frac{\sum f_i x_i}{n} \quad where\ n = \sum f_i$

(iv) Subtract the mean $\bar{x}$ form each of the scores x_i to calculate deviations. $x_i - \bar{x}$. Write these deviations in a separate column, say column 3. Sum all these deviations and see if the sum is zero (excepting the rounding errors.)

(v) Square each deviation obtained in step (iii) and writes the squared amounts in a separate column say column 4.

(vi) Sum all entries in col-4 to obtain a quantity $\sum f_i(x_i - \bar{x}_i)^2$

(vii) Take the square root of variance in step (v) to obtain the standard deviation.

Example 2.5

"How accurate are eyewitness reports of accidents?" Social scientists have studied this question in detail. In one experiment, subject viewed a film of an accident in which a car ran a stop sign and hit a parked car. The speed of the car was 31 miles per hour. After viewing the film, subjects were asked to estimate the speed of the car was 31 miles per hour. After viewing the film, subjects were asked to estimate the speed of the car. Ten subjects gave the following estimates:

15, 40, 32, 18, 35, 20, 37, 35, 28, 40

Calculate the mean and standard deviation for these data. How accurate were the estimates considering the mean score across all subjects? How does the SD help in interpret the mean?

Solution:

To calculate SD, it is useful to make the following table

x_i	f_i	f_ix_i	x_i^2	$f_ix_i^2$
15	1	15	225	225
40	2	80	1600	3200
32	1	32	1024	1024
18	1	18	324	324
35	2	70	1225	2500
20	1	20	400	400
37	1	37	1369	1369
28	1	28	784	784

Total	10	300		9826

Mean

$$\bar{X} = \frac{\sum f_i x_i}{n} = \frac{300}{10} = 30$$

Sample variance

$$S_x^2 = \left[n\left(\sum f_i x_i^2\right) - \left(\sum f_i x_i\right)^2\right] / (n)(n-1)$$

$$= \frac{[10(9826) - (300)^2]}{(10)(10-1)} = (98260 - 9000)/(10)(9)$$

$$= \frac{8260}{90} = 91.78$$

Standard deviations = $\sqrt{S^2} = \sqrt{(91.78)} = 9.58$

Both variance and standard deviation are based on two important properties of the mean: (i) Sum of the differences of scores from the mean in a distribution equals zero. It is due to this property that the deviations from the mean are squared. (ii) The sum of the squared differences of each value in a distribution from the mean of the distribution yields a minimum value, $\sum f_i (x_i - \bar{x}_i)^2$ = minimum.

There are two other important characteristics of frequency distribution that provide useful information about its nature. They are known as Skewness and Kurtosis

2.4. Skewness

Skewness is the degree of asymmetry of the distribution. In some frequency distributions scores are more concentrated at one end of the scale. Such a distribution is called a skewed distribution. Thus, Skewness refers to the extent to which a distribution of data points is concentrated at one end or the other. Skewness and variability are usually related, the more the Skewness the greater the variability. Skewness has both, direction as well as magnitude. In actual practice, frequency distributions are rarely symmetrical; rather they show varying degree of asymmetry or Skewness. In perfectly symmetrical distribution, the mean, median and mode coincide, whereas this is not the case in a distribution that is asymmetrical or skewed.

By skewness of a frequency distribution we mean the degree of its departure from symmetry. The frequency distribution of a discrete variable x is called symmetrical about the value x_o if the frequency of x_o-h is the same as the frequency x_o+h of whatever h may be.

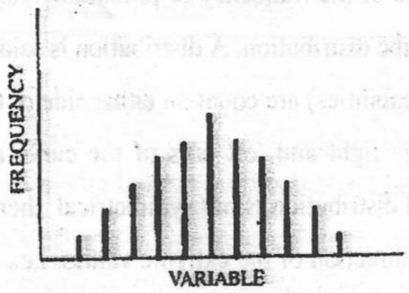


Fig. 2.1a A symmetrical distribution (discrete variable).

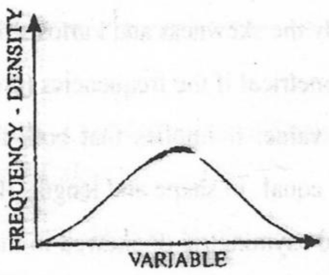


Fig. 2.1b A symmetrical distribution (continuous variable).

In the case of a continuous variable, the term ‘symmetry’ should be used in relation to its frequency curve. The frequency curve of a continuous variable is said to be symmetrical about x_o if the frequency density at x_o-h is the same as the frequency-density x_o+h at whatever h may be. Figure 2.1a and 2.1b show two symmetrical distributions.

A distribution which is not symmetrical is called asymmetrical or skew. This skewness is said to be positive if the longer tail of the distribution is towards the higher values of the variable (Fig. 2.2a), negative if the longer tails is towards the lower values of the variable (Fig. 2.2b)

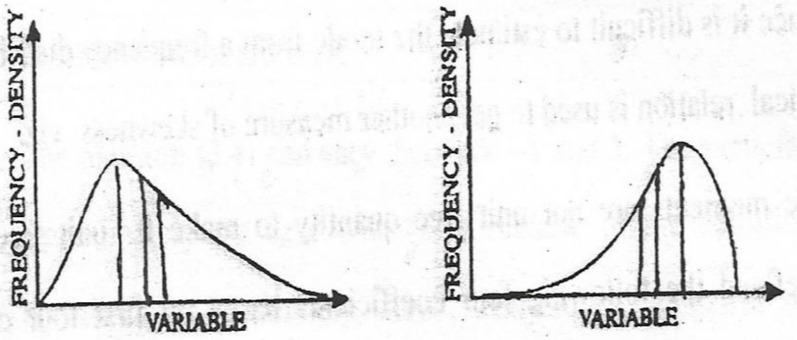


Fig. 2.2a A positively skew distribution.

Mode < Median < Mean

Fig. 2.2b A negatively skew distribution.

Mode < Median < Mean

An important point to be noted in this connection is that all odd-order central moments are zero for a symmetrical distribution, positive for a positively skew distribution and negative for a negatively skew distribution. Any such moment may, therefore, be considered a measure of the skewness of a distribution except, of course, m_1 which a necessarily zero for any distribution-symmetrical or otherwise. The simplest of these measures is m_3.

2.5. Kurtosis

The term ‘kurtosis’ refers to the ‘peakedness’ or flatness of a frequency distribution curve when compared with normal distribution curve. The Kurtosis of a distribution is the curvedness or peakedness of the graph.

Another method of describing a frequency distribution is to specify its degree of peakedness or kurtosis. Two distributions may have the same mean and the same standard deviation and may be equally skew, but one of them may be more peaked than the other.

2.5.1. Measure of Kurtosis β_2 and γ_2

This feature of the frequency distribution is measured by

$$\beta_2 = \frac{m_4}{m_2^2}, \quad and$$

$$\gamma_2 = \beta_2 - 3 \qquad \dots\dots\dots..(2.7)$$

Obviously, it is a pure number. For a normal distribution, β_2 = 3 and γ_2 =0. A positive value of γ_2 indicates that the distribution has high concentration of value near the central tendency and has high tails, in comparison with a normal distribution with the same standard deviation. In the same way, a negative value γ_2 means that the distribution has low.

β_2 = 3 implies that γ_2 = 0, the kurtosis is same as that of normal curve. The curve is *Mesokurtic*.

$\beta_2 > 3 \rightarrow \gamma_2 > 0$ the Kurtosis is said to be positive and curve called the *Leptokurtic*.

$\beta_2 < 3 \rightarrow \gamma_2 < 0$ the Kurtosis is said to be negative and curve called the *Platykurtic.*

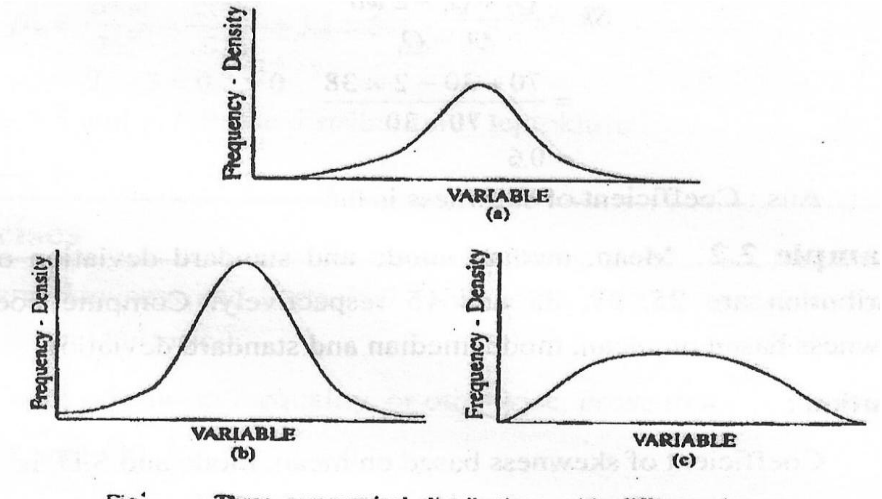


Fig. Three symmetrical distributions with different degrees of kurtosis : (a) mesokurtic, (b) leptokurtic, (c) platykurtic.

Concentration of values in the neighborhood of the central tendency and low tails, compared to a normal distribution with the same standard deviation. A normal curve is said to be mesokurtic (i.e. having medium kurtosis). A distribution with positive γ_2 is called leptokurtic, and one with negative γ_2 is known as platykurtic. The quantities $\beta_1 and\ \beta_2$ themselves are sometimes used as measures of skewness and kurtosis, respectively.

That the fourth central moment (m_4) may be used in measuring kurtosis becomes obvious from the fact that higher the kurtosis, the higher will be the effect of the large deviations (from the mean) in the tails when raised to the fourth power. Division of m_4 by s^4 makes the measure a pure number.

Actually, however, $\beta_1(or\, \gamma_2)$ will be appropriate as a measure of kurtosis or peaked-ness only if we confine our attention to the class of the usual bell-shaped (or unimodal) distributions. Otherwise, it may only serve to distinguish a unimodal distribution from a bimodal.

USE OF DESCRIPTIVE STATISTICS

Descriptive statistics are used to describe the basic features of the data in a study.

They provide simple summaries about the sample and the measures.

Together with simple graphical analysis, they form the basis of virtually every quantitative analysis of data.

With descriptive statistics one is simply describing what is in the data or what the data shows.

- Descriptive Statistics are used to present quantitative descriptions in a manageable form.
- In any analytical study we may have lots of measures. Or we may measure a large number of people on any measure.
- Descriptive statistics help us to simplify large amounts of data in a sensible way.
- Each descriptive statistic reduces lots of data into a simple summary.
- Every time you try to describe a large set of observations with a single indicator, you run the risk of distorting the original data or losing important detail.
- Even given these limitations, descriptive statistics provide a powerful summary that may enable you to make comparisons across people or other units.

2.6. Summary

Various measures of central tendency have been defined in this unit. There is found a tendency in the data to cluster around a central value. This value is known as measure of central tendency. These are mean, median and mode. Mean is obtained by dividing the sum of observations by number of observations. Median is that variants value which divides the given data or frequency distribution in two equal halves. Various measures of dispersion have been defined and formulas for their calculation are given in this unit. Once data have been represented by a measure of central tendency, are may like to know the scatter of the given data around this measure of central tendency. The various measures of dispersion are range, quartile deviation mean deviation, standard deviation and variance. Coefficient of variation for consistency of data or frequency distribution is defined as the ratio of standard deviation to arithmetic mean.

2.7. Terminal questions

Q.1. What do you mean by central tendency? Write about mean, mean and mode.

Answer: --
--

Q.2. Discuss the calculation of median of ungrouped data.

Answer: --
--

Q.3. What is the variance? Discuss the computation of variance.

Answer: --
--

Q.4. What are the measures of dispersion? Discuss the range and its advantages.

Answer: --
--

Q.5. What is the standard deviation? How to calculate standard deviation of given data.

Answer: --
--

Q.6. What are skewness and kurtosis? Give some suitable measures for skewness and kurtosis.

Answer: ---

2.8. Further suggested readings

1 Kenney, J.F and Keeping, E.S.: Mathematics of Statistics, Part I (Ch. 7) Van Nostrand, 1954 and Affiliated East- West Press.

2 Mills, F.C.: Statistical Methods (Ch.5) H. Holt, 1955.

3 Yule G.U. and Kerdall, M.G.: Introduction to the Theory of Statistics (Ch.6) Charles Griffin, 1953.

4 Fundamentals of Statistics volume I by Goon, Gupta Dasgupta.

5 Nagar, A. L., and Das, R. K. (1983). *Basic Statistics*. Oxford University Press, Delhi.

Unit-3: Correlation and Regression Analysis

Contents

3.1. Introduction

Suppose we are given the data on the scores obtained by 30 students in Statistics. In the previous units that we have studied so far, we know how to form the frequency distribution of the data. We have also seen how to compute the average score and the standard deviation. Now suppose the scores obtained by these students in environmental sciences are also given to him. Apart from finding the average score in environmental sciences, we may also like to know whether there is any relationship between a student's score in Statistics and higher score in environmental sciences. It is quite reasonable to think that a student good at Statistics would also be good at environmental sciences. But is this fact borne out by the data? In this unit, we are going to discuss some methods which will enable us to answer this question. We'll start by discussing the tabular and diagrammatic representation of bivariate data, i.e., of data pertaining to two variables.

After this, we'll talk about the nature and the degree of relationship between the observations on two variables. We would also like to see if we can build up an equation on the basis of the given data, which can help us in predicting one of the variables when the other is given.

Prediction is a goal common to all sciences, including social and behavior sciences. Prediction is based on relationship between or among variables. It is the fact that crime rate and cost of living are correlated that enables us to predict crime rate from cost of living, or vice versa. Thus, the correlation between variables sets the stage for predicting one variable from another. The stronger the correlation between two variables, the more accurately can one be predicated from the other, and weaker the correlation, less accurate is the predictions. A powerful statistical tool for translating correlations into predictions is referred to as bivariate regression analysis or simple linear regression. Regression analysis is a method of analyzing the variability of a dependent variable by resorting to information available on one or more independent variables. An answer is sought to the questions: What are the expected changes in Y as a result of changes in X? Bivariate regression refers to the case in which only one X and one Y are being analyzed at a time. It is through bivariate regression analysis that the correlation between X and Y is used in predicting one variable from the other variable, called the criterion variable or dependent variable or Y variable, from the other variable called the predictor variable or independent variable or X variable. One way of facilitating predictions is to obtain a simple linear equation that fits, or represents the available data. This equation can then be used to study how a change in X variable relates to a change in Y variable. In this chapter we discuss simple linear regression that helps us find such a prediction equation.

Objectives

After going through this unit you shall be able to

- After reading this unit, we will be able to :
- draw a scatter diagram corresponding to the given bivariate frequency distribution
- explain the meaning of "correlation" and "regression"
- Understand regression and obtain regression lines

- fit a regression line to the given data
- compute the correlation coefficient for grouped and ungrouped data
- Derive some relationship between the correlation and regression coefficient.
- Use regression coefficients and their properties

3.2. Scatter Diagram

The simplest mode of diagrammatic representation of bivariate data is the use of *scatter diagram* (or dot diagram). Taking two perpendicular axes of co-ordinates, one for x and the other for y, each pair of values $\{(x_i, y_j), i, j=1,2,3,\ldots.n)\}$ is plotted as a point on graph paper or xy- plane. The whole set of each (x_i, y_j) is represented as a point taken together constitutes the scatter diagram or 'dot' diagram. After obtaining the raw data on two quantitative variables, these data must first be arranged and paired as $(x_i, y_j, i, j=1,2,\ldots.N)$. It is not efficient to make a contingency table because both variables can potentially take a large number of values. Instead these data are arranged in the form of a graph known as Scatter Diagram. A scatter diagram graphically summarizes the data on two quantitative variables by showing the joint distribution of the values on two variables. Each point in a scatter plot represents individual's scores on the two variables represented by the two axes of the graph. A given case's point is located at the intersection of that case's values on each variable. To construct a scatter plot: (i) Draw an X-axis or horizontal axis and label it with the name and values of the independent variable. (ii) Draw a Y-axis or vertical axis and label it with the name and values of the dependent variable. To keep the graph from appearing either too flat or too steep, keep the height of the vertical axis equal to two thirds the length of horizontal axis. (iii) Plot the pairs of points (x_i, y_j) as dots between the two axes. For larger data sets, use SPSS to construct scatter plots.

Example 3.1

A researcher is interested in studying the relationship between level of education (measured in years) and income (measured in thousands of rupees). Let level of education be the independent variable denoted by X and income be the dependent variable dented by Y. For a sample of 15 people, the data are given below:

Case # :	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X:	7	12	8	12	14	9	18	14	8	12	17	10	16	10	13
Y:	9	16	14	12	11	16	19	13	13	14	14	16	15	10	18

For these data, make a scattered diagram.

Solution:

A scattered diagram of data on income and education is shown by Fig. 1.1 on next page

3.2.1. Reading and Interpreting a Scatter Diagram.

A scatter diagram provides several important pieces of information about the relationship between X and Y. In fact scattered diagram can often provide information about the relationship that is missed from a casual inspection of the statistical (numerical data) usually computed in connection with a correlation analysis. For this reason a scattered diagram should always be constructed as the first step in correlation analysis. Using a scatter diagram one can address the following questions:

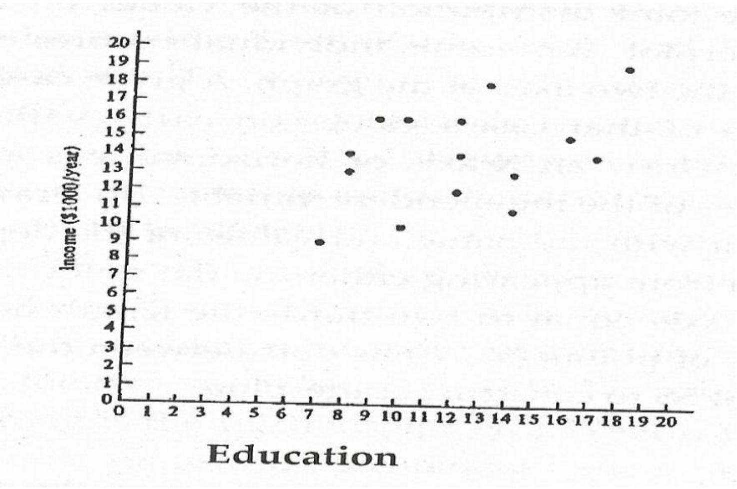


Figure 3.1 An Escambled Scatter Diagram

a) Is the relationship linear? If an examination of the scatter diagram suggests a football of egg-shaped pattern, two variables are likely to be related in a linear manner.

b) Does there exist a relationship? If the football is tilted, two variables are expected to be related. If the football is either standing vertically or hanging horizontally, if the points are parallel to either of the axes, then there are no relationships between the two variables since the change in one variable will not imply a change in the other variable.

c) How strong is the relationship? If all the points fall on a straight line (that is, there is no scatter of points), the relationship is perfect. And if the pattern in the scatter is either circular or random, or nonlinear, then there is no linear relationship between X and Y. The thinner the football, stronger the relationship and thicker it be, weaker the relationship. Scatter diagram reveals the extent to which variables co-vary, and the amount of variability found in each variable considered singly.

d) What is the nature of the relationship? If the football is tilted from left to right, it indicates a positive relationship and if the football is tilted from right to left it is indicative of a negative relationship.

e) ***Are there any outliers in the data?*** An outlier is a case that is radically different from the majority of other cases in terms of its combined or joint values on X and Y. Although an outlier may show scores on X and Y that are each well within the normal ranges for those variables the pair of scores may make the case quite deviant. In a scatter diagram the outliers will appear conspicuously removed from the rest of the points. Identifying outliers is important for several reasons. First, an outlier may represent a case whose scores on X and or Y have been recorded incorrectly. Second outliers may show deviant combined scores on X and Y because they did not understand (or deliberately disobeyed) instructions given during data collection. Third, outliers represent cases to which statements about the relationship between X and Y draw from the majority of cases do not apply. Fourth, outliers exert a disproportionate effect on computed correlations.

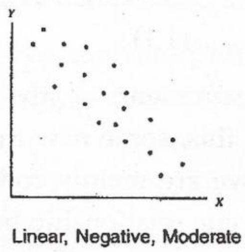


Fig. 1.2a

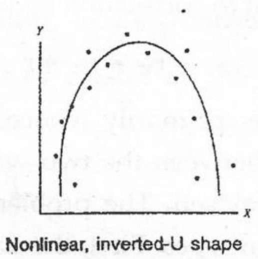


Fig. 1.2 b

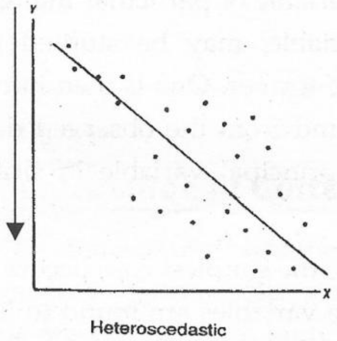


Fig. 1.2c

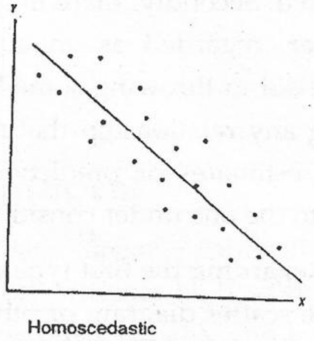


Fig.. 1.2d

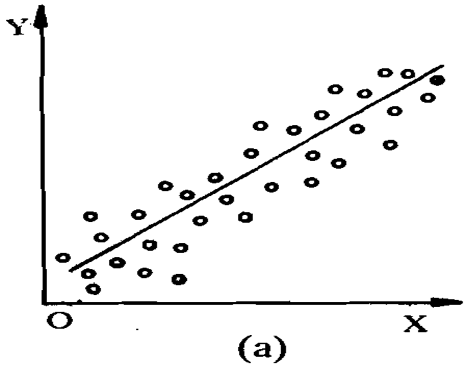
3.3. Karl Pearson's Coefficient of Correlation

An examination of the scatter diagram in example 3.1 suggests a linear, positive and moderately strong relationship between education and income. But examination of a scatter diagram is only visual, approximate and subjective way of measuring the relationship even the change of scales of axes has an effect on the slope of the curve. We would manner. That is we would like to measure the relationship is more accurate, concrete and objective manner. That is we would like to quantify the scatter diagram. A measure used to qualify the degree of linear relationship between two variables is called Pearson product moment correlation coefficient, denoted by r. Just be observed in graphed frequency distributions, Pearson r gives a more precise indication of the linear relationship between X and Y that is available from inspection a scatter diagram. Pearson r is a symmetric statistic that is its value and nature does not depend upon whether X is independent and Y is dependent or vice versa. The correlation between X and Y is equal to correlation between Y and X. That is

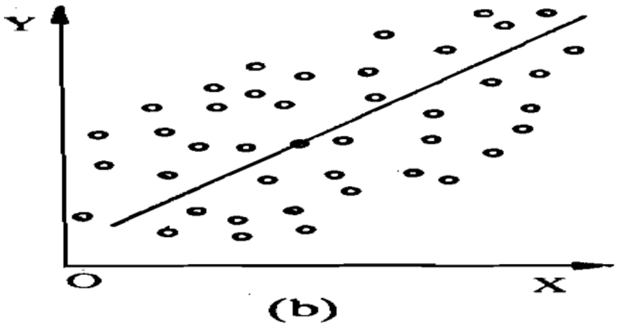
$$r_{xy} = r_{yx} = r \text{………………………..(1.1)}$$

Here we are primarily concerned with measurement of the linear relationship between the two variables and for this some new methods have to be devised. The problems with which we are mainly concerned may be of two types. First, the data may reveal some relationship between the two variables and we may want to measure the extent to which they are related. Secondly, there may be one variable, may be studied for its possible aid in throwing some light on the former. One is then interested in using any relationship that may be found from the observed data for making estimates or predictions of the principal variable in situations similar to the one under consideration.

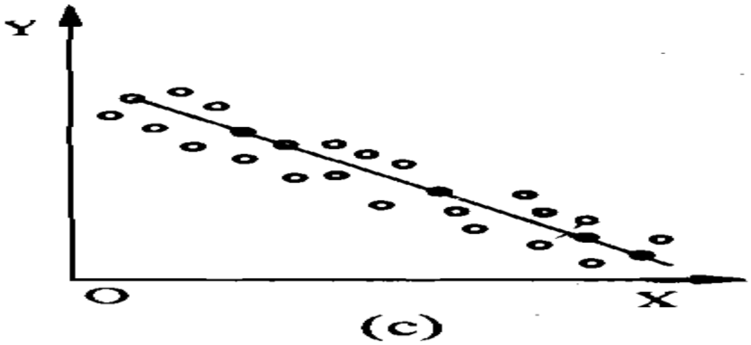
Regarding the first type of problem the simplest case occurs when from the scatter diagram or otherwise, the variables are found to linearly related, at least approximately. If it is found that as one variable increases the other also increases, in general or on the average there will be said to be *positive correlation* between them. This will be the case, for example, when the data relate to the height and weight of people or the score in mathematics and the score in statistics of students in a college. On the other hand, as one variable increases, the other may decrease on the average; we then say that there is *negative correlation* between them. There may still be a third situation where as one variable increase, the other remains constant on the average. This is the case of zero or no correlation and the two variables are then said to be uncorrelated. A near zero correlation is expected when we have data on height and IQ of students in HS institution. Following scatter diagrams shows the nature of correlation between two variables x and y.



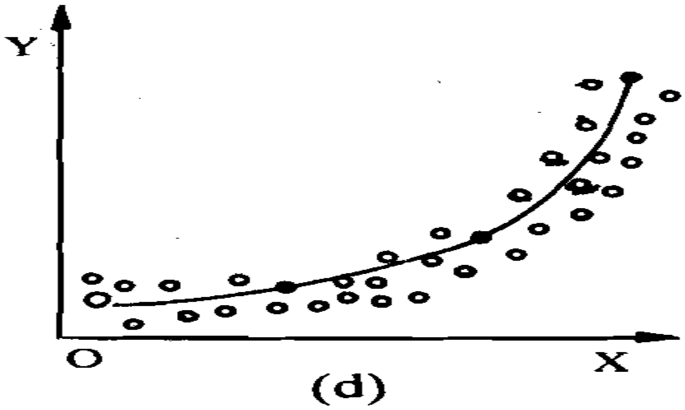
Positive linear relationship



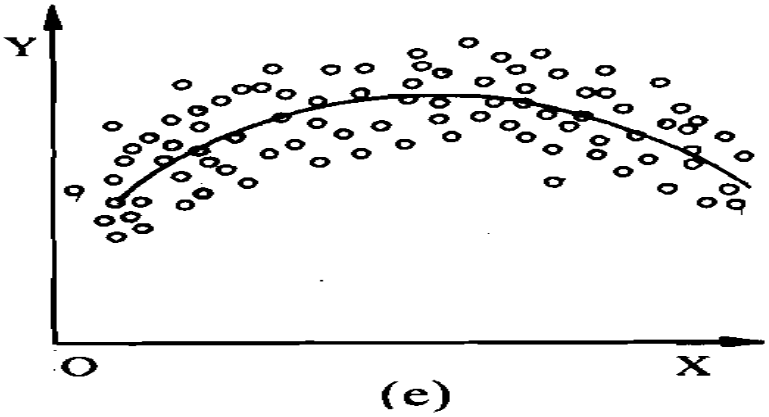
Positive linear relationship



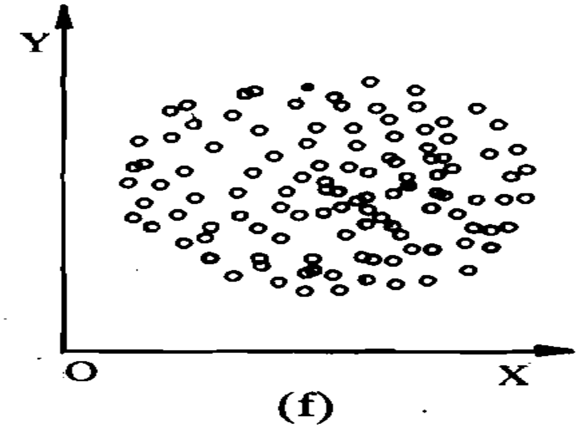
Negative linear relationship



Curvilinear relationships



Curvilinear relationships



No relationship (Zero correlation)

In figure (a), (b) and (c), you can see that all the points are quite close to a straight line path. In (a) and (b), this straight line has positive slope, while the one in (c) has negative slope. We say that the data in (a) and (b) show a **positive linear relationship.** Of course, the points in (b) are more scattered than those in (a). The scatter diagram in (c) exhibits a **negative linear relationship.** From (d) and (e), we can say that the data indicate curvilinear relationships. But, from the scatter diagram in (f), we cannot think of any relationship existing between the variables in the data.

We can interpret the way scatter of cluster as the properties of relationship between the two variables.

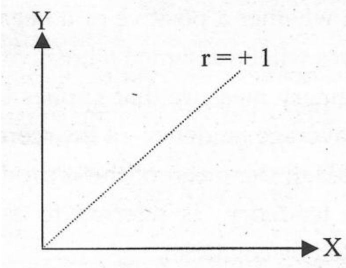


Fig. 1.4a

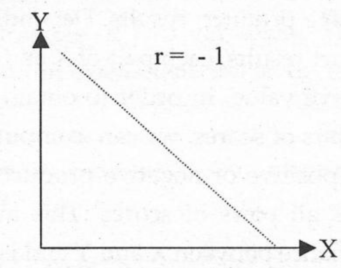


Fig. 1.4b

8.4.1 Calculation of Correlation Coefficient

In determining whether a systematic linear relationship exists between two variables, one seeks to find out whether high scores on one variable are paired with high scores on the other variable and low scores are paired with low scores or whether high scores on one variable are paired with low scores on the other variable and low scores are paired with high scores or neither. In order to label a particular X scores as either high or low relative to the other X scores, we compare it to the mean $\bar{X}$ of the set of all the scores. Similarly, we compare each particular Y scores to other Y's with the mean $\bar{Y}$ of all Y scores. If the X score is high, then it will be higher than $\bar{X}$ and the difference X- $\bar{X}$ will be positive. If the X score is low, than X and the difference X- $\bar{X}$ will be negative. Likewise for the Y scores. In this way we calculate all the deviations $X_i - \bar{X}$ and $Y_i - \bar{Y}$.

To determine whether a high score on X is paired with a high score on Y, and vice versa form the product $(X_i - \bar{X})(Y_i - \bar{Y})$ of the two deviations terms for each pair of (X_i, Y_i) values $(X_i - \bar{X})\ and\ (Y_i - \bar{Y})$ i=1,2,3,……..N. If both X and Y scores are high with regard to their respective distributions, then both terms $(X_i - \bar{X})\ and\ (Y_i - \bar{Y})$ in the product will be positive and the product $(X_i - \bar{X})(Y_i - \bar{Y})$ itself will be positive. Like wise, if both the X an Y scores are low with regard to their respective distributions, then both terms $(X_i - \bar{X})\ and\ (Y_i - \bar{Y})$ in the product will be negative but the product $(X_i - \bar{X})(Y_i - \bar{Y})$ will again be positive. Thus a positive product results when either both terms are positive or both negative. When one (not both) of the differences is negative, signifying either a high X score paired with a low Y score or a low X score paired with high Y score negative product results. Depending on whether a positive or a negative product results each pair of X and Y scores will be assigned a positive or a negative value. In order to obtain a summary measure that applies to all the pairs of scores, we can compute the average tendency of the scores to have positive or negative products by taking the mean of these products across all pairs of scores. This average tendency is referred to as the covariance between X and Y is denoted by cov (X,Y);

$$\mu_{12} = cov\ (X, Y)$$

$$= \frac{(X_i - \bar{X})(Y_i - \bar{Y}) + (X_2 - \bar{X})(Y_2 - \bar{Y}) + \cdots + (X_n - \bar{X})(Y_n - \bar{Y})}{N}$$

$$= \frac{1}{N}\sum_{i=1}^{N}[(X_i - \bar{X})(Y_i - \bar{Y})]$$

$$= \frac{1}{N}\sum x_i y_i \ \ldots \ldots \ldots \ldots \ldots (1.2)$$

Where $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$ for i=1, 2, 3,……N

The covariance measures how X and Y vary together. If the covariance between X and Y is positive, a positive linear relationship exists between X and Y. Cov (X,Y)>0 implies that positive x_iy_i's dominates negative x_iy_i's. If the covariance between X and Y is negative, a negative linear relationship exists between X and Y. If the covariance between X and y is 0, no linear relationship exists between X and Y and the sum of positive x_iy_i's equals in magnitude to sum of negative x_iy_i's help us to interpret the strength of the relationship between the two variables. However, because there are not bounds on the magnitude of the covariance term, it is difficult to know what covariance value constitutes a strong relationship and what value constitutes a weak relationship.

The manner in which a score is evaluated as either high or low depends only on the score and the mean. Therefore if the standard deviations of X and Y differ, the same difference value in raw points on (X_i-X) and (Y_i-Y) would not in general mean the same real distance above or below the respective means. And since we are comparing two distributions of scores pairwise, it is important to standardize the differences so that it means the same things in both distributions. This standardizing can be done in various ways. One way is to divide the total covariance by the product of total variance of two variables. What we obtain is called the Pearson's correlation coefficient r_{xy} is or product moment correlation coefficient. That is,

$$r_{xy} = \frac{cov(X,Y)}{\sqrt{var(X)}\sqrt{var\ (Y)}} = \frac{\mu_{12}}{\mu_{11}\mu_{22}} \quad \ldots\ldots\ldots\ldots\ldots\ldots.(1.3)$$

$$r_{xy} = \frac{\frac{1}{N}\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left[\frac{1}{N}\sum(X_i - \bar{X})^2\right]\left[\frac{1}{N}\sum(Y_i - \bar{Y})^2\right]}}$$

$$r_{xy} = \frac{\sum[(X_i - \bar{X})(Y_i - \bar{Y})]}{\sqrt{(X_i - \bar{X})^2}\ \sqrt{(Y_i - \bar{Y})^2}} \quad \ldots\ldots..(1.4)$$

Again

$$N\ cov\ (X,Y) = \sum(X_i - \bar{X})(Y_i - \bar{Y}) = \sum X_iY_i - N\overline{XY}$$

$$= \sum X_iY_i - \left(\sum X_i\right)\left(\sum Y_i\right)/N \ldots\ldots\ldots.(1.5)$$

$$N\ var\ (X) = \sum(X_i - \bar{X})^2 = \sum X_i^2 - N\bar{X}^2$$

$$= \sum X_i^2 - \frac{(\sum_i X_i)^2}{N} \qquad (1.6)$$

And similarly,

$$N\ var\ (Y) = \sum Y_i^2 - \left(\sum Y_i\right)^2/N \ldots\ldots\ldots\ldots.(1.7)$$

Hence r may be expressed in the alternative forms;

$$r_{xy} = \frac{\sum[x_iy_i\ /N] - \overline{XY}}{\{[\sum_i X_i^2/N] = \bar{X}^2\}^{1/2}\ \{[\sum_i Y_i^2/N] = \bar{Y}^2\}^{1/2}} \qquad \ldots\ldots..(1.8a)$$

$$= \frac{N\sum_i X_iY_i - (\sum_i X_i)(\sum_i Y_i)}{\{N\sum_i X_i^2 - (\sum_i X_i)^2\}^{1/2}\ \{N\sum_i Y_i^2 - (\sum_i Y_i)^2\}^{1/2}} \qquad \ldots\ldots..(1.8)$$

The form (1.4) will be found to be the most convenient for remembering the def; Either Form (1.8) or more convenient for computation work. If any one of $\sum X_iY_i, \sum X_i^2\ or\ \sum_i Y_i^2$ is quite large in magnitude, then we use (1.8a) computing r from raw data.

These formulae require the computation of only five different sums:

$$\sum X_i, \sum Y_i, \sum X_i Y_i, \sum X_i^2, \sum Y_i^2$$

Example 3.2

Using the data in Example 1.1 calculate r.

Solution:

The necessary information is presented in the following table. The first two columns contain the given X and Y score.

Table 1.1

<table>
<tr><th colspan="2">X_i</th><th>Y_i</th><th>X_iY_i</th><th>X_i^2</th><th>Y_i^2</th></tr>
<tr><td colspan="2">7</td><td>9</td><td>63</td><td>49</td><td>81</td></tr>
<tr><td colspan="2">12</td><td>16</td><td>192</td><td>144</td><td>256</td></tr>
<tr><td colspan="2">8</td><td>14</td><td>112</td><td>64</td><td>196</td></tr>
<tr><td colspan="2">12</td><td>12</td><td>144</td><td>144</td><td>144</td></tr>
<tr><td colspan="2">14</td><td>11</td><td>154</td><td>196</td><td>121</td></tr>
<tr><td colspan="2">9</td><td>16</td><td>144</td><td>81</td><td>256</td></tr>
<tr><td colspan="2">18</td><td>19</td><td>342</td><td>324</td><td>361</td></tr>
<tr><td colspan="2">14</td><td>13</td><td>182</td><td>196</td><td>169</td></tr>
<tr><td colspan="2">8</td><td>13</td><td>104</td><td>64</td><td>169</td></tr>
<tr><td colspan="2">12</td><td>14</td><td>168</td><td>144</td><td>196</td></tr>
<tr><td colspan="2">17</td><td>14</td><td>238</td><td>289</td><td>196</td></tr>
<tr><td colspan="2">10</td><td>16</td><td>160</td><td>100</td><td>256</td></tr>
<tr><td colspan="2">16</td><td>15</td><td>240</td><td>256</td><td>225</td></tr>
<tr><td colspan="2">10</td><td>10</td><td>100</td><td>100</td><td>100</td></tr>
<tr><td colspan="2">13</td><td>18</td><td>234</td><td>169</td><td>324</td></tr>
<tr><td>Total</td><td>180</td><td>210</td><td>2577</td><td>2320</td><td>3050</td></tr>
</table>

Putting N=15, $\sum X$=180, $\sum Y$=210, $\sum X^2$=2320, $\sum Y^2$= 3050, $\sum X\,Y$=2577 in equation (1.8) and (1.1), we get

$$r_{xy} = \frac{N(\sum X_iY_i) - (\sum X_i)(\sum Y_i)}{\sqrt{[\{N(\sum X_i^2) - (\sum X_i)^2\}\,\{(N\sum Y_i^2) - (\sum Y_i)^2\}]}}$$

$$r = r_{xy} = r_{yx} = \frac{15(2577) - (180)(210)}{\sqrt{[(15 \times 2577 - 180^2)(15 \times 2320 - 210^2)]}} = 0.43$$

as a measure of degree of relationship between linearly related variable X & Y.

3.4. Properties of Correlation Coefficient

1. Correlation coefficient is a pure number having no unit.
2. The limits of correlation coefficient r lies between -1 to +1, that is -1 ≤ r ≥1
3. Correlation coefficient is independent of change of Origin and scale.
4. It is not affected by linear transformation of variables, that is, if u=X-A,v=Y-B, Then r_{xy}= r_{uv} and is independent of A and B.
5. If correlation coefficient between X and Y be r and regression coefficient b_{yx} and b_{xy}, then $r = \sqrt{b_{yx}b_{xy}}$
6. If two variables x and y are independent then r =0, but converse is nottrue.

3.5. Spearman's Coefficient

First, let us suppose that there is no tie, i.e., no two individuals are ranked equal in either variable. The ranks x's and y's take values 1, 2, 3,….. n in some order. Hence

$$\sum x_i = \sum y_i = 1 + 2 + \cdots ..n(n+1)/2.$$

and means are

$$\bar{x} = \bar{y} = \frac{n+1}{2} \qquad \ldots \ldots ..(3.1)$$

Sum of squares are given as

$$\sum x_i^2 = \sum y_i^2 = 1^2 + 2^2 + \cdots..n^2 = (n+1)(2n+1)/6.$$

and

$$\sigma_x^2 = V(x) = \left(\frac{1}{n}\sum X^2 - \bar{x}^2\right)$$

$$= \left\{\frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2\right\}$$

$$= \frac{n^2-1}{12} = V(Y) = \sigma_y^2 \qquad \ldots\ldots\ldots..(3.2)$$

Let d$_i$ = x$_i$ – y$_i$,

Since $\bar{x} = \bar{y}$, d$_i$ can be written as d$_i$ = $(x_i - \bar{x}) - (y_i - \bar{y})$

Squaring both sides and summing over i from 1 to n we get

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n}\{(x_i - \bar{x}) - (y_i - \bar{y})\}^2$$

$$= \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2 - 2\sum_{i=1}^{n}(x_i - \bar{x}) - (y_i - \bar{y})$$

Dividing both sides by n sides n we have

$$\frac{1}{n}\sum_{i=1}^{n} d_i^2 = \sigma_x^2 + \sigma_y^2 - 2Cov(x,y) \qquad \ldots\ldots.(3.3)$$

Let ρ be the correlation coefficient between the ranks x and y, then

$$\rho = \frac{Cov\ (x,y)}{\sigma_x\sigma_y}\ or\ \ Cov(x,y) = \rho\sigma_x\sigma_y \qquad \ldots\ldots(3.4)$$

From (3.3) and (3.4)

$$\frac{1}{n}\sum_{i=1}^{n} d_i^2 = \sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y$$

Since $\sigma_x^2 = \sigma_y^2$, we have

$$\frac{1}{n}\sum_{i=1}^{n} d_i^2 = 2\sigma_x^2 - 2\rho\sigma_x^2$$

$$= 2\sigma_x^2\,(1-\rho)$$

$$(1-\rho) = \frac{\sum_{i=1}^{n} d_i^2}{2n\sigma_x^2}$$

or

$$\rho = 1 - \frac{\sum d_i^2}{2n\sigma_x^2}$$

Putting the value of σ_x^2 from (3.2) we get

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \qquad \ldots\ldots\ldots.(3.5)$$

This is Spearman's formula for the rank correlation coefficient. Since the rank correlation coefficient is simple product moment correlation coefficient between two series of ranks, it always lies between -1 to +1.

For perfect agreement $x_i = y_i$ for each i and $d_i = 0$ for all i,

$$So \qquad \sum d_i^2 = 0 \qquad and\ \rho = 1.$$

For perfect disagreement $y_i = n - x_i + 1$ ∀ i

$$d_i = x_i - y_i = x_i - (n - x_i + 1)$$

$$= 2x_i - (n+1)$$

$$\sum d_i^2 = 4\sum_{i=1}^{n}\left(x_i - \frac{n+1}{2}\right)^2 = 4n\sigma_x^2 = \frac{n(n^2-1)}{3}$$

$$\rho = 1 - \frac{6\frac{n(n^2-1)}{3}}{n(n^2-1)} = 1 - 2 = -1$$

Difference between regression and correlate coefficient

Consider the following situations:

i) The advertising manager of a firm collects data about the money spent on advertising and the sales in each year during 1980-90.

ii) A doctor collects data about the extent of cellular damage induced by exposure to differing intensities of radiation.

iii) A social worker collects data about the number of children, the ages of parents at the time of marriage and their educational status.

In each of these cases, data are collected to explore the possible relationship between the variables. The advertising manager wants to know whether there is any relationship between the money spent on advertising and the sale figures. He would also like to know what the relationship is, for it will help him decide how much more money he should spend td reach a particular target of sales. Similarly, the doctor is concerned about the extent of damage caused by exposure to radiation, and would want to have a clear idea before prescribing the dose.

The social worker wants to know what kind of relationship, if any, exists between the number of children born to a couple, and the ages of the parents at the time of their marriage and also their educational status.

In **regression analysis,** we deal with statistical methods which help us in formulating models which describe relationships among variables. These models are eventually used for prediction. The term **simple regression** is used when we are exploring the relationship between two variables (as in the first two situations above). When we are predicting one variable on the basis of information on more than one 'predictor' variables (as in the third situation), we use the term, **multiple regression.**

Principle of Least Square

Let Y and X be the dependent and independent variables respectively and we have a set of values (X_1 , Y_1),(X_2 , Y_2),..., X_n, X_n), i.e. observations are taken from n individuals on X and Y. We are interested in studying the function Y = f(X). If Yi is the estimated value of Y obtained by the function and Y_i is the observed value of Y at xi then we can define residual. The difference between y_i and Y_i i.e. the difference between observed value and estimated value is called error of the estimate or residual forY_i.

Principle of least squares consists in minimizing the sum of squares of the residuals, i.e. according to principle of least squares

$W = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ Should be minimum

Let us consider a curve (function) of the type $Y = a + bX + cX^2 + \cdots \ldots \ldots tX^k \ldots \ldots \ldots \ldots (A)$

Where, Y is dependent variable, X is independent variable and a, b, c,…,t are unknown constants. Suppose we have (X_1, Y_1) (X_2, Y_2),…….,(X_n, Y_n) values of two variables (X, Y) i.e. data for variables X and Y. These variables may be height and weight, sales and profit, rainfall and production of any crop, etc. In all these examples, first variables, i.e. height, sales and rainfall seem to be independent variables, while second variables, i.e. weight, profit and production of crop seem to be dependent variables. With the given values have (X_1, Y_1) (X_2, Y_2),…….,(X_n, Y_n) curve (function) given in equation (A) produces set of n equations.

$$Y_1 = a + bX_1 + cX_1{}^2 + \cdots \ldots \ldots tX_1{}^k$$

$$Y_2 = a + bX_2 + cX_2{}^2 + \cdots \ldots \ldots tX_2{}^k$$

.

. …….

……………………………. (B)

.

$$Y_n = a + bX_n + cX_n{}^2 + \cdots \ldots \ldots tX_n{}^k$$

Our problem is to determine the constants a, b, c,…, t such that it represents the curve of best fit given by equation (A) of degree k.

If n = k+1, i.e. number of equations and number of unknown constants are equal, there is no problem in determining the unknown constants and error can be made absolutely zero. But more often n > k+1 i.e. number of equations is greater than the number of unknown constants and it is impossible to do away with all errors i.e. these equations cannot be solved exactly which satisfy set of equations. Therefore, we try to determine the values of a, b, c,…, t which satisfy set of equations (B) as nearly as possible.

Substituting X_1, X_2,…….,X_n for X in equation (A) we have

$$Y_1 = a + bX_1 + c{X_1}^2 + \cdots \ldots \ldots t{X_1}^k$$

$$Y_2 = a + bX_2 + c{X_2}^2 + \cdots \ldots \ldots t{X_2}^k$$

.

. ...

.... (C)

.

.

$$Y_n = a + bX_n + c{X_n}^2 + \cdots \ldots \ldots t{X_n}^k$$

The Quantities Y_1 , Y_2 ,.......,Y_n are called expected or estimated values of Y_1 , Y_2 ,.......,Y_n (given values of Y) for the given values of X_1 , X_2 ,.......,X_n. Here Y_1 , Y_2 ,.......,Y_n are the observed values of Y.

Let us define a quantity W, the sum of squares of errors i.e.

$W = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

$W = \sum_{i=1}^{n}\left(Y_i - a - bX_i - c{X_i}^2 - \cdots \ldots \ldots t{X_i}^k\right)^2$............................(D)

According to the principle of least squares the constant a, b,…, t are chosen in such a way that the sum of squares of residuals is minimum. According to principle of maxima and minima (theorem of differential calculus), the extreme value (maximum or minimum) of the function W are obtained by

$$\frac{dW}{da} = 0 = \frac{dW}{db} = 0 = \frac{dW}{dc} = 0 \qquad \ldots .. = \frac{dW}{dt} = 0$$

(Provided that the partial derivatives exist)

Let us take

$$\frac{dW}{da} = 0$$

$$\frac{d}{da}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = 0$$

$$\frac{d}{da}\sum_{i=1}^{n}\left(Y_i - a - bX_i - c{X_i}^2 - \cdots \ldots \ldots t{X_i}^k\right)^2 = 0$$

$$2\sum_{i=1}^{n}(Y_i - a - bX_i - c{X_i}^2 - \cdots \ldots \ldots t{X_i}^k) - (1) = 0$$

$\sum_{i=1}^{n} Y_i = na + b\sum X_i + c\sum {X_i}^2 + \cdots \dots \dots + t\sum {X_i}^k$...................................(E)

$$\frac{d}{db}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = 0$$

$$\frac{d}{db}\sum_{i=1}^{n}\left(Y_i - a - bX_i - c{X_i}^2 - \cdots \dots \dots t{X_i}^k\right)^2 = 0$$

$$2\sum_{i=1}^{n}(Y_i - a - bX_i - c{X_i}^2 - \cdots \dots \dots t{X_i}^k) - (X_i)=0$$

$$\sum_{i=1}^{n} Y_i X_i = a\sum X_i + b\sum {X_i}^2 + c\sum {X_i}^3 + \cdots \dots \dots + t\sum {X_i}^{k+1} \dots \dots \dots \dots \dots .(F)$$

Therefore, by the conditions

$$\frac{dW}{da} = 0 = \frac{dW}{db} = 0 = \frac{dW}{dc} = 0 \qquad \dots .. = \frac{dW}{dt} = 0$$

Ultimately we get the following (k+1) equations

$$\sum_{i=1}^{n} Y_i = na + b\sum X_i + c\sum {X_i}^2 + \cdots \dots \dots + t\sum {X_i}^k.$$

$\sum_{i=1}^{n} Y_i X_i = a\sum X_i + b\sum {X_i}^2 + c\sum {X_i}^3 + \cdots \dots \dots + t\sum {X_i}^{k+1}$.................... (G)

$$\sum_{i=1}^{n} Y_i {X_i}^k = a\sum {X_i}^k + b\sum {X_i}^k + c\sum {X_i}^k + \cdots \dots \dots + t\sum {X_i}^{2k}$$

These equations are solved as simultaneous equations and give the value of (k+1) constants a, b, c, …, t. Substitution of these values in second order partial derivatives gives positive value of the function. Positive value of the function indicates that the values of a, b, c,…, t obtained by solving the set of equations (G), minimize U which is sum of squares of residuals . With these values of a, b, c,…, t , curve in equation (A) is the curve of best fit.

3.6. Regressions

Regression analysis is a method of analyzing the variability of a dependent variable by resorting to information available on one or more independent variables.

Simple regression: The regression analysis confined to the study of only two variables at a time is called the simple regression.

Multiple Regression: The regression analysis for studying more than two variables at a time is known as multiple regression.

This chapter is confined only to the study of simple regression.

Linear Regression: If the regression curve is a straight line, then there is a linear regression between the variables under study. In other words, in linear regression the relationship between the two variables X and Y is linear.

Non linear Regression: If a curve or regression is not a straight line, i.e., not a first degree equation in the variables X and Y then it is called a non-linear or curvilinear regression. In this case the regression equation will have a functional relation between the variables X and Y involving terms in X and Y of the degree higher than one, i.e., involving terms of the type X^2, Y^2, X^3, Y^3,XY, etc.

Utility of Regression Analysis:

1. The cause and effect relations are indicated from the study of regression analysis.
2. It establishes the rate of change in one variable in terms of the changes in another variable.
3. It is useful in economic analysis as regression equation can determine an increase in the cost of living index for a particular increase in general price level.
4. It helps in prediction and thus one can estimate the values of unknown quantities.
5. It helps in determining the coefficient of correlation as: $r = \sqrt{b_{yx} \times b_{xy}}$
6. It enables us to study the nature of relationship between the variables.
7. It can be useful to all natural. Social and physical sciences where the data are in functional relationship.

Purpose of Regression Analysis

An important use of statistical methods is to forecast or predict future events. Insurance companies sometimes set premiums on the basis of statistical predictions. The cost of automobile insurance for minors is greater than that for adults because age often correlates with frequencies of accidents. Colleges usually admit and reject applicants primarily on the basis of predictions about their probated future scholastic performance made from the scholastic aptitude tests and academic performance in high school.

Delinquency and dropout prevention programs frequently use early indicators or predictors in identifying persons who appear likely to become delinquents or dropouts. In vocational counseling and personnel selection, implicit or explicit predictions of various job related criteria are made from variables such as age interests aptitudes, sex and experience. These examples involve prediction. The degree or reliance on statistical considerations in making these predictions varies greatly from one application to another. Insurance companies rely heavily on statistical predictions, whereas the selection of employees is rarely made on purely statistical considerations.

By using statistical methods, the accuracy of predictions of a dependent variable (a criterion or outcome variable) from one or more independent (Predictor) variables can be maximized. In statistical terms, the dependent variable Y is said to be a function of the independent variable X. No causal relationship is assumed. Indeed, causation is beside the point in forecasting. The higher the correlation, the better the prediction; the lower the correlation, the greater the margin of error in predictions. The simplest type of prediction involves predicting a dependent variable Y from only one independent variable X when both X and Y are normally distributed.

3.7. Linear Regression Model

Suppose X and Y are perfectly linearly positively correlated, the simplest type of causal effect, the liner effect is represented by a straight line:

$$Y = a + bX \qquad \text{…………(2.1)}$$

Y is also known as dependents variable, response variable and endogenous variable, whereas X is known as in dependent variable, predictor variable exogenous variable. The line called the regression line of Y on X.

Where constant a = the value of Y when X = 0.

This constant ‘a’ is also called the Y- intercept because (when a line is plotted on a graph), ‘a’ is the value of Y at the point where the line crosses the Y-axis. We know that the equation of Y axis is X = 0.

b = the slope of the line.

The **slope** represents the change Y per unit increase in X. If X were a cause of Y, then a one unit increase in X would cause Y to change by b units. Thus b represents the effect of X on Y. If b > 0, Y is increasing with X. If b < 0, Y is decreasing with X.

The following figure shows a graph of a perfect positive linear relationship between X and Y where a = 1 and b = 0 .50. Note that if there are only two points (a sample of size n = 2), the relationship will always be perfect. The line is known if constants a and b are given. For example, a = 1, b = 0.50 gives the line as

$$Y = 1 + 0.50\ X$$

We can plot this line on the graph paper, by taking only two points (i.e., A sample of size n=2), the relationship will be perfect.

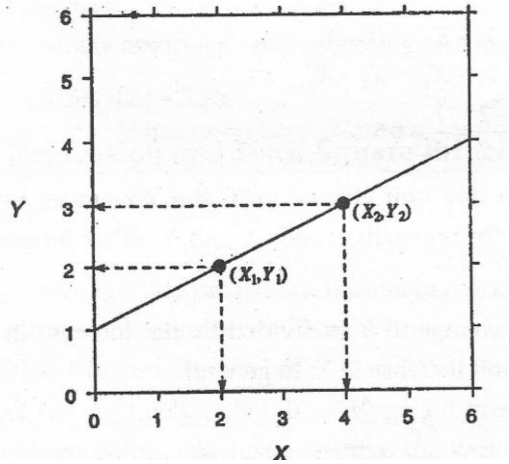


Fig.2.1

The line can be plotted by first taking any two values of X and computing their corresponding values of Y from the formula. The two pairs of X, Y values form two pairs of coordinates (X,Y) through which the line must pass. Suppose our sample on X variable yields. The coordinates in the above figure are determined by arbitrarily choosing X_1=2 and X_2=4.

If X_1= 2, Y_1= 1+.50(2)=2 giving the point (X_1,Y_1) = (1,2)

If X_2= 2, Y_2= 1+.50(4)=3 giving the point (X_2,Y_2) = (4,3)

The plotted line Y = 1 + 0.50 X passes through these points (1,2), (4,3). This line can be used for further prediction of Y value for any X value.

Conversely, if we were given the straight line drawn on the graph but did not know equation of the line we could determine it from the graph.

The slop ‘b’ could be determined by first selecting any two points on the line. One of these points could be the point at which the line crosses the Y-axis which is (0, a). In the above figure it is (0,1) from each of the two points drop a vertical line down to the X-axis to determine the X-values and run a horizontal line to the Y axis to determine Y values. The value of the slope will then be computed by subtracting the smaller X value from the larger X value to get the increase in X and then dividing the corresponding change in Y by this increase in X, as follows:

$$b_{yx} = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{Y_1 - Y_2}{X_1 - X_2}$$

Here

$$b = \frac{3-2}{4-2} = \frac{1}{2} = 0.5$$

Thus with one unit change in X and the Y changes by 0.50 unit, and since the sign of ‘b’ is positive, both X and Y either increase or decrease simultaneously.

When the change in Y is divided by the increase in X we get the change in Y per unit increase in X. In general

$$b_{yx} = \frac{\Delta y}{\Delta x} \qquad \dots\dots\dots\dots(2.2)$$

Where, Δy is change in Y and Δx is the corresponding change in X values.

Example 3.3

Suppose you are told that for a group of 20 students, there is a perfect linear relationship between grade point average (Y) and scores on and intelligence test (X). Suppose you are also told that the equation describing the relationship is:

Y = 1.00 + 0.025(X)

If an individual obtained a score of 100 on the inelegance test, (i) what must his or her grade point average be? What must the students’ grade point average be if he/she obtained an intelligence test score of 97? And of 108?

Solution:

Here X = 100

The grade point average associated with an intelligence test score to 100 is

Y =1.00 +(0.25)(100) =1.00 + 2.50 = 3.50

For X = 97:

The grade point average associated with an intelligence test score of 97 is

Y= 1.00 + (0.25) (97) = 3.425

For X = 108:

The grade point average associated with intelligence test score of 108 is

Y = 1.00 + (0.25) (108) = 3.70

3.9. Regression Lines

Suppose that there are N pairs of observations (X_i, Y_i), i= 1, 2, 3...N. Let the regression line of Y on X be

Y= a+bx

So that error sum of squares = $\sum e_i = \sum(Y_i - \hat{a} - b\bar{X}_i)^2 = S^2$ ……… (2.7)

The desired estimates of a and b are obtained by solving the simultaneous equations, called the normal equations.

$$\frac{\delta S^2}{\delta a} = 0, \frac{\delta S^2}{\delta b} = 0$$

Or
$$\left.\begin{aligned} \textstyle\sum_i (Y_i - a - bX_i) = 0 \\ \textstyle\sum X_i\,(Y_i - a - bX_i) = 0 \end{aligned}\right\}$$

$$and\ i.e., \quad \sum_i Y_i = Na + b\sum_i X_i$$

$$and \sum_i X_i Y_i = a \sum_i X_i + b \sum_i X_i^2 \Bigg\} \qquad \ldots \ldots \ldots (2.8)$$

The roots of the equations are

$$b = \frac{N \sum_i X_i Y_i - (\sum_i X_i)(\sum_i Y_i)}{N \sum_i X_i^2 - (\sum_i X_i)^2} \qquad \ldots \ldots \ldots (2.9)$$

$$= \frac{\sum_i \frac{X_i Y_i}{N} - \overline{XY}}{\frac{(\sum_i X_i)^2}{N} - \bar{X}^2}$$

$$\frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})/N}{\sum_i (X_i - \bar{X})^2/N}$$

$$= \frac{Cov\ (X,Y)}{var\ (X)} = r \frac{\sigma_y}{\sigma_x} \qquad \ldots \ldots \ldots \ldots (2.10)$$

And $\hat{a} = \bar{Y} - \hat{b}\bar{X}.$ ....(2.11)

Substituting these values in (2.3) we have the desired prediction formula:

$$Y = \bar{Y} + r \frac{\sigma_y}{\sigma_x} (X - \bar{X}) \qquad \ldots \ldots \ldots .. (2.12)$$

The line given by Eqn. (2.12) is known as regression line of Y and X.

The coefficient b is the amount by which the predicated value Y increases for a unit increment in the value of X. It is called the regression coefficient of Y on X. It is also written as b_{yx}.

The regression line of Y on X is used to get the best estimate of variable Y for any specified value of X.

The regression line of Y on X is also written as

$$Y - \bar{Y} = \mathrm{b_{yx}}\ (X - \bar{X})$$

or

$$Y = \bar{Y} + r \frac{\sigma_y}{\sigma_x} (X - \bar{X})$$

Similarly if we are interested in predicting X from Y, we use the regression lien of X on Y, which has the equation.

$$X = \bar{X} + r\frac{\sigma_y}{\sigma_x}(Y - \bar{Y}) \qquad \dots\dots\dots\dots\dots (2.13)$$

$r\frac{\sigma_y}{\sigma_x}$ is the amount by which the predicated value X increases for a unit increment in Y, is the regression coefficient of X on Y. It is denoted by $\mathrm{b_{xy}}$.

It may be noted that both the regression lines pass through the point which is their point ($\bar{X}, \bar{Y}$) of intersection.

We usually designate one variable as dependent (Y) and the other as independent (X) when using regression to estimate the effect of X on Y. It is also possible to calculate a regression equation that uses the variable labeled Y as predictor of the variable labeled X. Two researchers, for instance, might disagree about the direction of effect.

It may be instructive to examine how regression coefficients differ in the two situations.

The regression equations or regression X on Y is written as:

$$X' = a_{xy} + b_{xy}(Y)$$

Where

$$b_{xy}\frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (Y_i - \bar{Y})^2} = \frac{n\sum_i X_i Y_i - (\sum_i X_i)(\sum_i Y_i)}{n\sum_i Y_i^2 - (\sum_i Y_i)^2} \dots\dots.. (2.14)$$

and $a_{xy} = \bar{X} - b_{xy}(\bar{Y})$

Thus intercept and slope are asymmetric statistics. If the variances of X and Y were equal, then $b_{xy} = b_{xy}$.

The fact that there are two different regression lines one when Y is treated as the dependent variable and another when X is treated as dependent, does not mean they can both be used as measures of the effects of each variable on the other. They cannot both be correct. We have to choose one variable as dependent and the other as dependent and the other as independent and then use the appropriate formulas.

A line of regression is the line which gives the best estimate of one variable for any given value of the other variable.

I. ***Line of regression of X on Y.***

It is the line which gives the best estimate for the values of X for a specified value of Y.

It is given by, $X = \bar{X} + r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$

Where $\bar{X}, \bar{Y}$ are means of X series and Y series respectively are S.D. of X and Y series respectively and r is the correlation coefficient between X and Y.

It can also be put in form:

$$X = \hat{\alpha} + \hat{\beta}_y$$

Where $\hat{\alpha}$ is intercept of the line and $\hat{\beta}$ is the slope of the line X on Y.

II. ***Line of Regression of Y on X.***

It is the line which gives the best estimates for the values of Y for any specific values of X.

Regression equation of Y on X is given by:

$$Y = \bar{Y} + r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

It can also be put in the form

Y= a + bX,

Where a is the intercept of the line and b is the slope of the line Y on X.

3.10. Regression Coefficients

The regression coefficient of Y on X is $b_{yz} = r\frac{\sigma_y}{\sigma_x}$ and that of x on y is $b_{xy} = r\frac{\sigma_y}{\sigma_x}$

3.11. Properties Relating to Regression

Consider any one of the regression lines; say that of Y on x, it has the following properties:

Property I. Regression coefficients are independent of change of origin but not of scale.

$$let\ u = \frac{x - A}{c}\ and\ v = \frac{y - B}{d}, where\ c > and\ d > 0 \qquad \ldots (2.15)$$

Then the regression coefficient of y on x denoted by b_{yx} for the sake of definiteness is

$$b_{yx} = \frac{cov(x, y)}{var(x)} = \frac{cd\ cov\ (u, v)}{c^2 var(u)} = \frac{d}{c} \times \frac{cov(u, v)}{var(u)} = \frac{d}{c} \times b_{vu} \qquad \ldots (2.16)$$

or

$$b_{yx} = \frac{d}{c} \times \frac{n \sum_i u_i v_i - (\sum_i u_i)(\sum_i v_i)}{n \sum_i u_i^2 - (\sum_i u_i)^2} \qquad \ldots \ldots (2.17)$$

The other constants in the regression equation are, in terms of u and v.

$\bar{y} = B + d\bar{v}$

And

$\bar{x} = A + c\bar{u}$

Property II.

Since $\quad \hat{y} = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x_i - \bar{x}), \qquad \ldots \ldots (2.18)$

$$\therefore \hat{Y} = \frac{1}{n}\sum_i \hat{Y}_i = \frac{\left[n\bar{y} + r\frac{\sigma_y}{\sigma_x}\ (x_i - \bar{x})\right]}{n} \qquad \ldots \ldots (2.19)$$

Dividing both sides by n and remembering that $\sum_i(x_i - \bar{x}) = 0$, we have

$\hat{\bar{\hat{Y}}} = \bar{y} \qquad \ldots\ldots\ldots\ (2.20)$

In words the mean of the observed values of y is equal to the mean of the corresponding predicted values.

Property II the mean of the errors of estimates, $e_i = y_i - \hat{Y}_i,$ is zero since

$$e = \frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\left(y_i - \hat{Y}_i\right) = \bar{y} - \hat{\bar{Y}} = 0$$

Property III.

Again the residual variance, var (e), is given by

$$n\ var\ (e) = \sum_i e_i^2 \qquad [\therefore \bar{e} = 0]$$

$$= \sum_i (y_i - Y_i)^2 \qquad \dots \dots \dots . . (2.21)$$

$$= \sum_i \left\{(y_i - \bar{Y}) - r\frac{\sigma_y}{\sigma_x}\ (x_i - \bar{x})\right\}^2$$

$$= \sum_i (y_i - \bar{Y}\ \ )^2 - 2r\frac{\sigma_y}{\sigma_x} \times \sum_i (x_i - \bar{x})(y_i - \bar{y}) + r^2\frac{\sigma_y}{\sigma_x}\ (x_i - \bar{x})^2$$

$$= ns_y^2 - 2r\frac{\sigma_y}{\sigma_x} \times nr\sigma_x\sigma_y + r^2\frac{\sigma_y}{\sigma_x} \times n\sigma_x^2 \qquad \dots \dots \dots \dots . . (2.22)$$

Hence, var (e) = $\sigma_x^2(1 - r^2)$ ……………. (2.23)

The standard deviation of e, which is called the standard error of estimate of y from its linear regression on X, is denoted by We have, then,

$$\sigma_{yx} = \sigma_y\sqrt{1 - r^2} \qquad \dots \dots \dots . (2.24)$$

Since var (e)≥ 0, we have

$$r^2 \leq 1 \quad or \qquad -1 \leq r \leq 1 \qquad \dots \dots \dots . (2.25)$$

A result which has already been proved in a different way.

Property IV.

We have seen that

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} \qquad \dots \dots \dots (2.26)$$

and

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} \qquad \text{… … (2.27)}$$

Hence

$$b_{yx} \times b_{xy} = r^2$$

Or

$$|r| = \sqrt{b_{yx} \times b_{xy}} \qquad \text{… …..(2.28)}$$

Thus numerically the correlation coefficient is the geometric mean of the two regression coefficients. As regards the sign of r, it is the same as the common sign of the two regression coefficients.

Property V.

If one of the regression coefficients is greater than unity, then the other is less than unity.

Proof.

Let b_{yx} >1. Also we know that

$$r^2 \leq 1 \quad and \quad r^2 = b_{yx} \times b_{xy}$$

$$b_{yx} \times b_{xy} \leq 1 \qquad [\therefore r \leq 1]$$

$$b_{yx} \leq \frac{1}{b_{yx}} < 1$$

Property VI.

Arithmetic mean of the regression coefficient is greater than the correlation coefficient.

Property VII.

Regression coefficients are Independent of change of origin but not of scale.

Property VIII.

Both regression coefficients are independent of change of origin but not of scale.

Property IX.

The sign of correlation is same as that of regression coefficients, i.e., r >0 if b_{xy} > 0; and r < 0, if regression coefficients are negative.

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{r\sigma_y\sigma_x}{\sigma_x} = \frac{Cov(X,Y)}{\sigma_x^2}$$

Similarly

$$b_{xy} = \frac{Cov(X,Y)}{\sigma_x^2}$$

And

$$r = \frac{Cov(X,Y)}{\sqrt{\sigma_x\sigma_y}}$$

Thus signb_{yx}, b_{xy} of and r is same as *Cov* (X,Y).

Example 3.4

Using the data in Example 12.11 the statistic on median education (X) and average teacher salary (Y) are given below:

Mean of X = 12.47	Mean of Y = 19991.5
Sum of squares for	Sum of squares for
X = 0.5629	Y = 197774971.6

Sum of cross products = 5110.49

(a) Use these statistics to find the regression equation for predicting average teacher salary from a state's median education level.

(b) Find the correlation coefficient and coefficient of determination and describe the relationship.

(c) In what state is the average teacher salary the farthest below what we would expect or predict, based on median education? In what state is the average teacher salary the farthest above what would predict?

Solution:

(a) b = 5110.49/ .5629= 9078.86

a = 19991.5-9078.86 (12.47) = 93221.9

Y = -93221.9 + 9078.86 (X)

(b) r = 5110.49/ $\left(\sqrt{.5629}\,\sqrt{19777497.6}\right)$ =.484

$r^2 = R^2 = 0.464^2 = 0.235$

There is a moderate positive relationship between both variable. States with higher median education levels tend also to have higher average teacher salaries. Median education level explains 23.5% of the variation in average teacher salaries.

(c) In Vermont the average teacher salary is relatively low (16299).

Based on Vermont's above average median education level 912.6), we would predict a much higher average salary: Y= 21172. Vermont has the largest negative residual (-4873) among the 21 states. Similarly, New York is found to have the largest residual.

3.12. Summary

While dealing with qualitative characteristics like intelligence, it is advisable not to use the actual measurements for the calculation of correlation coefficient. Not only those different examiners in such a case will award different marks for the same intelligence, the same examiner at two different times may award different marks for the same intelligence. In such cases, therefore ordinal numbers are allotted to actual measurements either in ascending or descending order. These ordinal numbers are called ranks. The correlation coefficient between the ranks of two such variables or characteristics is known as rank correlation coefficient. Regression is mainly concerned with bringing out the nature of relationship between variables and using it to know the best approximate value of one variable corresponding to a known value of the other variable. The relationship between any two variable may be linear or nonlinear. A relationship may be described by means of a straight line or a curve. If it is best explained by a straight line. It is called linear regression. If it is described more appropriately by a curve, it is said to be non linear regression. The correlation coefficient between the members of same class is known as intra class correlation coefficient.

3.13. Terminal questions

Q.1. What is the scatter diagram? Disuses it with examples.

Answer: --

Q.2. What is the karl Pearson’s coefficient of correlation write the properties of correlation coefficient.

Answer: --

Q.3. Discuss the spearman’s coefficient with examples.

Answer: --

Q.4. What are the regressions? Discuss the linear regression model.

Answer: --

Q.5. What are the lines of regression? Write the properties of regression coefficients.

Answer: --

3.14. Further Readings

1. Goon, Gupta & Dasgupta: Fundamentals of Statistics Vol. I, The World Press Pvt. Ltd., Kolkata.

2. Yule G.U. and Kendall, M.G.: An Introduction to the Theory of Statistics, Charles Griffin and Company Ltd.

3. C.E. Weatherburn: Mathematical Statistics.

4. Introduction toStatistics, David Lane, Rice University

5. Basic Statistics, B.L. Agrawal, New Age International Private Limited

6. Basic Statistics, Thomas Higher Education Textbooks

Block-2 PGEVS-106N

Rajarshi Tandon Open

University, Prayagraj

Numerical

and

Statistical Computing

Block- 2

Probability and Testing of Hypothesis

Unit-4

Probability & Distribution Theory

Unit-5

Theory Of Hypothesis

Unit-6

ANOVA (Analysis Of Variance)

Rajarshi Tandon Open

University, Prayagraj

Numerical and Statistical Computing

Course Design Committee

Prof. Ashutosh Gupta **Chairman**

School of Science, UPRTOU, Prayagraj

Dr. Uma Rani Agarwal **Member**

Rtd. Professor, Department of Botany

CMP Degree College, Prayagraj

Dr. Ayodhaya Prasad Verma **Member**

Red. Professor, Department of Botany

B.S.N.V. P.G. College, Lucknow

Dr. Sudhir Kumar Singh **Member**

Assistant Professor

K. Banerjee Centre for Atmospheric and Ocean Studies

University of Allahabad, Prayagraj

Dr. Ravindra Pratap Singh **Member**

Assistant Professor (Biochemistry)

School of Science, UPRTOU, Prayagraj

Dr. Dharmveer Singh **Course Coordinator**

Assistant Professor (Biochemistry)

School of Science, UPRTOU, Prayagraj

Course Preparation Committee

Dr. Anuj Kumar Singh **Author** **Block-1** (Unit: 1)

Assistant Prof. (Statistics)

School of Sciences, UPRTOU, Prayagraj

Dr. Upasana Singh **Author** **Block-1&2** (Unit: 2-6)

Assistant Professor-Zoology

Prof. Rajendra Singh Rajju Bhaiya

University, Prayagraj

Dr. Jaspal Singh Assistant Professor Department of Environmental Science, Bareilly College, Bareilly	**Author**	**Block-1&3,4** (Unit: 1,7,8,9,10,11)
Dr. Nishtha Seth Associate Professor Department of Environmental Science, Bareilly College, Bareilly	**Author**	**(All blocks and units)**

Dr. Dharmveer Singh
(Course and SLM Coordinator)

School of Sciences, UPRTOU, Prayagraj

Introduction

This second block of numerical and statistical computing, this consists of following three units:

Unit-4: This subject covers probability and distribution theory, including additive and multiplicative laws of probability, conditional probability, probability mass functions, and so on. Probability density functions, binomial distributions, poisson distributions, and normal distributions are all explored.

Unit-5: This unit discusses hypotheses and their types, as well as their significance levels, critical regions, and degrees of freedom. The P-value, types of errors, Z-test, t-test, F-test, and Chi-Square Tests, as well as their applications, are all described.

Unit-6: This unit gives an introduction to ANOVA. It discusses variance analysis for one- and two-way classifications. This unit also discusses the ANOVA with missing observations (one and two), as well as the analysis of covariance.

Unit-4: Probability & Distribution Theory

Content

4.1. Introduction

In this unit will cover the topics Probability theory, Additive and multiplicative law of probability**,** Conditional probability, Probability mass functions, Probability density functions Binomial distribution, Poisson distribution, Normal distribution.

Data generally demands an inference to be drawn and statistical inferences are never certain as in deterministic experiments. Statisticians tend to use prefixes or suffixes as "likely" or "almost likely". Understanding of probability ad its concepts is, thus, very important to interpret inferential statistics. In randorn experiment set all possible outcomes may be known but it is not possible to predetermine the outcome.

Objectives:

After reading this unit, you will be able:

- to understand the concept of probability
- to understand the laws of probability including Bayes' Theorem
- to understand the concept of probability distributions
- to understand the concept of binomial distribution
- to understand the concept of Poisson distribution
- to understand the concept of Normal distribution.

4.1. Probability

Probability is a discipline of mathematics that studies the chance of events occurring. It is a metric that quantifies the uncertainty or probability of a specific event occurring. Probability values range from 0 to 1, with 0 representing an impossible event and 1 indicating a certain event. A higher likelihood of an event means that it is more likely to occur.

Sample Spaces

A set S that consists of all possible outcomes of a random experiment is called a sample space, and each outcome is called a sample point. Often there will be more than one sample space that can describe outcomes of an experiment, but there is usually only one that will provide the most information

Events

An event is a subset A of the sample space S, i.e., it is a set of possible outcomes. If the outcome of an experiment is an element of A, we say that the event A has occurred. An event consisting of a single point of S is often called a simple or elementary event.

Axioms and Theorems of Probability

In axiomatic approach, probability is introduced as a function of outcomes of an experiment, under certain restrictions. These restrictions are called Axioms of Probability

Axiom 1*:* The probability *of* an event is a non-negative real number; that is, $0 \leq P(A) \leq 1$. for any subset *A* of S.

Axiom 2*: P(S)* = *1,* i.e. the probability of outcome space is 1

Axiom 3*: If* $A_1, A_{2,} \ldots \ldots A_n$ is a finite or infinite sequence of mutually exclusive events of S, then

$$P(A_1 \cup A_2 \cup A_3 \ldots \ldots \ldots \ldots \cup A_n) = P(A_1) + P(A_2) + \cdots \ldots \ldots \ldots \ldots P(A_n)$$

The first axiom implies that the probability of an event is a non-negative number less than or equal to unity. The second axiom implies that the probability of an event that is certain to occur must be equal to unity and the last axiom gives a basic rule of addition of probabilities.

Bayes' Theorem:

Before stating the theorem it is necessary to know about *prior* and *posterior probabilities.*

Prior probabilities*:* These are the probabilities, assigned to events on the basis of conditions, past experience or judgment.

Posterior probabilities*:* These are the revised probabilities in the light of some additional information.

Bayes' theorem states that if an event B can occur in combination with any of the n mutually exclusive and exhaustive events A_i's and if B is found to have occurred, then the probability that it was preceded by a particular event A_i is given by

$$P(A_i/B) = \frac{P(A_i)P({}^{B}/_{A_i})}{\sum_{i=1}^{n} P(A_i)P({}^{B}/_{A_i})}$$

Example: The contents of urns I, II & III are as follows:

1 white, 2 black and 3 red balls

2 white, 1 black and 1 red balls

4 white, 5 black and 3 red balls

They happen to be white and red. What is the probability that they come from urn II?

Solution:

LetE_1, E_2 and E_3 denote the events that the urn I, II and III is chosen, respectively, and let A be the event that the two balls taken from the selected urn are white and red.

$$P(E_1) = P(E_2) = P(E_3) = \frac{1}{3}$$

$$P\left({}^{A}/_{E_1}\right) = \frac{1}{5}$$

$$P\left({}^{A}/_{E_2}\right) = \frac{1}{3}$$

$$P\left({}^{A}/_{E_3}\right) = \frac{2}{11}$$

$$P(E_2/A) = \frac{P(E_2)P\left({}^{A}/_{E_2}\right)}{\sum_{i=1}^{3} P(E_i)P\left({}^{A}/_{E_i}\right)}$$

$$= \frac{\mathbf{55}}{\mathbf{1}}$$

Event - A possible outcome of a trial is called an event. Thus, head is an event that may result from tossing a coin. Similarly, the occurrence of five or the occurrence of an odd number is a possible event of the trial of having a dice. The latter example indicates that an event may consist of one or more possible outcomes of an experiment. The event of getting an odd number, in fact, consists of three possible outcomes of rolling a dice. We should that *km* event .consisting of only one possible outcome is often called an elementary event

Equally Likely Outcomes:

In some experiments all the outcomes have the same chance of happening. If we roll a fair die the chances are the same for rolling a two or rolling a five. If we draw a single card from a well shuffled deck of cards, each card has the same chance of being selected. We call outcomes like these equally likely. Drawing names from a hat or drawing straws are other examples of equally likely outcomes. The tack tossing example did not have equally likely outcomes since the probability of the tack landing point up is different than the probability of the tack landing point down.

An experiment has **equally likely outcomes** if every outcome has the same probability of occurring.

For equally likely outcomes, the **probability of outcome A**, *P(A)*, is:

$$P(A) = \frac{\text{number of ways for A to occur}}{\text{total number of outcomes}}.$$

Example 1: Simple Probabilities with Cards

Draw a single card from a well shuffled deck of 52 cards. Each card has the same chance of being drawn so we have equally likely outcomes. Find the following probabilities:

a. *P*(card is red)

$$P(\text{card is red}) = \frac{\text{number of red cards}}{\text{total number of cards}} = \frac{26}{52} = \frac{1}{2}$$

The probability that the card is red is $\frac{1}{2}$.

b. P(card is a heart)

$$P(\text{card is a heart}) = \frac{\text{number of hearts}}{\text{total number of cards}} = \frac{13}{52} = \frac{1}{4}$$

The probability that the card is a heart is $\frac{1}{4}$.

c. P(card is a red 5)

$$P(\text{card is a red 5}) = \frac{\text{number of red fives}}{\text{total number of cards}} = \frac{2}{52} = \frac{1}{26}$$

The probability that the card is a red five is $\frac{1}{26}$.

Example 2 : Simple Probabilities with a Fair Die

Roll a fair die one time. The sample space is S = {1, 2, 3, 4, 5, 6}. Find the following probabilities.

a. P(roll a four)

$$P(\text{roll a four}) = \frac{\text{number of ways to roll a four}}{\text{total number of ways to roll a die}} = \frac{1}{6}$$

The probability of rolling a four is $\frac{1}{6}$.

b. P(roll an odd number)

The event roll an odd number is E = {1, 3, 5}.

$$P(\text{roll an odd number}) = \frac{\text{number of ways to roll an odd number}}{\text{total number of ways to roll a die}} = \frac{3}{6} = \frac{1}{2}$$

The probability of rolling an odd number is $\frac{1}{2}$.

c. P(roll a number less than five)

The event roll a number less than five is F = {1, 2, 3, 4}.

$$P\left(\text{roll a number less than five}\right) = \frac{\text{number of ways to roll number less than five}}{\text{total number of ways to roll a die}} = \frac{4}{6} = \frac{2}{3}$$

The probability of rolling a number less than five is $\frac{2}{3}$.

Three Ways of Finding Probabilities:

There are three ways to find probabilities.

1.Theoretical probability: A theoretical probability is based on a mathematical model where all outcomes are equally likely to occur. The probability of getting a red jack in a card game or rolling a five with a fair die can be calculated from mathematical formulas. These are examples of theoretical probabilities.

2.Empirical probability: An empirical probability is based on an experiment or observation and is the relative frequency of the event occurring. In the tack tossing example we calculated the probability of the tack landing point up by doing an experiment and recording the outcomes. This was an example of an empirical probability**.**

3.Subjective probability: A subjective probability is an estimate (a guess) based on experience or intuition. Saying that there is an 80% chance that you will go to the beach this weekend is a subjective probability. It is based on experience or guessing.

Complements:

If there is a 75% chance of rain today, what are the chances it will not rain? We know that there are only two possibilities. It will either rain or it will not rain. Because the sum of the probabilities for all the outcomes in the sample space must be 100% or 1.00, we know that

P(will rain) + P(will not rain) = 100%.

Rearranging this we see that

P(will not rain) = 100% - P(will rain) = 100% - 75% = 25%.

The events E = {will rain} and F = {will not rain} are called complements.

The complement of event E, denoted by $\bar{E}$, is the set of outcomes in the sample space that are not in the event E. The probability of $\bar{E}$ is given by $P\left(\bar{E}\right)=1-P\left(E\right)$.

4.2. Probability Distributions:

A probability distribution (probability space) is a sample space paired with the probabilities for each outcome in the sample space. If we toss a fair coin and see which side lands up, there are two outcomes, heads and tails. Since the coin is fair these are equally likely outcomes and have the same probabilities. The probability distribution would be *P*(heads) = 1/2 and *P*(tails) = 1/2. This is often written in table form

Table: Probability Distribution for a Fair Coin

Outcome	Heads	Tails
Probability	1/2	1/2

A **probability distribution** for an experiment is a list of all the possible outcomes and their corresponding probabilities

Example: Probabilities for the Sum of Two Fair Dice

In probability problems when we roll two dice, it is helpful to think of the dice as being different colors. Let’s assume that one die is red and the other die is green. We consider getting a three on the red die and a five on the green die different than getting a five on the red die and a three on the green die. In other words, when we list the outcomes the order matters. The possible outcomes of rolling two dice and looking at the sum are given in Table 3.1.8.

Table: All Possible Sums of Two Dice

1+1 = 2	1+2 = 3	1+3 = 4	1+4 = 5	1+5 = 6	1+6 = 7
2+1 = 3	2+2 = 4	2+3 = 5	2+4 = 6	2+5 = 7	2+6 = 8
3+1 = 4	3+2 = 5	3+3 = 6	3+4 = 7	3+5 = 8	3+6 = 9

4+1 = 5	4+2 = 6	4+3 = 7	4+4 = 8	4+5 = 9	4+6 = 10
5+1 = 6	5+2 = 7	5+3 = 8	5+4 = 9	5+5 = 10	5+6 = 11
6+1 = 7	6+2 = 8	6+3 = 9	6+4 = 10	6+5 = 11	6+6 = 12

Table: Probability Distribution for the Sum of Two Fair Dice

Sum	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
Reduced Probability	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

Mutually Exclusive Events:

An experiment consists of drawing one card from a well shuffled deck of 52 cards. Consider the events *E*: the card is red, *F*: the card is a five, and *G*: the card is a spade. It is possible for a card to be both red and a five at the same time but it is not possible for a card to be both red and a spade at the same time. It would be easy to accidentally count a red five twice by mistake. It is not possible to count a red spade twice.

Remark: Two events are **mutually exclusive** if they have no outcomes in common.

Example 3.2.2: Mutually Exclusive with Dice

Two fair dice are tossed and different events are recorded. Let the events *E*, *F* and *G* be as follows:

E = {the sum is five} = {(1, 4), (2, 3), (3, 2), (4, 1)}

F = {both numbers are even} = {(2, 2), (2, 4), (2, 6), (4, 2), (4, 4), (4, 6), (6, 2), (6, 4), (6, 6)}

G = {both numbers are less than five} = {(1, 1), (1, 2), (1, 3), (1, 4), (2, 1), (2, 2), (2, 3), (2, 4), (3, 1), (3, 2), (3, 3), (3, 4), (4,1), (4, 2), (4, 3), (4,4)}

a. Are events *E* and *F* mutually exclusive?

   Yes. *E* and *F* are mutually exclusive because they have no outcomes in common. It is not possible to add two even numbers to get a sum of five.

b. Are events *E* and *G* mutually exclusive?

   No. *E* and *G* are not mutually exclusive because they have some outcomes in common. The pairs (1, 4), (2, 3), (3, 2) and (4, 1) all have sums of 5 and both numbers are less than five.

c. Are events *F* and *G* mutually exclusive?

   No. *F* and *G* are not mutually exclusive because they have some outcomes in common. The pairs (2, 2), (2, 4), (4, 2) and (4, 4) all have two even numbers that are less than five.

Addition Rule for “Or” Probabilities:

The addition rule for probabilities is used when the events are connected by the word “or”. Remember our teacher in Example 3.2.1 at the beginning of the section? She wanted to know the probability that her students were taking either art or English. Her problem was that she counted some students twice. She needed to add the number of students taking art to the number of students taking English and then subtract the number of students she counted twice. After dividing the result by the total number of students she will find the desired probability. The calculation is as follows

$$\begin{aligned} P\left(\text{art or English}\right) &= \frac{\text{\# taking art + \# taking English - \# taking both}}{\text{total number of students}} \\ &= \frac{13+21-9}{30} \\ &= \frac{25}{30} \approx 0.833 \end{aligned}$$

The probability that a student is taking art or English is 0.833 or 83.3%.

When we calculate the probability for compound events connected by the word “or” we need to be careful not to count the same thing twice. If we want the probability of drawing a red card or a five we cannot count the red fives twice. If we want the probability a person is blonde-haired or blue-eyed we cannot count the blue-eyed blondes twice. The addition rule for probabilities adds the number of blonde-haired people to the number of blue-eyed people then subtracts the number of people we counted twice.

Addition Rule for “Or” Probabilities

If *A* and *B* are any events then, *P(A* or *B) = P(A) + P(B) – P(A* and *B).*

If *A* and *B* are mutually exclusive events then *P(A* and *B) = 0,* so then *P(A* or *B) = P(A) + P(B).*

Example 3.2.3: Additional Rule for Drawing Cards

A single card is drawn from a well shuffled deck of 52 cards. Find the probability that the card is a club or a face card.

There are 13 cards that are clubs, 12 face cards (J, Q, K in each suit) and 3 face cards that are clubs.

$$\begin{aligned} P(\text{club or face card}) &= P(\text{club}) + P(\text{face card }) - P(\text{club and face card}) \\ &= \frac{13}{52} + \frac{12}{52} - \frac{3}{52} \\ &= \frac{22}{52} = \frac{11}{26} \approx 0.423 \end{aligned}$$

The probability that the card is a club or a face card is approximately 0.423 or 42.3%.

Independent Events:

Sometimes we need to calculate probabilities for compound events that are connected by the word "and." We have two methods to choose from, independent events or conditional probabilities (Section 3.3). Tossing a coin multiple times or rolling dice are independent events. Each time you toss a fair coin the probability of getting heads is ½. It does not matter what happened the last time you tossed the coin. It's similar for dice. If you rolled double sixes last time that does not change the probability that you will roll double sixes this time. Drawing two cards without replacement is not an independent event. When you draw the first card and set it aside, the probability for the second card is now out of 51 cards not 52 cards.

"Two events are independent events if the occurrence of one event has no effect on the probability of the occurrence of the other event."

Multiplication Rule for "And" Probabilities: Independent Events

If events *A* and *B* are independent events

Then $P(A \text{ and } B) = P(A) \cdot P(B)$.

Example: Independent Events for Tossing Coins

Suppose a fair coin is tossed four times. What is the probability that all four tosses land heads up?

The tosses of the coins are independent events. Knowing a head was tossed on the first trial does not change the probability of tossing a head on the second trial.

$$\begin{aligned} P(\text{four heads in a row}) &= P(\text{1st heads and 2nd heads and 3rd heads and 4th heads}) \\ &= P(\text{1st heads}) \cdot P(\text{2nd heads}) \cdot P(\text{3rd heads}) \cdot P(\text{4th heads}) \\ &= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \\ &= \frac{1}{16} \end{aligned}$$

The probability that all four tosses land heads up is $\frac{1}{16}$

At Least Once Rule for Independent Events:

Many times we need to calculate the probability that an event will happen at least once in many trials. The calculation can get quite complicated if there are more than a couple of trials. Using the complement to calculate the probability can simplify the problem considerably. The following example will help you understand the formula.

At Least Once Rule

If an experiment is repeated n times, the n trials are independent and the probability of event A occurring one time is P(A) then the probability that A occurs at least one time is:

$P\left(A \text{ occurs at least once in n trials}\right)=1-P\left(\bar{A}\right)^{n}$

<u>“And” Probabilities from Two-Way</u>:

“And” probabilities are usually done by one of two methods. If you know the events are independent you can use the rule $P\left(A \text{ and } B\right)=P\left(A\right)\cdot P\left(B\right)$. If the events are not independent you can use the conditional probabilities in Section 3.3. There is an exception when we have data given in a two-way table. We can calculate “and” probabilities without knowing if the events are independent or not.

4.4. Conditional Probabilities

What do you think the probability is that a man is over six feet tall? If you knew that both his parents were tall would you change your estimate of the probability? A conditional probability is a probability that is based on some prior knowledge.

“A **conditional probability** is the probability that an event will occur if some other condition has already occurred. This is denoted by $P\left(A \mid B\right)$, which is read “the probability of *A* given *B*.”

Example: Conditional Probability for Drawing Cards without Replacement

Two cards are drawn from a well shuffled deck of 52 cards without replacement. Find the following probabilities.

(a) The probability that the second card is a heart given that the first card is a spade.

Without replacement means that the first card is set aside before the second card is drawn and we assume the first card is a spade. There are only 51 cards to choose from for the second card. Thirteen of those cards are hearts.

It’s important to notice that the question only asks about the second card.

$$P\left(\text{2nd heart} \mid \text{1st spade}\right) = \frac{13}{51}$$

The probability that the second card is a heart given that the first card is a spade is $\frac{13}{51}$

(b) The probability that the first card is a face card and the second card an ace.

Notice that this time the question asks about both of the cards.

There are 12 face cards out of 52 cards when we draw the first card. We set the first card aside and assume that it is a face card. Then there are four aces out of the 51 remaining cards. We want to draw a face card and an ace so use multiplication.

$$P\left(\text{1st face card and 2nd ace}\right) = \frac{12}{52} \cdot \frac{4}{51} = \frac{48}{2652} \approx 0.018$$

The probability that the first card is a face card and the second card an ace is approximately 0.018 or 1.8%.

(c) The probability that one card is a heart and the other a club.

There are two ways for this to happen. We could get a heart first and a club second or we could get the club first and the heart second.

$$\begin{aligned} P(\text{heart and club}) &= P(\text{heart 1st and club 2nd or club 1st and heart 2nd}) \\ &= P(\text{heart 1st and club 2nd}) + P(\text{club 1st and heart 2nd}) \\ &= \frac{13}{52}\cdot\frac{13}{51}+\frac{13}{52}\cdot\frac{13}{51} \\ &\approx 0.127 \end{aligned}$$

The probability that one card is a heart and the other a club is approximately 0.127 or 12.7%.

Multiplication Rule for "And" Probabilities: Any Events

For events A and B, $P(A \text{ and } B) = P(A)\cdot P(B \mid A)$

Conditional Probability:

For events A and B, $P(B \mid A) = \dfrac{P(A \text{ and } B)}{P(A)}$

4.5. Binomial Distribution:

Suppose an experiment is repeated 'n' times and each trail is independent.

Let us assume that each trail results in two possible mutually exclusive and exhaustive outcomes i.e. success and failure.

Let X is random variable represents total no. of successes in ‘n’ trails. Let the probability of success in each trail is p and the probability of failure is q=1-p and p remains constant from trail to trail.

Now, we have to find out the probability of x successes in n trails.

Let us suppose that a particular order of outcomes of x successes in n repetitions be as follows

SSSSSFFFSSFS………FS (x number of successes and n-x failures)

Since, the trails are all independent the probability for the joint occurrence of the event is

P p p p p q q q p p q p……..q p

= (pppppp…..x times) (qqqqqq…… (n-x) times)

= $p^x q^{n-x}$

Further in a series of n trails x successes and n-x failures can occur in ${}^{n}c_{x}$ ways. So, the required probability is

Probability of x successes in n trails is

$$P(X=x) = {}^{n}c_{x}\ p^{x}q^{n-x}, \qquad x = 0,1,2,........,n$$

This is called probability distribution of Binomial random variable X or simply Binomial distribution. Symbolically this can be written as b(X; n, p)

A random variable X is said to be follow a binomial distribution if its probability mass function is given by

$$P(X=x) = {}^{n}c_x \; p^x q^{n-x}, \; x = 0,1,2,........,n$$

And p + q =1

Where n and p are called parameters of the binomial distribution

Properties: The sum of the probabilities of the binomial distribution is unity.

Proof: For a binomial distribution the probability function is given by

$$P(X=x) = {}^{n}c_x \; p^x q^{n-x}, \; x = 0,1,2,........,n$$

Now, $\sum_{x=0}^{n} p(X = x) = \sum_{x=0}^{n} {}^{n}c_x p^x q^{n-x}$

$$= {}^{n}c_0 \; p^0 q^{n-0} + {}^{n}c_1 \; p^1 q^{n-1} + {}^{n}c_2 \; p^2 q^{n-2} + \ldots\ldots\ldots + {}^{n}c_n \; p^n q^{n-n}$$

$$= (q+p)^n = 1 \qquad [q+p=1]$$

Mean of the binomial distribution:

For a binomial distribution the probability function is given by

$$P(X=x) = {}^{n}c_x \; p^x q^{n-x}, \; x = 0,1,2,........,n$$

Now, the mean of the Binomial distribution is

$$E(X) = \sum_{x=0}^{n} x \, P(X = x)$$

$$= \sum_{x=0}^{n} x\ {}^{n}c_x p^x q^{n-x}$$

$$= \sum_{x=0}^{n} x\ {}^{n}c_x p^x q^{n-x}$$

$$= \sum_{x=0}^{n} x \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} x \frac{n(n-1)!}{x(x-1)!(n-x)!} p\, p^{x-1} q^{n-x}$$

$$= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} q^{n-x}$$

$$= np \sum_{x=1}^{n} {}^{n-1}c_{x-1}\ p^{x-1} q^{n-x}$$

$$= np(q+p)^{n-1}$$

$$= np(1)^{n-1} \quad [\because\ q+p=1]$$

$$= np$$

The mean of the binomial distribution is np

Variance of the Binomial distribution:

The variance of the Binomial distribution is

$$V(X) = E(X^2) - [E(X)]^2$$

$$= E(X^2) - (np)^2 \text{ …………….. (A)} \qquad [E(X) = np]$$

Now,

$$E(X^2) = \sum_{x=0}^{n} x^2 \; {}^{n}c_x \, p^x q^{n-x}$$

$$= \sum_{x=0}^{n} [x(x-1)+x] \; {}^{n}c_x \, p^x q^{n-x}$$

$$= \sum_{x=0}^{n} x(x-1) \frac{n!}{x!(n-x)!} p^x q^{n-x} + \sum_{x=0}^{n} x \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} x(x-1) \frac{n(n-1)(n-2)!}{x(x-1)(x-2)!(n-x)!} p^2 \; p^{x-2} q^{n-x} + E(X)$$

$$= n(n-1)p^2 \sum_{x=2}^{n} \frac{(n-2)!}{(x-2)!(n-x)!} p^{x-2} q^{n-x} + np$$

$$= n(n-1)p^2 \sum_{x=2}^{n} \; {}^{n-2}c_{x-2} \; p^{x-2} q^{n-x} + np$$

$$= n(n-1)p^2 (q+p)^{n-2} + np$$

$$= n(n-1)p^2 \, (1)^{n-2} + np \qquad [\because \; q+p=1]$$

$$= n(n-1)p^2 + np \; \ldots\ldots\ldots\ldots \text{(B)}$$

Putting (B) in (A) we get

$$V(X) = n(n-1)p^2 + np - (np)^2$$

$$= \; np(np - p + 1 - np)$$

$$= \; np(1-p)$$

$$= npq$$

The variance of the Binomial distribution is npq

Remark: In Binomial distribution since mean = np and variance = npq and p + q = 1 therefore ***mean* > *variance***

Moments of the Binomial distribution

Non central moments (about zero):

$$\mu_1^{1} = E(X) = Mean = np$$

$$\mu_2^{1} = E(X^2) = n(n-1)p^2 + np$$

$$\mu_3^{1} = E(X^3) = n(n-1)(n-2)p^3 + 3n(n-1)p^2 + np$$

$$\mu_4^{1} = E(X^4) = n(n-1)(n-2)(n-3)p^4 + 6n(n-1)(n-2)p^3 + 7n(n-1)p^2 + np$$

Central moments (about mean):

$$\mu_1 = 0$$

$$\mu_2 = Variance = V(X) = E(X-np)^2 = npq$$

$$\mu_3 = E(X-np)^3 = npq(q-p)$$

$$\mu_4 = E(X-np)^4 = npq[1+3(n-2)pq]$$

Measure of skewness of Binomial distribution

$$=\sqrt{\beta_1} = \sqrt{\frac{{\mu_3}^2}{{\mu_2}^3}}$$

$$=\sqrt{\frac{n^2p^2q^2(q-p)^2}{n^3p^3q^3}}$$

$$=\sqrt{\frac{(q-p)^2}{npq}}$$

$$=\sqrt{\frac{(1-2p)^2}{npq}} \quad [\because q+p=1]$$

The Skewness of Binomial Distribution is $\sqrt{\frac{(1-2p)^2}{npq}}$

Remark: The Binomial distribution is called ***Symmetric Binomial Distribution*** if $p=\frac{1}{2}$

Kurtosis of Binomial Distribution $\beta_2 = \frac{\mu_4}{{\mu_2}^2}$

$$=\frac{npq[1+3(n-2)pq]}{n^2p^2q^2}$$

$$=\frac{[1+3(n-2)pq]}{npq}$$

$$=3+\frac{1-6pq}{npq}$$

Measure of Kurtosis of B.D is $3+\dfrac{1-6pq}{npq}$

Mode of the Binomial Distribution:

Case-1: The Binomial Distribution has unique mode if (n+1) p is not an integer and the value of mode is m, the integral part of (n+1) p

Case-2: The Binomial Distribution has two modes i.e. bimodal if (n+1)p is an integer and the modes are m and m-1

Properties of binomial distribution:

1) Binomial distribution is a discrete probability distribution with two parameters n and p and finite range from 0 to n
2) The mean and the variance of the **binomial distribution** are np and npq respectively and mean > variance
3) The measure of Skewness of B.D is $\sqrt{\beta_1}=\sqrt{\dfrac{(1-2p)^2}{npq}}$

If $p=\dfrac{1}{2}$, the distribution is symmetric

$p<\dfrac{1}{2}$, the distribution is positively skewed

$p>\dfrac{1}{2}$, the distribution is negatively skewed

4) The measure of Kurtosis of **binomial distribution** is $\beta_2 = 3 + \frac{1-6pq}{npq}$

If $pq = \frac{1}{6}$, the distribution is mesokurtic

$pq < \frac{1}{6}$, the distribution is leptokurtic

$pq > \frac{1}{6}$, the distribution is plattykurtic

and for a symmetric binomial distribution i.e. for $p = \frac{1}{2}$, the kurtosis of B.D is

$3 - \frac{2}{n}$

5) For $p = \frac{1}{2}$, the binomial distribution has maximum probability at $x = \frac{n}{2}$, if n is even and

$x = \frac{n-1}{2}$ and $x = \frac{n+1}{2}$, if n is odd

6) Under certain conditions the B.D approaches to Poisson and Normal distributions

4.6. Poisson distribution

Poisson distribution is a discrete probability distribution, which is the limiting case of the binomial distribution under certain conditions.

1. When n is very indefinitely very large
2. Probability of success is very small.
3. np = λ is finite, $\lambda \in R^+$

A discrete random variable X is said to be follow a Poisson distribution if the probability mass function is given by

$$p(X = x) = P(x;\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}, x = 0,1,2,3.........\infty$$

Where $\lambda > 0$

λ is called the ***parameter*** of the Poisson distribution.

Applications of Poisson distribution:

This distribution is used to describe the behavior of the rare events like

1. The number of blind born per year in a large city.
2. The number of printing mistakes per page in a large volume of a book.
3. The number of air pockets in a glass sheet.
4. The number of accidents occurred annually at a busy crossing of city.
5. The number of defective articles produced by a quality machine.
6. This is widely used in waiting lines or queuing problems in management studies.
7. It has wide applications in industrial quality control.
8. In determining the number of deaths in a given period by a rare disease.

Properties: The sum of the probabilities of the Poisson distribution is unity i.e.

$$\sum_{x=0}^{\infty} P(x;\lambda) = 1$$

Proof: For a Poisson distribution the probability mass function is given by

$$p(X = x) = P(x;\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}, x = 0,1,2,3..........\infty$$

Now, $\sum_{x=0}^{\infty} P(x;\lambda) = \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^{x}}{x!}$

$$= e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^{x}}{x!}$$

$$= e^{-\lambda}\left[\frac{\lambda^{0}}{0!}+\frac{\lambda^{1}}{1!}+\frac{\lambda^{2}}{2!}+\frac{\lambda^{3}}{3!}...........+\infty\right]$$

$$= e^{-\lambda}e^{\lambda}$$

=1

Mean and Variance of Poisson distribution:

For a Poisson distribution the probability mass function is given by

$$p(X = x) = P(x;\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}, x = 0,1,2,3..........\infty$$

Now,

Mean = E(X) = $\sum_{x=0}^{\infty} x\,P(x;\lambda)$

$$= \sum_{x=0}^{\infty} x\frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$= e^{-\lambda}\sum_{x=0}^{\infty} x\frac{\lambda\lambda^{x-1}}{x(x-1)!}$$

$$= e^{-\lambda}\lambda\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!}$$

$$= e^{-\lambda}\lambda\left[\frac{\lambda^0}{0!}+\frac{\lambda^1}{1!}+\frac{\lambda^2}{2!}+\frac{\lambda^3}{3!}..........+\infty\right]$$

$$= e^{-\lambda}\lambda e^{\lambda}$$

$$= \lambda$$

Mean of the Poisson distribution is λ

Variance = V(X) = $E(X^2)-[E(X)]^2$

$= E(X^2)-[\lambda]^2$ .... (3)

From equation number (3)

$$E(X^2) = \sum_{x=0}^{\infty} x^2\, P(x;\lambda)$$

$$= \sum_{x=0}^{\infty} x^2\,\frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \sum_{x=0}^{\infty}[x(x-1)+x]\frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \sum_{x=0}^{\infty}x(x-1)\frac{e^{-\lambda}\lambda^x}{x!} + \sum_{x=0}^{\infty}x\frac{e^{-\lambda}\lambda^x}{x!}$$

$= e^{-\lambda}\sum_{x=1}^{\infty}x(x-1)\frac{\lambda^2\lambda^{x-2}}{x(x-1)(x-2)!}+\lambda$ (Mean of the Poisson distribution is λ)

$$= e^{-\lambda}\lambda^2 \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} + \lambda$$

$$= e^{-\lambda}\lambda^2 \left[\frac{\lambda^0}{0!} + \frac{\lambda^1}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} .... + \infty \right] + \lambda$$

$$= e^{-\lambda}\lambda^2 e^{\lambda} + \lambda$$

$$E(X^2) = \lambda^2 + \lambda \; .... \qquad (4)$$

Putting (4) in (3) we get

$$V(X) = E(X^2) - [\lambda]^2$$

$$= \lambda^2 + \lambda - \lambda^2$$

$$= \lambda$$

Variance of the Poisson distribution is λ

Remark: In Poisson distribution the mean and variance are equal i.e. λ

Moments of the Poisson distribution:

Non central moments (about zero):

$$\mu_1^1 = Mean = E(X) = \lambda$$

$$\mu_2^1 = E(X^2) = \lambda^2 + \lambda$$

$$\mu_3^1 = E(X^3) = \lambda^3 + 3\lambda^2 + \lambda$$

$$\mu_4{}^1 = E(X^4) = \lambda^4 + 6\lambda^3 + 7\lambda^2 + \lambda$$

Central moments (about mean):

$$\mu_1 = 0$$

$$\mu_2 = Variance = \lambda$$

$$\mu_3 = \lambda$$

$$\mu_4 = 3\lambda^2 + \lambda$$

Skewness of Poisson Distribution:

Measure of Skewness of Poisson distribution is given by

$$\sqrt{\beta_1} = \sqrt{\frac{\mu_3{}^2}{\mu_2{}^3}} = \sqrt{\frac{\lambda^2}{\lambda^3}} = \sqrt{\frac{1}{\lambda}}$$

Kurtosis of Poisson distribution:

Measure of Kurtosis of Poisson distribution is given by

$$\beta_2 = \frac{\mu_4}{\mu_2{}^2} = \frac{3\lambda^2 + \lambda}{\lambda^2} = 3 + \frac{1}{\lambda}$$

Remark:

1) Since Skew ness = $\sqrt{\frac{1}{\lambda}}$ >0, Poisson distribution is always positively skewed.

2) Since Kurtosis = $3 + \frac{1}{\lambda}$ >3, Poisson distribution is always Leptokurtic

Property 1: Prove that the Poisson distribution is the limiting case of binomial distribution stating the required conditions**.**

Sol: The Poisson distribution can be limiting case of a binomial distribution under certain conditions.

1. Number of trails i.e. n is indefinitely large i.e. $n \to \infty$
2. p , the probability of success in each trail is indefinitely small i.e $p \to 0$
3. $np = \lambda$ is finite.

If X is a binomial variate then the probability mass function is given by

$$\text{P(X=x)} = {}^n c_x \; p^x q^{n-x}, \; x = 0,1,2,........,n$$

Under the above conditions

$$\lim_{n\to\infty} B(x;n,p) = \lim_{n\to\infty} {}^n c_x \, p^x q^{n-x}$$

$$= \lim_{n\to\infty} \frac{n!}{x!(n-x)!}\left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x} \quad [\because np = \lambda]$$

$$= \lim_{n\to\infty} \frac{n(n-1)(n-2)...........(n-x+1)(n-x)!}{x!(n-x)!}\left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$= \frac{\lambda^x}{x!} \lim_{n\to\infty} \frac{n^x\left[1\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right).................\left(1-\frac{x-1}{n}\right)\right]}{n^x} \frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^x}$$

$$= \frac{\lambda^x}{x!} \lim_{n\to\infty} 1 \, . \, \frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^x}$$

$$= \frac{\lambda^x}{x!} \frac{\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n}{\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^x} = \frac{\lambda^x}{x!}\frac{e^{-\lambda}}{1^x} = \frac{e^{-\lambda}\lambda^x}{x!}$$

Mode of the Poisson distribution:

Case: 1) When λ is not an integer, the distribution is uni model and the value of mode is the integral part of the λ .

Case: 2) When λ is an integer, the distribution is bimodal and the value of modes are λ and $\lambda - 1$.

Property 2: Show that in a Poisson distribution with unit mean, mean deviation is 2/e times the standard deviation**.**

Sol: The Poisson distribution with unit mean i.e. $\lambda = 1$ is given by

$$P(x;\, 1) = \frac{e^{-1}1^x}{x!} = \frac{e^{-1}}{x!}$$

Now, we have to show that

Mean deviation about mean = $\frac{2}{e}$ Standard deviation

$$E\left|X - E(X)\right| = \frac{2}{e} E(X^2) - [E(X)]^2$$

$$E\left|X-1\right| = \frac{2}{e}\sqrt{1} \qquad \text{[mean = 1]}$$

$$E\left|X-1\right| = \frac{2}{e}$$

Now,

$$E\left|X-1\right| = \sum_{x=0}^{\infty}\left|x-1\right|\frac{e^{-1}}{x!}$$

$$= \frac{1}{e}\left[1+\sum_{x=1}^{\infty}\frac{x-1}{x!}\right]$$

$$= \frac{1}{e}\left[1+\sum_{x=1}^{\infty}\frac{1}{(x-1)!}-\sum_{x=1}^{\infty}\frac{1}{x!}\right]$$

$$= \frac{1}{e}\left[1+e-(e-1)\right]$$

$$= \frac{2}{e}$$

Hence in Poisson distribution with unit mean, mean deviation are 2/e times the standard deviation.

Remark:

1. Poisson distribution is a discrete probability distribution with single parameter λ .
2. Both mean and variance of the Poisson distribution are equal to λ .
3. The distribution is positively skewed and leptokurtic.

4. It is asymptotic form of binomial distribution when p is small, n is large and np is finite.
5. the normal distribution is a limiting form of a Poisson distribution as $\lambda \to 0$
6. The distributio0n of rare events generally approximates to a Poisson distribution.
7. If X_1 and X_2 are two independent Poisson variates with mean λ_1 and λ_2 respectively, then $X = X_1 + X_2$ is also a Poisson variate with mean $\lambda_1 + \lambda_2$.

4.7.Normal Distribution

Normal distribution is an approximation of binomial distribution under certain conditions.

1. n, the number of trails is indefinitely large, i.e. $n \to \infty$
2. Neither p nor q is very small.

A continuous random variable X is said to have a normal distribution with parameters μ and σ^2 if its density function is given by the probability law

$$f(x/\mu,\sigma^2) = N(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2}\left\{\frac{x-\mu}{\sigma}\right\}^2\right]$$

Where $-\infty < X < \infty$

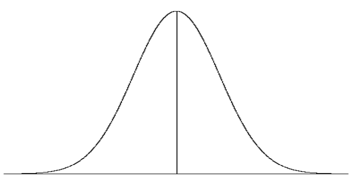
$$-\infty < \mu < \infty,\ \sigma > 0$$

Here μ and σ^2 are the mean and variance of the normal distribution respectively.

A random variable X with mean μ and variance σ^2 and following the normal law can be expressed by $X \sim N(\mu,\sigma^2)$

Properties of normal distribution:

The following points are important properties of normal distribution.



Normal curve

1. The normal curve is symmetrical and bell shaped.
2. The range of the distribution is $-\infty \; to \; \infty$
3. The value of mean, median, mode will coincide

as the distribution is symmetrical.

i.e. mean = median = mode

∞

4. The parameters μ and σ^2 represent the mean and variance of the distribution.
5. For different values of the parameters we get different normal distributions.
6. It has only one mode i.e. the distribution is unimodal and it occurs at $x = \mu$.
7. The skewness of the distribution is $\beta_1 = 0$ and kurtosis is $\beta_2 = 3$.
8. The odd ordered moments about mean vanishes i.e. $\mu_{2r+1} = 0$
9. The even ordered moments about mean $\mu_{2r} = (2r-1)\sigma^2 \mu_{2r-2}$.
10. The mean deviation from mean is $\sigma\sqrt{\frac{2}{\pi}} \cong \frac{4}{5}\sigma$.

Remark: The total area bounded by the curve and horizontal axis is equal to 1.

11. The maximum ordinate occurs at $x = \mu$ and its value is $\frac{1}{\sqrt{2\pi}\sigma}$.

12. The quartile deviation is $\frac{Q_3 - Q_1}{2} = 0.6745\sigma$

13. The first quartile $Q_1 = \mu$ -0.6745 σ and third quartile $Q_3 = \mu$ +0.6745 σ

14. The co-efficient of quartile deviation $\frac{Q_3 - Q_1}{Q_3 + Q_1} = 0.6745\frac{\sigma}{\mu}$.

15. A linear function $a_1X_1 + a_2X_2 + a_3X_3 + ... + a_nX_n$ of n independent normal variables $X_1, X_2, X_3, ..., X_n$ with means $\mu_1, \mu_2, \mu_3, ..., \mu_n$ and variances $\sigma_1{}^2, \sigma_2{}^2, \sigma_3{}^2, ..., \sigma_n{}^2$ is also a normal variable with mean $a_1\mu_1 + a_2\mu_2 + a_3\mu_3 + ... + a_n\mu_n$ and variance $a_1{}^2\sigma_1{}^2 + a_2{}^2\sigma_2{}^2 + a_3{}^2\sigma_3{}^2 + ... + a_n{}^2\sigma_n{}^2$.

Proof: Let Z= $a_1X_1 + a_2X_2 + a_3X_3 + ... + a_nX_n$

Now, $E(Z) = E[a_1X_1 + a_2X_2 + a_3X_3 + ... + a_nX_n]$

$$= a_1E[X_1] + a_2E[X_2] + a_3E[X_3] + ... + a_nE[X_n]$$

$$= a_1\mu_1 + a_2\mu_2 + a_3\mu_3 + ... + a_n\mu_n$$

And $V(Z) = V[a_1X_1 + a_2X_2 + a_3X_3 + ... + a_nX_n]$

$$= a_1{}^2V(X_1) + a_2{}^2V(X_2) + a_3{}^2V(X_3) + ... + a_n{}^2V(X_n)$$

$$= a_1{}^2\sigma_1{}^2 + a_2{}^2\sigma_2{}^2 + a_3{}^2\sigma_3{}^2 + ... + a_n{}^2\sigma_n{}^2$$

16. If X is a normal variate with mean μ and standard deviation σ, then the distribution o $Z = \frac{X - \mu}{\sigma}$ is also normal with mean 0 and variance 1. Here Z is called standard normal variable.

Symbolically if $X \sim N(\mu, \sigma^2)$ then $Z = \frac{X - \mu}{\sigma} \sim N(0,1)$

Proof: $E(Z) = E\left[\frac{X-\mu}{\sigma}\right] = \frac{1}{\sigma}\left[E(X)-\mu\right] = \frac{1}{\sigma}\left[\mu-\mu\right] = 0$

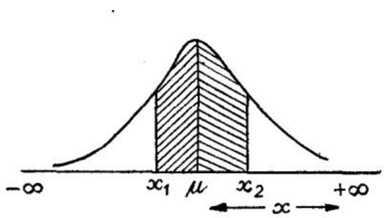
$$V(Z) = V\left[\frac{X-\mu}{\sigma}\right] = \frac{1}{\sigma^2}\left[V(X)\right] = \frac{1}{\sigma^2}\left[\sigma^2\right] = 1.$$

Remark: The probability density function of standard normal variable Z is

$$f(z/0,1) = N(0,1) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}z^2} , -\infty < Z < \infty$$

Area under normal curve:

As the normal variable is a continuous random variable, the probability that the random variable X assumes a value $x = x_1$ and $x = x_2$ is represented by the area under the probability curve bounded by the values x_1 *and* x_2 can be defined as



$$\Pr ob(x_1 < x < x_2) = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Since the normal curve depends on two parameters μ and σ^2, the area represented by $\Pr ob(x_1 < x < x_2)$ is also dependent on μ and σ^2. Though theoretically this probability can be calculated by using the method of integral calculus, normal integral tables are available for the use of practicing statisticians. It is very voluminous work to compile tables for all possible values of μ and σ^2. In fact such tables would be infinitely many because $-\infty < \mu < \infty$, $\sigma > 0$.

To facilitate the preparation of tables, the normal variable is standardized or is transformed to a new variable which is also normal, but having mean 0 and variance 1. Thus if X is normal variable with mean μ and variance σ^2, then $Z = \dfrac{X-\mu}{\sigma}$ is a standardized normal variable having mean 0 and variance 1

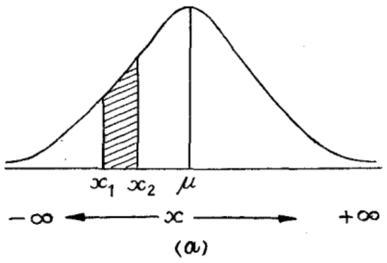
And thus $\Pr ob(x_1 < x < x_2) = \Pr ob\left(\dfrac{x_1-\mu}{\sigma} < \dfrac{x-\mu}{\sigma} < \dfrac{x_2-\mu}{\sigma}\right)$

$$= \Pr ob(z_1 < z < z_2)$$

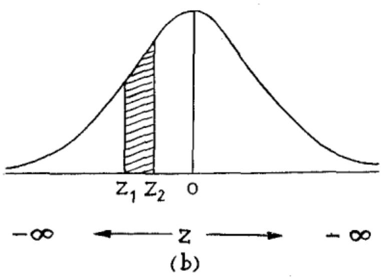
i.e.

$$\Pr ob(x_1 < x < x_2) = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$= \int_{z_1}^{z_2} \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}(z)^2} dz = \Pr ob(z_1 < z < z_2)$$



Original normal curve



Standardised normal curve

Cumulative

distribution functions of Z:

The cumulative distribution function of Z is defined as

$$\phi(t) = P(Z \le t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}(z)^2} dz$$

And

$$P(Z > t) = 1 - P(Z \le t)$$

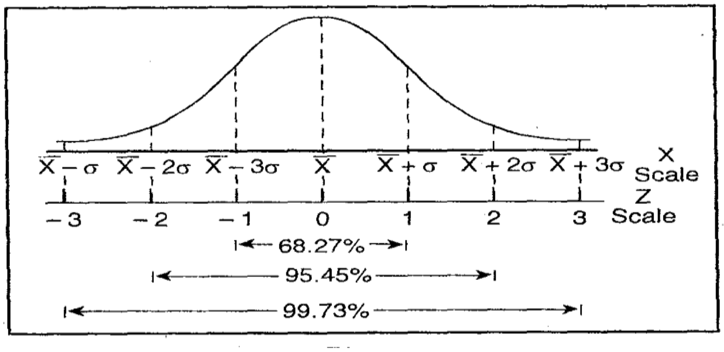
Remark: since the normal curve is symmetrical

$$P(Z > t) = P(Z < -t)$$

Some of the areas of standardized normal curve:

Distance from the mean ordinate	Area under curve
±0.6745	50%
±1	68.27%
±1.96	95%
±2	95.45%
±2.58	99%

±3	99.73%



<u>Applications of Probability</u>

Probability is the mathematical theory, which is used to describe and quantify uncertainty. Uncertainty can be due to our lack of knowledge, deliberate amalgamation, or due to the essential randomness of "Nature". In any case, we measure the uncertainty of events on a scale from zero (impossible events) to one (certain events or no uncertainty).

Probability axioms form the basis for mathematical probability theory. Probability applications include even more than Statistics, which is usually based on the idea of probability distributions. Probability theory plays a critical role in the development of statistical theory. Statistics is a branch of applied mathematics, which includes planning, summarizing, and interpreting uncertain observations. Whatever knowledge one has (about a process) is depicted mathematically and an attempt to learn more from whatever one can observe is made. This requires to:

a) Plan observations to control their variability;

b) Summarize a collection of observations to feature their community by suppressing details; and

c) Reach consensus about what the observations tells about the world under observation.

Probability distributions are very useful in understanding the bibliometric "laws" that have been used to characterize counts of document-related preferences.

It helps in understanding the phenomenon used in explaining the bibliometric "laws". Probability distributions also have applications in modeling of behavior that identify other factors influencing the formation of individual preference. It forms the basis of testing of hypotheses for validating the model identified for a particular process. It has deep applications in webometrics, which is the application of traditional bibliometric techniques in analyses of the structure of the World Wide Web. In short, probability forms the basis for many Statistics methods, which are used in Library and Information Science.

4.8. Summary:

The theory of probability is a study of random experiment. An event with probability of 1 can be considered a certain event and an event with a probability of 0 can be considered an impossible event. In this Unit you have learnt about finding the probability of occurrence of an event. You have learnt about the various definitions, needed to understand the concept of probability. You also learnt about the additive and multiplicative property of the probability. You have understood what is meant by the probability distribution of a random variable and you have also studied about the joint probability and margins probability distribution. Finally you have studied about some special distributions, which are of immense importance in statistics. You would appreciate that a strong base in probability would help in understanding further issues like testing of hypothesis and decision theory better.

4.9. Terminal questions:

Q.1: Two fair dice are rolled. Find the odds against rolling a sum of eight.

Answer:---

Q.2: A single card is drawn from a well-shuffled deck of 52 cards. Find the odds that the card is a nine.

Answer:---

Q.3: A single card is drawn from a well-shuffled deck of 52 cards. Find the odds against the card being a face card.

Answer:---

Q.4: A candy dish contains 12 chocolate candies, 18 butterscotch candies, 8 caramels, and 15 peppermints. A single candy is selected at random. What are the odds that the candy is a chocolate candy?

Answer:---

Q.5: An urn contains 10 red balls, 15 white balls, and 20 black balls. A single ball is selected at random. Find the odds against drawing a white ball.

Answer:---

Q.6: A real estate agent has kept records of the number of bedrooms in the houses he has sold for the last year. The data is listed in the following table.

Number of Bedrooms	1 or 2	3	4	5 or more
Number of Houses	5	12	25	3

If a house sold by the agent is selected at random, find:

a. the odds for the house having four bedrooms.

b. the odds for the house having less than three bedrooms.

c. the odds against the house having five or more bedrooms.

d. the odds against the house having more than three bedrooms.

Answer:--

4.10. Further Readings

1. A.M. Mood, F. Ar. Graybill & D.C. Boes. Introduction to the theory of Statistics, III. Editions Pub: Mac.Graw Hill.
2. Rahtagi V.K. (1984): An Introduction to Probability theory and Mathematical Statistics chapter VIII, IX & X Pub; John Wiley & Sons, New York.
3. Goon A.N., Gupta M.K. & Das Gupta B (1987) Fundamentals of Statistics Vol. I The World Press Pvt. Ltd., Kolkata.
4. Kapoor V.K. & S.C. Saxena: Fundamentals of Mathematical Statistics, Chapter Seventeen, Pub: S. Chand.

Unit-5: Theory of Hypotheses

Contents

5.1. Introduction

In this unit, we have discussed one part of statistical inference, that is, estimation and we have learnt how we estimate the unknown population parameter(s) by using point estimation and interval estimation. We will focus on the second part of statistical inference which is known as testing of hypothesis. Whenever you are trying to estimate the value of a parameter, we try to take a sample and we observe that the different samples give different values of the estimates and the value of the Estimates may or may not be the same as of the unknown value of the parameter in the Population. And the difference in the estimated values which are obtained from different samples May be due to some random variation. In simple words, we just want to take the closeness of estimated value with some hypothetical value and then what are we trying to do? We are trying to find the difference between the estimated value and the hypothetical value and if this difference is small, this is less then we can expect to accept it and would say that there is not much significant difference between the two values. Example, In our day-to-day life, we see different commercials advertisements in television, newspapers, magazines, etc. such as

(i) The television of certain brand saves up to 20% electric bill.

(ii) The car of certain brand gives 60 km/litter mileage.

(iii) A detergent of certain brand produces the cleanest wash, etc.

The Procedure of testing such type of claims or statements or assumptions is known as testing of hypothesis. The truth or falsity of a claim or statement is never known unless we examine the entire population. But practically it is not possible in mostly situations so we take a random sample from the population under study and use the information contained in this sample to take the decision whether a claim is true or false.

A customer of car wants to test whether the claim of car of certain brand gives the average mileage 60 km/liter is true or false. The decision maker is interested in making inference about the population parameter(s). However, he/she is not interested in estimating the value of parameter(s) but he/she is interested in testing a claim or statement or assumption about the value of population parameter(s). Such claim or statement is postulated in terms of hypothesis. In statistics, a hypothesis is a statement or a claim or an assumption about the value of a population parameter (e.g., mean, median, variance, proportion, etc.). Similarly, in case of two or more populations a hypothesis is comparative statement or a claim or an assumption about the values of population parameters. (e.g., means of two populations are equal, variance of one population is greater than other, etc.). The plural of hypothesis is hypotheses. Thus test is a two actions decision – rule where the actions are either to accept or to reject the hypothesisH_0. The truth or falsity of the statistical hypothesis H depends upon whether the information contained in the sample is consistent with or hypothesis. If the sample information is inconsistent with the hypothesis then the hypothesis is rejected; otherwise it is accepted**.**

Objectives

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

- Understand testing of hypothesis.
- Decide a null hypothesis and alternative hypothesis.
- Understand the meaning of level of significance, size of a test and power of a test.

- Understand the most powerful (MP) test and uniformly most powerful (UMP) test
- We shall define some terms associated with testing of hypothesis.

5.2. Hypothesis and its types

In hypothesis testing problems first of all we should being identifying the claim or statement or assumption or hypothesis to be tested and write it in the words. Once the claim has been identified then we write it in symbolical form if possible.

Statistical hypothesis is some assumption or statement about a population, or probability distribution characterizing the given population. It is frequently denoted by H.

A hypothesis is not accepted without being supported by evidence from the population. A hypothesis needs to be verified and is therefore, put to test; and based on the evidences provided by a random sample (which are set of independent observations) from the population a decisions is taken to accept or reject it. In fact the evidence provided by the sample is the value of a test statistic which will decide the course of the action – to accept or to reject H.

Definition: A test of a statistical hypothesis H is a rule or procedure, based on the observed values of random sample from the population to accept or reject the hypothesis H.

For example if a random variable X ~N (μ, 5) then the statement that the mean of the population is greater than 8, is a statement about the population mean with known variance σ^2 = 5 and therefore is a hypothesis. We write H: μ>8.

5.2.1. Simple Hypotheses

In general sense, if a hypothesis specifies only one value or exact value of the population parameter then it is known as simple hypothesis.

In case of normal population N (μ, σ^2) the hypothesis.

(a) $Ho: \mu = \mu_0, \quad \sigma^2 = \sigma o^2$

is a simple hypothesis. It specifies values to both parameter μ and $\sigma^2$and therefore, it completely specifies the distribution.

Examples, the hypothesis postulated μ = 40 is simple hypothesis because it gives a single value of parameter (μ = 40).

5.2.2. Composite Hypotheses

If a hypothesis specifies not just one value but a range of values that the population parameter may assume is called a composite hypothesis.

On the other hand each of the following hypotheses is composite hypothesis:

(ii) $Ho: \mu = \mu_0$ (No statement about σ^2)

(iii) H: $\sigma^2 = \sigma o^2 \quad (\mu\ is\ not\ specified)$

(i) H: $\mu < \mu_0, \quad \sigma^2 = \sigma o^2$

(ii) $H: \mu > \mu_0, \quad \sigma^2 = \sigma o^2$

(iii) $H: \mu = \mu_0, \quad \sigma^2 > \sigma o^2$

$and\ so\ on$

Examples, the hypothesis postulated μ > 40 is composite hypothesis because it does not specify the exact average value. It may be 60, 50, 100 or any other.

5.2.3. Null Hypotheses:

As we have discussed in hypothesis testing problems first of all we identify the claim or statement to be tested and write it in symbolical form. After that we write the complement or opposite of the claim or statement in symbolical form. In example, the claim is μ = 40 then its complement is μ ≠ 40 km. In (ii) the claim is μ > 40 then its complement is μ ≤ 40. If the claim is μ < 40 gm then its complement is μ ≥ 40 gm. The claim and its complement are formed in such a way that they cover all possibility of the value of population parameter.

Once the claim and its compliment have been established then we decide of these two which is the null hypothesis and which the alternative hypothesis is. The thump rule is that the statement containing equality is the null hypothesis. That is, the hypothesis which contains symbols = or ≤ or ≥ is taken as null hypothesis and the hypothesis which does not contain equality i.e. contains ≠ or < or > is taken as alternative hypothesis. The null hypothesis is denoted by H_0 and alternative hypothesis is denoted by H_1.

The hypothesis which we wish to test is called as the null hypothesis. According to Prof. R.A. Fisher, "**A null hypothesis is a hypothesis which is tested for possible rejection under the assumption that it is true."** The hypothesis which complements to the null hypothesis is called alternative hypothesis.

For example in sampling from normal population N (μ, σ^2) the hypothesisHo: $\mu = \mu_0$ is a null hypothesis if it is to be tested. It is said to be a null hypothesis since it states that there is no difference between Ho: $\mu\ and\ \mu o$.

5.2.4. Alternative Hypotheses:

It is very important to state the alternative hypothesis H_1 explicitly in respect to any null hypothesis Ho because the acceptance or rejection of Ho is meaningful only if it is being tested against the rival hypothesis H_1.

The concept of simple and composite hypothesis applies also to alternative hypothesis. For example in comparing the mean effect on the yield of soybean of two fertilizer say A and B, we may formulate the null and alternative hypothesis as Ho: $\mu A = \mu B$ against H_1: $\mu A < \mu B$. Ho is a simple hypothesis and H_1 is a composite hypothesis.

The alternative hypothesis has two types:

(i) Two-sided (tailed) alternative hypothesis.

(ii) One-sided (tailed) alternative hypothesis.

One-sided (tailed) Tests

We have seen that rejection (critical) region lies at one-tail or two-tails on the probability curve of sampling distribution of the test statistic its depend upon the form of alternative hypothesis. Similarly, the test of testing the null hypothesis also depends on the alternative hypothesis. A test of testing the null hypothesis is said to be two-tailed test if the alternative hypothesis is two-tailed whereas if the alternative hypothesis is one-tailed then a test of testing the null hypothesis is said to be one-tailed test. For example, if our null and alternative hypothesis are

H0:θ = θ0 and H1: θ ≠ θ0

Two-sided (tailed) Tests

Then the test for testing the null hypothesis is two-tailed test because the alternative hypothesis is two-tailed that means, the parameter θ can take value greater than θ0 or less than θ0.

If the null and alternative hypotheses are H0:θ <= θ0 and H1: θ > θ0

Then the test for testing the null hypothesis is right-tailed test because the alternative hypothesis is right-tailed. Similarly, if the null and alternative hypotheses are

H0:θ >= θ0 and H1: θ <θ0

Then the test for testing the null hypothesis is left-tailed test because the alternative hypothesis is left-tailed.

If we want to test the null hypothesis Ho that the population N (μ, σ^2)

Case1: specified mean μ_0 let be σ^2known then

H$_1$: $\mu = \mu_0$ is a simple hypothesis and all the possible alternative:

(i) $H_{11}: \mu \neq \mu_0\ (i.e. \mu > \mu_0\ or\ \mu < \mu_0)$

(ii) $H_{12}: \mu > \mu_0$

(iii) $H_{13}: \mu < \mu_0$

are composite hypothesis: while

(iv) $H_1: \mu = \mu_1$ is a simple hypothesis.

The normal population: X~ N (μ, σ^2)

Case 2: both $\mu \; and \; \sigma^2$ are unknown has parameter space.

$Q: \{(\mu, \sigma^2): -\infty < \mu < \infty, \sigma^2 > 0\}$. Here

The null hypothesis

$Ho: \mu = \mu_0, \sigma^2 > 0$

and alternative hypothesis

$Ho: \mu > \mu_0, \sigma^2 < 0$

and both composite hypothesis.

If $\sigma^2 = \sigma o^2$be known, then the null hypothesis

$Ho: \mu = \mu_0$

is a simple hypothesis.

The above definition can be symbolically written as follows:

For population: X~f (x;θ), $\theta \in \theta$

Ho: $\theta \in \theta_0 \subset \theta$

is simple if $\theta_0$is singleton set otherwise it is composite hypothesis. Similarly the alternative

H1: $\theta \in \theta \; where \; \theta_1 = \theta\,' - '\theta_0$

is simple if $\theta_1 = \theta\,' - '\theta_0$is a singleton set otherwise it is composite.

It is worth to note that sometimes one has to formulate the hypothesis exactly opposite of what is to be tested in the problem. For instance if it is required to show that the students of one school has a higher IQ than those of another school, i.e. $\mu_1 > \mu_2$. In this case also the null hypothesis must be Ho:$\mu_1 = \mu_2$ instead of H: $\mu_1 > \mu_2$we formulate the null hypothesis that there is “no difference” in the IQ’s of the two schools. Now a days the null hypothesis is being used to “any hypothesis we may want to test.”

5.3. Critical (Rejection) Region

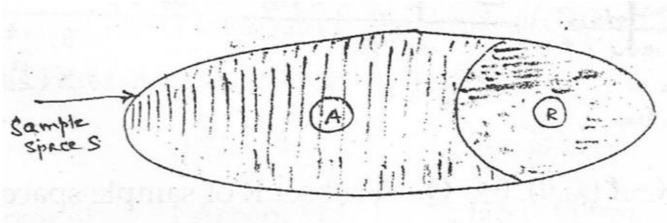
Let the population be X~ f (x; θ) $\theta \in Q$ where Q is the parameter space of the parameter θ .

Let $\underline{x}$: ($x_1, x_2 \dots \dots x_n$)be n independent sample observations corresponding to a random sample $\underline{X}$: ($X_1, X_2 \dots \dots X_n$)of size n from the population.

The n-dimensional space S which is the aggregate of all sample points $\underline{x}$: ($X_1, X_2 \dots \dots X_n$) is called a *sample space and* is denoted by S.

In order to test a hypothesis, the entire sample space is partitioned into two disjoint sub-spaces, say, u and S-u. If calculated value of the test statistic lies in u, then we reject the null hypothesis and if it lies in S-w, then we do not reject the null hypothesis. The region is called a “rejection region or critical region” and the region is called a “non-rejection region”. Therefore, we can say that “A region in the sample space in which if the calculated value of the test statistic lies, we reject the null hypothesis then it is called critical region or rejection region.

The test for a hypothesis divides the whole sample S into two disjoint (mutually exclusive) regions; one region A for acceptance of hypothesis H and another region R (or C) for rejection of hypothesis H.



A- *Figure 3.1 Sample Space, Acceptance (A) and Rejection (R) region*

B- Acceptance region (ACS)

R- Rejection (or critical) region (RCS)

With A∪R = S and A∩R = ∅

Thus the test for hypothesis H is;

Rejected Ho if ($x_1, x_2 \ldots \ldots x_n$) ∈ R

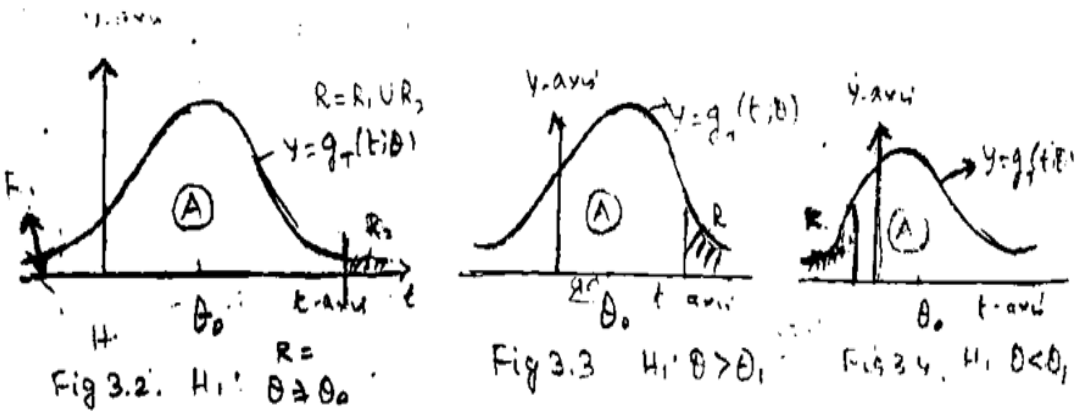
Accepted Ho if($x_1, x_2 \ldots \ldots x_n$) ∈ A

Definition 3.6: If a statistics T = T($X_1, X_2 \ldots \ldots X_n$) is used as an estimator for a parameter (-), then T=T ($\underline{X}$) is known as estimator of θ. if a statistic T is used to define a test of a hypothesis H, then it is known as a test statistic for H.

Thus a statistic associated with the test is called a test statistic.

A statistic R =T ($X_1, X_2 \ldots \ldots X_n$) condenses the experimental data $\underline{x}$: ($x_1, x_2 \ldots \ldots x_n$)to a point t-T($\underline{x}$) = T ($x_1, x_2 \ldots \ldots x_n$). In other words it maps the n-dimensional sample space S into a real line (one-dimensional) R_1: $(-\infty, \infty)$. There will be a region R and a region A on he real line R_1 corresponding to region R and region A, respectively in sample space S. Thus, a test γ partitions the real R_1 or the range of the test statistic T($\underline{x}$) into two disjoint sets: The acceptance region A and rejection Region R.

If g (t, θ) be the sampling distribution of the test statistic T($\underline{\text{X}}$) or test of hypothesis H, then we may get following types of rejection regions for H: $\theta = \theta_0$



Figure

Definition 3.7: Let X∼ f (x; θ) $\theta \in Q$. A subset R of sample space S, such that if R then Ho is rejected (with probability 1) is called the *critical region (or rejection region)* C of the test, where

$$C = \{\underline{x} \in S: H_0 \; is \; rejected \; if \; \underline{x} \in R\}$$

The complementary set A or R is said to *acceptance region* of the test.

5.4. Errors in Hypotheses

A test statistic is calculated on the basis of observed sample observations. But a sample is a small part of the population about which decision is to be taken. A random sample may or may not be a good representative of the population. A faulty sample misleads the inference (or conclusion) relating to the null hypothesis. For example, an engineer infers that a packet of screws is substandard when actually it is not. It is an error caused due to poor or inappropriate (faulty) sample. Similarly, a packet of screws may infer good when actually it is sub-standard.

Two kinds of Error

The decision of the test for hypothesis H is taken on the basis of the information of a sample from the population: X∼ f (x; θ) $\theta \in Q$. As such there is an element of risk – the risk of taking wrong decisions. In any test procedure, there are four possible mutually exclusive and exhaustive decisions:

(i) Reject Ho when actually Ho is not true (false)

(ii) Accept Ho when it is true

(iii) Reject Ho when it is true

(iv) Accept Ho when it is false

So we can commit two kinds of errors while testing a hypothesis

The decisions in (i) and (ii) are correct while the decisions (iii) and (iv) are wrong decisions. These decisions may be expressed in the following dichotomons table.

True state in The nature	Decision		
	Ho True	Wrong (Type I error)	Correct
	Ho False (H_1 True)	Correct	Wrong (Type II error)

Thus in testing hypothesis may lead to following two kinds of errors.

Definition: An error of type I is made if the null hypothesis Ho is rejected when Ho is true; and the error of Type II is made if the null hypothesis Ho is accepted when Ho is false.

Type I Error = [Reject Ho | Ho is true]

Type II Error = [Accept Ho | Ho is false]

= [Accept Ho | Ho is true]

= [Reject H_1 | Ho is true]

The probabilities of type I and Type II errors are denoted by α and β respectively.

Definition: The size of a Type I error is the probability of type I error **α** similarly, the size of a type II error is the probability of type II error **β.**

Thus,

α = P [Type I Error]

= Prob. [Reject H_0 |H_0 is True]

= Prob. [$\underline{x} \in R| \, Ho$] where $\underline{x} = (\, x_1, x_2 \ldots \ldots x_n)$

Similarly,

β = P [Type II Error]

= Prob. [Accept H_0 |H_1 is True]

= = Prob. [$\underline{x} \in A| \, H_1$] where$\underline{x} = (x_1, x_2 \ldots \ldots x_n)$.

Power of the test:

It is the probability with which the test reject H_0 when H_1 is true. It is denoted by 1- β where β is the probability of type II error.

Power of the test = 1- P [Type II Error]

= 1-β

= Prob. ($\underline{x} \in R|H_1$)

The power of the test provides a basis for the comparison of two or more tests for simple hypothesis Ho against the sample alternative H_1. The power function P() is a function P(θ) is a function of θ ; therefore the power curve will be the basis for comparison betweens tests for Ho Vs. H_1.

5.5. Level of Significance

We have discussed the hypothesis, types of hypothesis, critical region and types of errors. In this part, we shall discuss very useful concept “level of significance”, which play an important role in decision making while testing a hypothesis.

The probability of type-I error is known as level of significance of a test. It is also called the size of the test or size of critical region, denoted by α.

If calculated value of the test statistic lies in rejection (critical) region, then we reject the null hypothesis and if it lies in non-rejection region, then we do not reject the null hypothesis. Also we note that when H_0 is rejected then automatically the alternative hypothesis H_1 is accepted. Now, one point of our discussion is that how to decide critical value(s) or cut-off value(s) for a known test statistic.

The level of significance the maximum of probability of the type I error with which one is prepared to reject Ho when Ho is true. It is also called the size of critical region.

α = P [Type I Error]

= Prob. [Reject H_0 $|H_0$ is True]

= Prob. [$\underline{x} \in R|\ H_0$] where $\underline{x} = (\ x_1, x_2 \dots \dots x_n)$

Remarks:

(i) An idea test would be one for which both of the probabilities $\alpha\ and\ \beta$ are zero, but there exists no test with fixed sample size n for which both $\alpha\ and\ \beta$ are zero. Consequently for fixed sample size n it is not possible to minimize both the error simultaneously. In general type I error is supposed to be more serious than type II error.

(ii) Hence for a fixe sample size n the usual practice in testing of a hypothesis Ho against alternative H1 is to keep α at a pre-determined low level say 0.01 or 0.05 and the test which has a more power or lesser β is said to be better than the other one.

(iii) A level of significance $\alpha = .05$, implies that if a very large number of samples, each of size n, be taken from the population the event [T ($\underline{x}$) $\in R$] is observed then in about 5 out of 100 cases, the hypothesis Ho is rejected when Ho is true.

5.6. General Procedure of Testing a Hypothesis

Testing of hypothesis is a huge demanded statistical tool by many discipline and professionals. It is a step by step procedure as you will see in next three units through a large number of examples. The aim of this section is just give you flavor of that sequence which involves following steps:

First of all, we have to setup null hypothesis H_0 and alternative hypothesisH_1. Suppose, we want to test the hypothetical / claimed / assumed value θ0 of parameter θ. So we can take the null and alternative hypotheses as

H_0:$\theta = \theta_0$ andH_1: $\theta \neq \theta_0$ {for two tail test}

Or

H_0:$\theta <= \theta_0$ and H_1: $\theta > \theta_0$ {for one tail test}

H_0:$\theta >= \theta_0$ and H_1: $\theta < \theta_0$

In case of comparing same parameter of two populations of interest, say, θ1 and θ2, then our null and alternative hypotheses would be

$H_0\theta_1 = \theta_2$ and H_1: $\theta_1 \neq \theta_2$ {for two tail test}

Or

H_0: $\theta_1 <= \theta_2$ and H_1: $\theta_1 > \theta_2$ {for one tail test}

H_0: $\theta_1 >= \theta_2$ and H_1: $\theta_1 < \theta_2$θ2

After setting the null and alternative hypotheses, we establish criteria for rejection or non-rejection of null hypothesis, that is, decide the level of significance (α), at which we want to test our hypothesis. Generally, it is taken as 5% or 1% (α = 0.05 or 0.01).

We choose an appropriate test statistic under H0 for testing the null hypothesis. After that, specify the sampling distribution of the test statistic preferably in the standard form like Z (standard normal), χ2 , t, F or any other well-known.

Calculate the value of the test statistic on the basis of observed sample observations.

Obtain the critical (or cut-off) value(s) in the sampling distribution of the test statistic and construct rejection (critical) region of size α. Generally, critical values for various levels of significance are putted in the form of a table for various standard sampling distributions of test statistic such as Z-table, χ 2 -table, t-table, etc.

After that, compare the calculated value of test statistic obtained, with the critical value(s) obtained and locates the position of the calculated test statistic, that is, it lies in rejection region or non-rejection region.

In testing of hypothesis ultimately we have to reach at a conclusion. It is done as explained below

(i) If calculated value of test statistic lies in rejection region at α level of significance then we reject null hypothesis. It means that the sample data provide us sufficient evidence against the null hypothesis and there is a significant difference between hypothesized value and observed value of the parameter.

(ii) If calculated value of test statistic lies in non-rejection region at α level of significance then we do not reject null hypothesis. Its means that the sample data fails to provide us sufficient evidence against the null hypothesis and the difference between hypothesized value and observed value of the parameter due to fluctuation of sample.

5.7. Degree of freedom

The greatest number of logically independent values that can fluctuate in a data sample is known as the degree of freedom. To compute degrees of freedom, take the number of elements in the data sample and subtract one. The maximum number of logically independent values that can fluctuate in a data sample is referred to as degrees of freedom. To compute degrees of freedom, take the number of elements in the data sample and subtract one. The early 1800s saw the development of the notion of degrees of freedom thanks to the contributions of astronomer and mathematician Carl Friedrich Gauss. Degrees of freedom are frequently addressed in relation to chi-squares and other statistical hypothesis testing methods. Situations in business when management has to make a decision that affects another variable's result are referred to as having degrees of freedom. The formula to determine degrees of freedom is:

$$D_f \quad = \quad N - 1$$

where:

D_f= degrees of freedom

N=sample size

If consider the assignment of choosing ten baseball players with a batting average requirement of.250. The sample size, N = 10, is the total number of participants that will comprise our data set. In this example, 9 (10 - 1) baseball players can be randomly picked, with the 10th baseball player having a specific batting average to adhere to the .250 batting average limit

Examples of Degrees of Freedom

Example 1: Let us consider a sample of data that consists of five positive numbers. There must be an average of six among the five integer numbers. Ten must be the fifth number in the data set if the first four elements are {3, 8, 5, and 4}. Four degrees of freedom are available since the first four integers can be selected at random.

Example 2: Let us consider a sample of data that consists of five positive numbers. If there is no known correlation between the numbers, they could be any number. Four degrees of freedom exist because any one of the five can be selected at random and without restriction.

5.8. Test of Significance

Suppose that the problem is to test the hypothesis that the mean μ of the normal population $N(\mu, \sigma^2)$with known variance $\sigma^2$is different fromμ_0.

As explained above the null hypothesis Ho and the alternative hypothesis H$_1$ will be set up as follows:

$$H_o: \mu = \mu_o \; against \; alternative \; H_1 = \mu \neq \mu_o$$

Let us chosen level of the significant at α which is a smaller number.

Let X$_1$, X$_2$, X$_3$, ……….X$_n$ be a random sample of size n drawn from the population. Now the problem is to obtain a test for

$$H_o: \mu = \mu_o \qquad Vs. \quad H_1 = \mu \neq \mu_o$$

at level of significance α based on a sample random of size n. It is reasonable to accept Ho if the estimate $\hat{\mu}$of μ is close enough to μ_o.

We know that the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an estimator of population mean and has sampling distribution.

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

So that

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Further, 100 (1- α)% confidence internal for mean μ is

$$\bar{x} + z_0\left(\frac{\sigma}{\sqrt{n}}\right)$$

Where,

$$P[|Z| < z_0] = 1 - \alpha$$

Or

$$P(-z_0 < Z < z_0) = 1 - \alpha$$

The value of z_0 may be obtained from the Normal area Table for given α.

It is reasonable to take $\mu = \mu_o$if the estimate $\bar{x}\ of\ \mu$is close enough to μ_o . Obviously, it is reasonable accept Ho if lies in the interval $\bar{x} \pm z_0 \left(\frac{\sigma}{\sqrt{n}}\right)$

That is,

$$\bar{x} - z_0 \frac{\sigma}{\sqrt{n}} < \mu_o \leq \bar{x} + z_0 \frac{\sigma}{\sqrt{n}}$$

Or equivalently,

$$\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| \leq z_0$$

In this case we do not reject the hypothesis Ho: if, on the other hand, is not in the interval or equivalently.

$$\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| \geq z_0$$

Then reject Ho.

Our test for $H_o: \mu = \mu_o\ \ Vs.\ \ H_1 = \mu \neq \mu_o$ at the level of significance α is,

$$Reject\ Ho \qquad\qquad if\ |\bar{x} - \mu_0| > z_0 \left(\frac{\sigma}{\sqrt{n}}\right)$$

$$Accept\ \ Ho \qquad\qquad if\ |\bar{x} - \mu_0| \leq z_0 \left(\frac{\sigma}{\sqrt{n}}\right)$$

Where, $\frac{\sigma}{\sqrt{n}}$is the standard error of $\bar{X}$.

Prob [Type I error] = P [Reject Ho |Ho] =

$$P\left[\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right| \geq z_0 | Ho\right] = P\left[\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right| \leq z_0\right] = \alpha$$

Hence critical region R and Acceptance A are

$$A = \left\{(\, x_1, x_2 \dots \dots x_n): \mu_0 - Z_0 \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \ \mu_0 + Z\left(\frac{\sigma}{\sqrt{n}}\right)\right\}$$

and

$$R = \left\{(\, x_1, x_2 \dots \dots x_n): \left[\bar{x} < \mu_0 - \frac{Z}{\sqrt{n}}\right], U \ \left[\bar{x} + \ \mu\left(\frac{\sigma}{\sqrt{n}}\right)\right]\right\}$$

Such a test is known as test of significance.

Here the sample value of the statistic differ from the given value $\mu_0$of the parameter by more than certain amount in our case, Z$_0$SE ($\bar{X}$) is held important or significant to reject Ho at level of Significance α.

The value of |Z| = $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ under the assumption that $H_o: \mu = \mu_o$holds is

|Z| $\alpha - 1 \left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right| . if\ this\ value\ \left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right|\ exceeds\ tabulated\ value\ of Z_0\ of\ Z$

at level of significance α then we say that computed value of Z is significant at α and we reject the null hypothesis Ho.

It may worth to note that the rejection of a hypothesis Ho at α- level of significance does not implies the disapproval of the hypothesis. It only implies that the data or the sample information does not support the hypothesis at level of significance α. Similarly, the acceptance to Ho should be understood. The acceptance of the implies that x does not deviate from $\mu_o$by so much amount that we reject Ho. It does not imply that is actually equal to $\mu_o$but that it is close to μ_o .

Further the difference $|\bar{X} - \mu_0|$ between $\bar{x}$ and μ_o is inevitable produce of sampling fluctuations. The acceptance of Ho implies that this difference $|\bar{X} - \mu_0|$ is due to sampling fluctuations alone.

Definition: A test of significance for hypothesis Ho: is a procedure to assess the difference between the sample statistic and the value of parameter given by Ho or differences between two independent statistic to be significant or to reject or accept Ho at the given level of significance α.

We say that

(i) the difference between a statistic and the corresponding population parameters.

Or

(ii) the difference between two independent statistics

is not significant at the given level of significance, say α if it can be attributed only to the sampling fluctuations; otherwise it is said to be significant.

The procedure to be adopted for test of significance is outlined below-

(1) Propose the null hypothesis Ho and alternative hypothesis H_1:
(2) Fix a level of significance α for the test and a sample size n.
(3) Then choose a statistic T(x) whose sampling distribution is known under H_o.
(4) Keeping the value of α in mind decide upon those values of the test statistic (i.e. rejection region) that lead to its acceptance. In other words, define the test for H_o Vs. H_1 at level α.
(5) Now draw a random sample of size in from the population and compute the value of the test statistic.
(6) Finally on the basis of the value of the test statistic take the decision to accept or reject Ho.

Example: The mean of sample of size 25 from a normal population with mean μ and s.d. 4 is found to be 15. Do you accept or reject Ho: μ = 20 at the 10% level of significance?

Solution: Here $\bar{x}$= 15, n = 25, σ =4, α = 0.1

Since

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) = N\left(\mu, \frac{4^2}{25}\right)$$

So that

|Z| = $\left[\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right] \sim N\ (0,1)$

Under Ho: μ = 20

|Z| = $\left[\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right] = \frac{15-20}{4/\sqrt{25}} = -\frac{25}{4} = -6.25$

Case I : Let Ho: μ = 20 against H$_1$: $\mu \neq$ 20. We have to use a two tailed test. The test is as follows:

$$Reject\ Ho,\ \ if\ \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} < Z_{-\alpha/2} \quad or \quad \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} > Z_{\alpha/2}$$

$$Accept\ Ho, \qquad if - Z_{\alpha/2} < \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} > Z_{\alpha/2}$$

For α = 0.01, $\alpha/2$= 0.05, we get from normal areas table,

$$P\left[Z > Z_{\frac{\alpha}{2}}\right] = 1 - \frac{\alpha}{2} = 0.95$$

gives

$$Z_{-\alpha/2} = -1.645, \qquad Z_{\frac{\alpha}{2}} = +1.645$$

Test at α = 0.10 level of significance is

$$Reject\ Ho \quad if\ Z > -1.645 \quad or \quad Z < 1.645$$

$$Accept\ Ho \quad if \quad -1.645 \ < Z < \ 1.645$$

Here, computed Z = -6.25

Hence we reject Ho at 10% level of significance.

Case II Let Ho = μ = 20 be tested against H_1: Here right tailed test is to be used.

The test is

$$Reject\ Ho \quad if \qquad Z > Z_\alpha$$

$$Accept\ Ho \quad if \qquad Z \leq \ Z_{\alpha/2}$$

Where $Z_\alpha$is obtained by,

0.1= Prob. [Reject Ho|Ho|] = P [Z >Z_α|Ho]

$$= Prob.\left[\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > Z_\alpha\right]$$

Which gives

Z_α= 1.282

Observed Z= - 6.25

We conclude that the does not support Ho at 10% level of significance. You may try the following problems-

E- 3.1 Let population be $\bar{X} \sim N(\mu, \sigma^2)$. To test Ho: μ = .5 we take a random sample of size n= 17 and observe that $\bar{x}$ = 78.8 and S = 12.8. Do you accept or reject Ho at 5% level of significance.

[Ans. 1.19<2.1 accept/ Ho 94 and times]

Hint: Here $\sigma^2$is unknown $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{x})^2$is an unbiased estimator of σ^2therefore

$$T = \frac{\bar{X} - \mu}{\sqrt{S^2/(n/2)}} \sim student - test.$$

E- 3.2 Assume that IQ scores for a certain population are approximately $N(\mu, 100)$. To test Ho: $\mu = \mu_0$ against the one sided alternative hypothesis H$_1$: μ> 110, we take a random sample of size n = 16 from this population and observer $\bar{x}$ = 113.5 Do you accept or reject Ho at-

(a) 5% significance level

(b) 10% significance level,

Ans.: (a) 1.4< 1.1645 accept

5.8.1. *Z*-test for a population means (variance known)

To investigate the significance of the difference between an assumed population mean μ_0 and a sample mean $\bar{X}$.

Limitations for *Z*-test:

1. It is necessary that the population variance σ^2 is known. (If σ^2 is not known, see the *t*-test for a population mean)

2. The test is accurate if the population is normally distributed. If the population is not normal, the test will still give an approximate guide.

Test Procedure:

From a population with assumed mean μ_0 and known varianceσ^2, a random sample of size *n* is taken and the sample mean $\bar{X}$ calculated. The test statistic

$$Z = \left(\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \right)$$

may be compared with the standard normal distribution using either a one- or two-tailed test, with critical region of size α.

Z-test for two populations means (variances known and equal)

Investigate the significance of the difference between the means of two populations.

Limitations

1. Both populations must have equal variances and this variance σ_1 must be known.

(If σ_2 is not known, see the *t*-test for two population means)

2. The test is accurate if the populations are normally distributed. If not normal, the test may be regarded as approximate.

Method

Consider two populations with means $\mu1$ and $\mu2$. Independent random samples of size

n_1 and $n_2$are taken which give sample means X_1 and X_2.

The test statistic

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{\frac{1}{2}}}$$

may be compared with the standard normal distribution using either a one- or two-tailed test.

Z-test for two populations means (variances known and unequal)

Investigate the significance of the difference between the means of two populations.

Limitations

1. It is necessary that the two population variances be known. (If they are not known, see the *t*-test for two population means)

2. The test is accurate if the populations are normally distributed. If not normal, the test may be regarded as approximate.

Method

Consider two populations with means μ_1 and μ_2 and variances ${\sigma_1}^2$and ${\sigma_2}^2$Independent random samples of size n_1 and n_2 are taken and sample means X_1 and X_2 are calculated.

The test statistic

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\left(\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}\right)^{\frac{1}{2}}}$$

may be compared with the standard normal distribution using either a one- or two-tailed test.

***t*-Test for two population means (variances unknown but equal)**

Investigate the significance of the difference between the means of two populations.

Limitations

1. If the variance of the populations is known, a more powerful test is available: the *Z*-test for two population means).

2. The test is accurate if the populations are normally distributed. If the populations are not normal, the test will give an approximate guide.

Method

Consider two populations with means $\mu_1$and μ_1 Independent random samples of size

n_1 and n_2 are taken from which sample means $\bar{x}_1$ and $\bar{x}_2$together with sums of squares

$${s_1}^2 = \sum_{i=1}^{n_1} (X_i - \bar{X}_1)^2$$

And

$$s_2{}^2 = \sum_{i=1}^{n_2}(X_i - \bar{X}_2)^2$$

are calculated.

The best estimate of the population variance is found as

$$s^2 = \frac{[(n_1 - 1)s_1{}^2 + (n_2 - 1)s_2{}^2]}{(n_1 + n_2 - 2)}$$

The test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{\frac{1}{2}}}$$

which may be compared with Student's *t*-distribution with $n_1 + n_2 - 2$ degrees of freedom. The test may be either one-tailed or two-tailed.

5.8.2. . t-Test for two population means (variances unknown and unequal)

The sample's t value must be determined and compared to a crucial value in order to conduct a t-test. A data set's t distribution with the degrees of freedom can be used to calculate the accurate critical value, which varies. Investigate the significance of the difference between the means of two populations.

Limitations

1. If the variances of the populations are known, a more powerful test is available: the

Z-test for two population means).

2. The test is approximate if the populations are normally distributed or if the sample sizes are sufficiently large.

3. The test should only be used to test the hypothesis $\mu_1 = \mu_2$

Method

Consider two populations with means μ_1 and μ_2 Independent random samples of size

n_1 and n_2 are taken from which sample means $\bar{x}_1$and $\bar{x}_2$and variances

$$s_1{}^2 = \frac{\sum_{i=1}^{n_1}(X_i - \bar{X}_1)^2}{n_1 - 1}$$

$$s_1{}^2 = \frac{\sum_{i=1}^{n_1}(X_i - \bar{X}_1)^2}{n_1 - 1}$$

are calculated. The test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\left(\frac{s_1{}^2}{n_1} + \frac{s_1{}^2}{n_2}\right)^{\frac{1}{2}}}$$

which may be compared with Student's *t*-distribution with degrees of freedom given

by

$$v = \left\{\frac{\left(\frac{s_1{}^2}{n_1} + \frac{s_2{}^2}{n_2}\right)}{\frac{s_1{}^4}{n_1{}^2(n_1 - 1)} + \frac{s_2{}^4}{n_2{}^2(n_2 - 1)}}\right\}$$

***t*-test for two population means (method of paired comparisons)**

Object

To investigate the significance of the difference between two population means, μ_1 and

μ_2 No assumption is made about the population variances.

Limitations

1. The observations for the two samples must be obtained in pairs. Apart from population differences, the observations in each pair should be carried out under identical, or almost identical, conditions.

2. The test is accurate if the populations are normally distributed. If not normal, the test may be regarded as approximate.

Method

The differences d_i are formed for each pair of observations. If there are n such pairs of observations, we can calculate the variance of the differences by

$$s^2 = \frac{\sum_{i=1}^{n} \left(d_i - \bar{d}\right)^2}{n - 1}$$

Let the means of the samples from the two populations be denoted by .*x*1 and .*x*2. Then the test statistic becomes

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{\frac{s}{n^{\frac{1}{2}}}}$$

which follows Student's *t*-distribution with $n - 1$ degrees of freedom. The test may be either one-tailed or two-tailed.

Z-test of a correlation coefficient

To investigate the significance of the difference between a correlation coefficient and a specified value ρ_0.

Limitations

1. The *x* and *y* values originate from normal distributions.

2. The variance in the *y* values is independent of the *x* values.

3. The relationship is linear.

When these conditions cannot be met, the user should turn to the Kendall rank correlation test.

Method

With *r* as defined in the *t*-test of a correlation coefficient, using the Fisher *Z*-transformation we have

$$Z_1 = \frac{1}{2} log_e \left(\frac{1+r}{1-r}\right) = 1.1513 log_{10} \left(\frac{1+r}{1-r}\right)$$

The distribution of *Z*1 is approximately normal with mean μZ_1 and standard deviation

σZ_1here

.

$$\mu Z_1 = \frac{1}{2} log_e \left(\frac{1+\rho_0}{1-\rho_0}\right) = 1.1513 log_{10} \left(\frac{1+r}{1-r}\right)$$

$$\rho Z_1 = \frac{1}{\sqrt{n-3}}$$

The test statistic is now

$$Z = \frac{Z_1 - \mu Z_1}{\rho Z_1}$$

***Z*-test for two correlation coefficients**

To investigate the significance of the difference between the correlation coefficients for a pair of variables occurring from two different samples and the difference between two specified values ρ_1 and ρ_2.

Limitations

1. The *x* and *y* values originate from normal distributions.

2. The variance in the *y* values is independent of the *x* values.

3. The relationships are linear.

Method

Using the notation of the *Z*-test of a correlation coefficient, we form for the first sample

$$Z_1 = \frac{1}{2} log_e \left(\frac{1 + r_1}{1 - r_1}\right) = 1.1513 log_{10} \left(\frac{1 + r_1}{1 - r_1}\right)$$

which has mean

$$\mu Z_1 = \frac{1}{2} log_e \left(\frac{1 + \rho_1}{1 - \rho_1}\right)$$

And varience

$$\rho Z_1 = \frac{1}{\sqrt{n_1 - 3}}$$

where n_1 is the size of the first sample; $Z_2$is determined in a similar manner. The test statistic is now

$$Z = \frac{(Z_1 - Z_2) - (\mu Z_1 - \mu Z_2)}{\sigma}$$

where$\sigma = \left({\sigma_{Z_1}}^2 + {\sigma_{Z_2}}^2\right)^{\frac{1}{2}}$.

Z is normally distributed with mean 0 and with variance 1.

5.8.3. F-test for two population variances (variance ratio test)

Object

investigate the significance of the difference between two population variances.

Limitations

The two populations should both follow normal distributions. (It is not necessary that they should have the same means.)

Method

Given samples of size $n_1$with values x_1 x_2 ……..x_n and size $n_2$with values y_1 y_2 ……..y_n the two populations, the values of

$$\bar{x} = \frac{\sum x_i}{n_1}$$

$$\bar{y} = \frac{\sum y_i}{n_2}$$

And

$${s_1}^2 = \frac{\sum (x_i - \bar{x})^2}{n_1 - 1}$$

$${s_2}^2 = \frac{\sum (y_i - \bar{y})^2}{n_2 - 1}$$

can be calculated. Under the null hypothesis that the variances of the two populations are equal the test statistic

$$F = \frac{{s_1}^2}{{s_2}^2}$$

follows the F-distribution with $(n_1 - 1, n_2 - 1)$ degrees of freedom. The test may be either one-tailed or two-tailed.

5.8.4. P Value

The P value is the likelihood of getting a result that is as extreme as or more extreme than what was actually observed, assuming that there is no impact or difference (null hypothesis). P, or probability, expresses the likelihood that any observed variation across groups is the result of chance. The least significance at which the null hypothesis would be rejected is shown by the P-value, which is used as an alternative to the rejection point. The alternative hypothesis is more strongly supported if the P-value is modest. The P-value table shows the hypothesis interpretations:

P-value	Decision
P-value > 0.05	The result is not statistically significant and hence don't reject the null hypothesis.
P-value < 0.05	The result is statistically significant. Generally, reject the null hypothesis in favour of the alternative hypothesis.
P-value < 0.01	The result is highly statistically significant, and thus rejects the null hypothesis in favour of the alternative hypothesis.

We are aware that the P-value is a statistical indicator that may be used to assess the validity of a hypothesis. A P-value is a value in the range of 0 and 1. The researcher should establish the predetermined threshold known as the level of significance, or α. usually, it is set to 0.05. The formula used to determine the P-value is

$$z = \frac{\hat{p} - p0}{\sqrt{\frac{po(1 - p0)}{n}}}$$

Where

$$\hat{p} = \text{Sample Proportion}$$

P0 = assumed population proportion in the null hypothesis

N = sample size

5.8.5. χ2-test for a population variance

The Greek letter chi, or χ2, which is used to denote this statistic in formulas and outcomes, is the name of the statistical test known as chi-square. The standard test of the null hypothesis for discrepancies between the observed and anticipated frequencies is the chi-square test. It is frequently necessary to compare the distribution of a categorical variable in one sample with that of another sample. To investigate the difference between a sample variance s^2 and an assumed population variance ${\sigma_0}^2$

Limitations

It is assumed that the population from which the sample is drawn follows a normal distribution.

Method

Given a sample of n values x_1 x_2 ……..x_n the values of

$$\bar{x} = \frac{\sum x_i}{n}$$

And $s^2 = \sum(x_i - \bar{x})^2$

are calculated. To test the null hypothesis that the population variance is equal to ${\sigma_0}^2$

The test statistic

$$\frac{(n-1)s^2}{{\sigma_0}^2}$$

will follow a χ2-distributkm with n−1 degrees of freedom.

The test may be either one-tailed or two-tailed.

A chi-square test can be classified into two types: an independence test, which poses a relationship-related question like, "Is there a relationship between gender and SAT scores?"; and the goodness-of-fit test, which poses a question similar to "Will a coin come up heads 50 times and tails 50 times if it is tossed 100 times?"

Degrees of freedom are used in these tests to assess if the overall number of variables and samples in the experiment allows for the rejection of a null hypothesis. For instance, a sample size of 30 or 40 students is probably not big enough to produce meaningful data when looking at students and course choice. It is more legitimate to obtain identical or comparable results from a study with a sample size of 400 or 500 pupils.

5.9. Summary

In this unit an attempt is made to explain the basis concepts related to the testing of hypotheses.

5.10.Terminal questions

Q.1. In a rat feeding experiment the following results were obtained gain inweight in gm.

High Protein	13	14	10	11	12	14	10	8	11	12	9	12
Low Protein	7	11	10	8	10	12	9					

Find if there is any evidence of superiority of one diet over the other.

Given t (5%) on 17 d.f. is 2.11

Answer:---

--

Q.2. A random sample of 900 members is found to have a menu of 3.4cms. Could it be regarded as a sample from a large population with mean 3.25cms? and s.d. 2.61 cm. at 5% level of significance.

Answer:--

Q.3. The mean of sample of size 25 from a normal population with mean μ and s.d. 4 is found to be Do you accept or reject Ho: μ = 20 at the 10%level of significance?

Answer:--

Q.4. What does a smaller P-value represent?

Answer:--

Q.5. What is chi-square χ2 and how is it used?

Answer:--

5.11. Further Readings

5. A.M. Mood, F. Ar. Graybill & D.C. Boes. Introduction to the theory of Statistics, III. Editions Pub: Mac.Graw Hill.

6. Rahtagi V.K. (1984): An Introduction to Probability theory and Mathematical Statistics chapter VIII, IX & X Pub; John Wiley & Sons, New York.

7. Goon A.N., Gupta M.K. & Das Gupta B (1987) Fundamentals of Statistics Vol. I The World Press Pvt. Ltd., Kolkata.

8. Kapoor V.K. & S.C. Saxena: Fundamentals of Mathematical Statistics, Chapter Seventeen, Pub: S. Chand.

Unit-6: ANOVA (Analysis Of Variance)

Contents

6.1. Introduction

Test of significance based on *t*- distribution is an adequate procedure only for testing the significance of the difference between two sample means. In a situation when we have three or more samples to consider at a time, an alternative procedure is needed for testing the hypothesis that all the samples are drawn from the same population. For example when 5 different fertilizers are applied to four plots each, then we may be interested in finding whether the fertilizers have any significant effect on the yield. In other words, we want to see whether the samples are coming from the same normal population.

In any set of observations, the variation is inherent in nature. The total variation in any set of numerical data is due to a number of causes, but mainly classified as

(i) The *assignable cause*

(ii) The *chance cause*.

Assignable cause of variation*:* The assignable cause of variation can be identified, measured and controlled.

Chance cause of variation*:* The chance causes of variations are beyond the control of human hand and cannot be traced separately.

Analysis of Variance consists of estimation of the amount of effects due to each of independent factors (causes) separately and compares the estimates of effects due to assignable factors (causes) with estimates of the effects due to chance factor (cause) or experimental error or simple error.

<u>Assumptions for ANOVA</u>

The following assumptions are made in any analysis of variance procedure.

(1) The observations are independent.

(2) Parent population from which observations are taken is normal; and

(3) Various treatment and environmental effects are additive in nature.

Objectives:

After reading this unit, you should be able to

- identify different sources of variation in a given problem;
- distinguish the one-way classification data from other types;
- write down the model for one-way classification problems;
- carry out the test of hypothesis on equality of all treatment means, equality of pairs of treatments means;
- Estimate the model parameters
- Draw inferences and conclusions from the analysis.

6. 2. Analysis of Variance of a One-way layout with fixed effects model

Suppose there are n observationsy_{ij}, $(i = 1,2,\cdots,k; j = 1,2,\cdots,n_i)$ of a random variable Y are grouped into k groups of size $n_1, n_2, \cdots, n_k$ respectively. Then $n = \sum_{i=1}^{k} n_i$ and the observation table is as follows.

Groups	Observations	Total	Mean
1	y_{11} y_{12} $\cdots$ y_{1n_1}	$T_{1.} = \sum_{j=1}^{n_1} y_{1j}$	$\bar{y}_{1.} = \frac{T_{1.}}{n_1}$
2	y_{21} y_{22} $\cdots$ y_{2n_2}	$T_{2.} = \sum_{j=1}^{n_2} y_{2j}$	$\bar{y}_{2.} = \frac{T_{2.}}{n_2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	y_{i1} y_{i2} $\cdots$ y_{in_i}	$T_{i.} = \sum_{j=1}^{n_i} y_{ij}$	$\bar{y}_{i.} = \frac{T_{i.}}{n_i}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	y_{k1} y_{k2} $\cdots$ y_{kn_k}	$T_{k.} = \sum_{j=1}^{n_k} y_{kj}$	$\bar{y}_{k.} = \frac{T_{k.}}{n_k}$
Total		$T_{..} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}$	$\bar{y}_{..} = \frac{T_{..}}{n}$

6. 2.1 Statistical Analysis of One-way classification:

The total variation in the observation can be split into the following two components.

(I) The variation between the classes or *assignable cause* of variation and

(II) The variation within the classes or *chance cause* of variation.

Hence the mathematical model is given by

$y_{ij} = \mu_i + e_{ij}; j = 1,2,\cdots,n_i, i = 1,2,\cdots,k,$

where μ_i is the average effect of the i^{th}group, which can be split as

$\mu_i = \mu + \mu_i - \mu = \mu + \alpha_i$ with $\alpha_i = \mu_i - \mu$, $i = 1,2,\cdots,k$ and $\mu = \frac{1}{n}\sum_{i=1}^{k} n_i\mu_i$.

Hence,

$y_{ij} = \mu + \alpha_i + e_{ij}; j = 1,2,\cdots,n_i, i = 1,2,\cdots,k;$

(1)

where y_{ij} is the j^{th} observation of i^{th} class; $j = 1,2,\cdots,n_i, i = 1,2,\cdots,k,$

μ is the general mean effect,

α_i is the additive effect due to i^{th}group and

e_{ij} is the error effect due to chance and these are assumed to be *iid* random variables

each following $N(0,\sigma_e^2); j = 1,2,\cdots,n_i, i = 1,2,\cdots,k.$

The side condition is $\sum_{i=1}^{k} n_i\alpha_i = \sum_{i=1}^{k} n_i(\mu_i - \mu) = n\mu - n\mu = 0$.

6. 2.2 Assumptions

The statistical analysis of this layout is based on the following assumptions.

(1) *All the observations are mutually independent.*

(2) *Different effects are additive in nature.*

(3) e_{ij}*'s are iid random variables each following* $N(0,\sigma_e^2)$*;* $j = 1,2,\cdots,n_i, i = 1,2,\cdots,k.$

The null hypothesis to be tested is ***H₀***: The groups do not differ significantly or there is no additive effect due to different groups. In other words,

$\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0.$

Summing (1) over j and dividing by n_i, we get

$\bar{y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} = \mu + \alpha_i + \bar{e}_{i.}, \forall\, i = 1,2,\cdots,k,$

(2)

where $\bar{e}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} e_{ij}$ are *iid* random variables each distributed as $N(0,\sigma_e^2/n_i)$.

Summing (1) over i and j and dividing by n, we get

$\bar{y}_{..} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \mu + \bar{e}_{..} = \mu + \bar{e}_{..},$

(3)

where $\bar{e}_{..} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}$ are *iid* random variables each distributed as $N(0,\sigma_e^2/n)$.

Now the total variation in each observation is given by the total sum of squares as

$T.S.S. = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{..}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\bar{y}_{i.} - \bar{y}_{..} + y_{ij} - \bar{y}_{i.}\right)^2$

$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2$

$= \sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2.$

Or *T.S.S. = S.S.G. + S.S.E,*

where *T.S.S* = Total sum of squares = $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{..}\right)^2$;*S.S.G* = Sum of squares due to groups = $\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$; and *S.S.E* = Sum of squares due to error or residuals = $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2$.

6.2.3 Degrees of freedom for various sums of squares

T.S.S = Total sum of squares = $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{..}\right)^2$ is computed from n quantities of the form $\left(y_{ij} - \bar{y}_{..}\right)$ with one constraint $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{..}\right) = 0$. Hence, *T.S.S* will have $n - 1$ degrees of freedom.

S.S.G = Sum of squares due to groups = $\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$ is computed from k quantities of the form $(\bar{y}_{i.} - \bar{y}_{..})$ with one constraint $\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..}) = 0$. Hence, *S.S.G* will have $k - 1$ degrees of freedom.

S.S.E = Sum of squares due to error or residuals = $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2$ is computed from n quantities of the form $\left(y_{ij} - \bar{y}_{i.}\right)$ with k constraints $\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right) = 0$ for all $i = 1,2,\cdots,k.$. Hence, *S.S.E* will have $n - k$ degrees of freedom.

6.2.4 Mean Sum of squares

The sum of squares divided by its degrees of freedom gives the corresponding mean sum of squares. Thus,

Mean sum of squares due to groups = *M.S.G.*= $\frac{S.S.G.}{k-1}$.

Mean sum of squares due to error = *M.S.E.*= $\frac{S.S.E.}{n-k}$

6.2.5 Least square estimates

In the mathematical model (1), μ and $\alpha_i, i = 1,2,\cdots,k$ are the unknown parameters which have to be estimated by the principle of least squares. Hence, we consider the sum of squares due to errors, which is given by

$$S.S.E = \sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \mu - \alpha_i\right)^2.$$

(4)

Differentiating (4) with respect to μ and α_i and equating to zero individually, we get

$$\frac{dS.S.E}{d\mu} = 0 \Rightarrow -2\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \mu - \alpha_i\right) = 0$$

$$\Rightarrow \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \mu - \alpha_i\right) = 0$$

$\Rightarrow \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = n\mu + \sum_{i=1}^{k} n_i\alpha_i = n\mu$ [$\because \sum_{i=1}^{k} n_i\alpha_i = 0$ by side condition.]

Hence, the estimate of μ is given by

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \bar{y}_{..}.$$

$$\frac{dS.S.E}{d\alpha_i} = 0 \Rightarrow -2\sum_{j=1}^{n_i}\left(y_{ij} - \mu - \alpha_i\right) = 0,\ i = 1,2,\cdots,k.$$

$$\Rightarrow \sum_{j=1}^{n_i}\left(y_{ij} - \mu - \alpha_i\right) = 0$$

$$\Rightarrow \sum_{j=1}^{n_i} y_{ij} = n_i\mu + n_i\alpha_i$$

$$\Rightarrow \hat{\alpha}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} - \hat{\mu} = \bar{y}_{i.} - \bar{y}_{..}.$$

6.2.6 Variance of the estimates

We have $\hat{\mu} = \bar{y}_{..}$ and $\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$.

$$V(\hat{\mu}\,) = E[\bar{y}_{..} - E(\bar{y}_{..} - \alpha_0)]^2 = E[\mu + \bar{e}_{..} - \mu]^2 = E[\bar{e}_{..}]^2 = E(\bar{e}_{..}^2)$$

$$= V(\bar{e}_{..}) = \frac{\sigma_e^2}{n}.$$

Also we have $\hat{\alpha}_i - E(\hat{\alpha}_i) = \bar{y}_{i.} - \bar{y}_{..} - E(\bar{y}_{i.} - \bar{y}_{..})$

$$= \mu + \alpha_i + \bar{e}_{i.} - \mu - \bar{e}_{..} - E(\mu + \alpha_i + \bar{e}_{i.} - \mu - \bar{e}_{..})$$

$$= \mu + \alpha_i + \bar{e}_{i.} - \mu - \bar{e}_{..} - \alpha_i$$

$$= \bar{e}_{i.} - \bar{e}_{...}$$

Hence, $V(\hat{\alpha}_i\,) = E[\bar{e}_{i.} - \bar{e}_{..}]^2 = E[\bar{e}_{i.}^2 + \bar{e}_{..}^2 - 2\bar{e}_{i.}\bar{e}_{..}]$

$$= E(\bar{e}_{i.}^2) + \; E(\bar{e}_{..}^2) - \; 2E(\bar{e}_{i.}\bar{e}_{..}).$$

Now, since $\bar{e}_{..} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij} = \frac{1}{n}\sum_{i=1}^{k} n_i\bar{e}_{i.}$ [as $\bar{e}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} e_{ij} \Rightarrow \sum_{j=1}^{n_i} e_{ij} = n_i\bar{e}_{i.}$]

$$E(\bar{e}_{i.}\bar{e}_{..}) = E(\bar{e}_{i.}\frac{1}{n}\sum_{i=1}^{k} n_i\bar{e}_{i.})$$

$$= \frac{1}{n}E[\bar{e}_{i.}(n_1\bar{e}_{1.} + \; \cdots \; + n_i\bar{e}_{i.} + \cdots + n_k\bar{e}_{k.})]$$

$$= \frac{1}{n}E[n_1\bar{e}_{i.}\bar{e}_{1.} + \cdots + \; n_i\bar{e}_{i.}^2 + \; \cdots + \; n_k\bar{e}_{i.}\bar{e}_{k.}]$$

$= \frac{n_i}{n}E(\bar{e}_{i.}^2)$ since $E(\bar{e}_{i.}\bar{e}_{h.}) = 0$ for $h \neq i$;

$$= \frac{n_i}{n}V(\bar{e}_{i.}) = \frac{n_i}{n}\frac{\sigma_e^2}{n_i} = \; \frac{\sigma_e^2}{n}.$$

Hence, $V(\hat{\alpha}_i \,) = \frac{\sigma_e^2}{n_i} + \; \frac{\sigma_e^2}{n} - \; 2\frac{\sigma_e^2}{n} = \frac{\sigma_e^2}{n_i} - \; \frac{\sigma_e^2}{n}$.

In particular if all group sizes are equal, say to r, *i.e.* if $n_i = r, \forall \; i = \; 1,2,\cdots,k$, then $n = rk$ and

$$V(\hat{\alpha}_i \,) = \frac{\sigma_e^2}{r} - \; \frac{\sigma_e^2}{rk} = \; \frac{\sigma_e^2}{r}\left(1 - \; \frac{1}{k}\right) = \frac{(k-1)\sigma_e^2}{rk}.$$

6. 2.7 Expectation of Sum of Squares

We have $y_{ij} = \mu + \; \alpha_i + e_{ij}; j = 1,2,\cdots,n_i, i = \; 1,2,\cdots,k;$

$\bar{y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} = \mu + \; \alpha_i + \bar{e}_{i.}, \forall \; i = \; 1,2,\cdots,k$, and

$$\bar{y}_{..} = \mu + \bar{e}_{..},$$

Then

$$E(y_{ij}^2) = E(\mu^2 + \; \alpha_i^2 + e_{ij}^2 + \; 2\mu\alpha_i + \; 2\mu e_{ij} + 2\alpha_i e_{ij})$$

$$= E(\mu^2) + \; E(\alpha_i^2) + E(e_{ij}^2) + \; 2\mu E(\alpha_i) + \; 2\mu E(e_{ij}) + 2E(\alpha_i)E(e_{ij})$$

$$= \mu^2 + \alpha_i^2 + \sigma_e^2 + 2\mu\alpha_i \, .$$

$$E(\bar{y}_{i.}^2) = E(\mu^2 + \alpha_i^2 + \bar{e}_{i.}^2 + 2\mu\alpha_i + 2\mu\bar{e}_{i.} + 2\alpha_i\bar{e}_{i.})$$

$$= E(\mu^2) + E(\alpha_i^2) + E(\bar{e}_{i.}^2) + 2\mu E(\alpha_i) + 2\mu E(\bar{e}_{i.}) + 2E(\alpha_i)E(\bar{e}_{i.})$$

$$= \mu^2 + \alpha_i^2 + \frac{\sigma_e^2}{n_i} + 2\mu\alpha_i.$$

$$E(\bar{y}_{..}^2) = E(\mu^2 + \bar{e}_{..}^2 + 2\mu\bar{e}_{..})$$

$$= E(\mu^2) + E(\bar{e}_{..}^2) + 2\mu E(\bar{e}_{..}) = \mu^2 + \frac{\sigma_e^2}{n}.$$

$E(S.S.G.) = E\{\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2\}$

$$= E\{\sum_{i=1}^{k} n_i\bar{y}_{i.}^2 - n\bar{y}_{..}^2\}$$

$$= \sum_{i=1}^{k} n_i E(\bar{y}_{i.}^2) - nE(\bar{y}_{..}^2)$$

$$= \sum_{i=1}^{k} n_i \left(\mu^2 + \alpha_i^2 + \frac{\sigma_e^2}{n_i} + 2\mu\alpha_i\right) - n(\mu^2 + \frac{\sigma_e^2}{n})$$

$$= n\mu^2 + \sum_{i=1}^{k} n_i\,\alpha_i^2 + k\,\sigma_e^2 + 2\mu\sum_{i=1}^{k} n_i\alpha_i - n\mu^2 - \sigma_e^2$$

$$= \sum_{i=1}^{k} n_i\,\alpha_i^2 + (k-1)\sigma_e^2.$$

Or $E(M.S.G.) = E\left(\frac{S.S.G}{k-1}\right) = \frac{1}{(k-1)}\sum_{i=1}^{k} n_i\,\alpha_i^2 + \sigma_e^2.$

Now $E(S.S.E.) = E\{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2\}$

$$= E\{\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{k} n_i\bar{y}_{i.}^2\}$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} E(y_{ij}^2) - \sum_{i=1}^{k} n_i E(\bar{y}_{i.}^2)$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mu^2 + \alpha_i^2 + \sigma_e^2 + 2\mu\alpha_i) - \sum_{i=1}^{k} n_i(\mu^2 + \alpha_i^2 + \frac{\sigma_e^2}{n_i} + 2\mu\alpha_i)$$

$= \quad n\mu^2 + \sum_{i=1}^{k} n_i\,\alpha_i^2 + n\sigma_e^2 + 2\mu \sum_{i=1}^{k} n_i\alpha_i - n\mu^2 - \sum_{i=1}^{k} n_i\,\alpha_i^2 - k\sigma_e^2 - 2\mu \sum_{i=1}^{k} n_i\alpha_i$

$= (n-k)\sigma_e^2.$

Or $E(M.S.E.) = E\left(\frac{S.S.E}{n-k}\right) = \sigma_e^2.$

Thus under $\boldsymbol{H_0}$, $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$. Hence,

$E(M.S.G.) = \sigma_e^2 = E(M.S.E.).$

Also under H0, *S.S.G.* follows a χ^2 distribution with $k - 1$ degrees of freedom and *S.S.E.* follows a χ^2 distribution with $n - k$ degrees of freedom.

Hence, for testing $\boldsymbol{H_0}$, the test statistic is given by $F = \frac{S.S.G/(k-1)}{S.S.E./(n-k)} = \frac{M.S.G}{M.S.E}$ which will follow a central *F* distribution with $k - 1$ and $n - k$ degrees of freedom.

<u>ANOVA Table for a one-way classified data with one observation per cell</u>

Sources of Variation	**Degrees of freedom**	**Sum of Squares**	**Mean Sum of Squares**	**Variance ratio**
Groups	$k - 1$	$S.S.G. = \sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$	$M.S.G = \frac{S.S.G.}{k-1}$	$F = \frac{M.S.G.}{M.S.E}$
Error	$n - k$	$S.S.E. = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2$	$M.S.E. = \frac{S.S.E}{n-k}$	
Total	$n - 1$	$T.S.S. = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{..}\right)^2$		

If the calculated value of *F* is greater than the tabulated value of *F* at $k - 1$ and $n - k$ degrees of freedom, then reject the null hypothesis $\boldsymbol{H_0}$ otherwise it may be accepted.

6. 2.8 For Practical calculations

We have *T.S.S* = $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{..}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}^2 + \bar{y}_{..}^{\,2} - 2y_{ij}\bar{y}_{..})$

$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 + n\bar{y}_{..}^{\,2} - 2\bar{y}_{..}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}$

$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 + n\bar{y}_{..}^{\,2} - 2n\bar{y}_{..}^{\,2}$ $\quad [\because \bar{y}_{..} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} \Rightarrow n\bar{y}_{..} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}\ ]$

$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - n\bar{y}_{..}^{\,2} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}\right)^2}{n}$

$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{T_{..}^{\,2}}{n}$

= *Raw Sum of Squares*(*R.S.S.*) – *Correction factor*(*C.F.*)

S.S.G = $\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2 = \sum_{i=1}^{k} n_i(\bar{y}_{i.}^{\,2} + \bar{y}_{..}^{\,2} - 2\bar{y}_{i.}\,\bar{y}_{..})$

$= \sum_{i=1}^{k} n_i\bar{y}_{i.}^{\,2} + n\,\bar{y}_{..}^{\,2} - 2\,\bar{y}_{..}\sum_{i=1}^{k} n_i\,\bar{y}_{i.}$

$= \sum_{i=1}^{k} n_i\bar{y}_{i.}^{\,2} + n\,\bar{y}_{..}^{\,2} - 2n\bar{y}_{..}^{\,2}$

$= \sum_{i=1}^{k} n_i\bar{y}_{i.}^{\,2} - n\,\bar{y}_{..}^{\,2} = \sum_{i=1}^{k} n_i\left(\frac{T_{i.}}{n_i}\right)^2 - n\left(\frac{T_{..}}{n}\right)^2$

$= \sum_{i=1}^{k}\frac{T_{i.}^2}{n_i} - \frac{T_{..}^2}{n} = \sum_{i=1}^{k}\frac{T_{i.}^2}{n_i} - C.F.$

S.S.E. = *T.S.S.* – *S.S.G.*

6. 3 .Analysis of Variance of a Two-way layout

Suppose there are n observations y_{ij},$(i = 1,2,\cdots,k; j = 1,2,\cdots,h)$ of a random variable Y are grouped into k rows and h columns respectively. Then $n = hk$ and the observation table is as follows.

Columns		*Row Totals*	*Row*

Rows \	*1* *2* $\cdots$ *j* $\cdots$ *h*		*Means*
1	y_{11} y_{12} $\cdots$ y_{1j} $\cdots$ y_{1h}	$T_{1.} = \sum_{j=1}^{h} y_{1j}$	$\bar{y}_{1.} = \frac{T_{1.}}{h}$
2	y_{21} y_{22} $\cdots$ y_{2j} $\cdots$ y_{2h}	$T_{2.} = \sum_{j=1}^{h} y_{2j}$	$\bar{y}_{2.} = \frac{T_{2.}}{h}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	y_{i1} y_{i2} $\cdots$ y_{ij} $\cdots$ y_{ih}	$T_{i.} = \sum_{j=1}^{h} y_{ij}$	$\bar{y}_{i.} = \frac{T_{i.}}{h}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	y_{k1} y_{k2} $\cdots$ y_{kj} $\cdots$ y_{kh}	$T_{k.} = \sum_{j=1}^{h} y_{kj}$	$\bar{y}_{k.} = \frac{T_{k.}}{h}$
Column Totals	$T_{.1} = \sum_{i=1}^{k} y_{i1}$ $T_{.2} = \sum_{i=1}^{k} y_{i2}$ $\cdots$ $T_{.j} = \sum_{i=1}^{k} y_{ij}$ $\cdots$ $T_{.h} = \sum_{i=1}^{k} y_{ih}$	$T_{..} = \sum_{i=1}^{k}\sum_{j=1}^{h} y_{ij}$	
Column Means	$\bar{y}_{.1} = \frac{T_{.1}}{k}$ $\bar{y}_{.2} = \frac{T_{.2}}{k}$ $\cdots$ $\bar{y}_{.j} = \frac{T_{.j}}{k}$ $\cdots$ $\bar{y}_{.h} = \frac{T_{.h}}{k}$		$\bar{y}_{..} = \frac{T_{..}}{hk}$

6.3.1 Statistical Analysis of two-way classification:

The total variation in the observation can be split into the following two components.

i. The variation between the classes or *assignable cause* of variation which are due to classification into different rows and column and

ii. The variation within the rows or columns or *chance cause* of variation.

Let y_{ij} denote the value of the observation in the $(i, j)^{\text{th}}$ cell and suppose that y_{ij}'s are *iid* random variables, distributed according to $N(\mu_{ij}, \sigma_e^2)$. Then the mathematical model is

$$(1)\ y_{ij} = \mu_{ij} + e_{ij}; i = 1,2,\cdots,k; j = 1,2,\cdots,h,$$

Where $e_{ij}'s$the error effect due to chance and these are are assumed to be *iid* random variables each following$N(0, \sigma_e^2)$; $i = 1,2,\cdots,k, j = 1,2,\cdots,h$.

μ_{ij} is further split into (1) $\mu = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{h}\mu_{ij} = \frac{1}{hk}\sum_{i=1}^{k}\sum_{j=1}^{h}\mu_{ij}$, the over all mean;

(2) the row effect $\alpha_i = \mu_{i.} - \mu$, where $\mu_{i.} = \frac{1}{h}\sum_{j=1}^{h}\mu_{ij}$; and

(3) the column effect $\beta_j = \mu_{.j} - \mu$, where $\mu_{.j} = \frac{1}{k}\sum_{i=1}^{k}\mu_{ij}$.

Thus,

$\mu_{ij} = \mu + \mu_{i.} - \mu + \mu_{.j} - \mu = \mu + \alpha_i + \beta_j$.

Obviously,

$\sum_{i=1}^{k}\alpha_i = \sum_{i=1}^{k}(\mu_{i.} - \mu) = \sum_{i=1}^{k}\mu_{i.} - k\,\mu = k\,\mu - k\,\mu = 0$.

Similarly,

$\sum_{j=1}^{h}\beta_j = \sum_{j=1}^{h}\left(\mu_{.j} - \mu\right) = \sum_{j=1}^{h}\mu_{.j} - h\,\mu = h\,\mu - h\,\mu = 0$.

Hence the mathematical model is given by

$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}; i = 1,2,\cdots,k; j = 1,2,\cdots,h$

(1)

Where y_{ij} is the observation of i^{th} row and j^{th} column? $i = 1,2,\cdots,k, j = 1,2,\cdots,h,$

μ is the general mean effect

α_i is the additive effect due to i^{th}row; $i = 1,2,\cdots,k$;

β_j is the additive effect due to j^{th}column; $i = 1,2,\cdots,k$; and

$e_{ij}'s$ are the error effect due to chance and these are assumed to be *iid* random variables each following $N(0, \sigma_e^2)$; $i = 1,2,\cdots,k, j = 1,2,\cdots,h$.

The side conditions are $\sum_{i=1}^{k} \alpha_i = \sum_{j=1}^{h} \beta_j = 0$.

Summing (1) over j and dividing by h, we get

$\bar{y}_{i.} = \frac{1}{h} \sum_{j=1}^{h} y_{ij} = \mu + \alpha_i + \bar{e}_{i.}, \forall\, i = 1,2,\cdots,k,$

(2)

and

$\bar{e}_{i.} = \frac{1}{h} \sum_{j=1}^{h} e_{ij}$ are *iid* random variables each distributed as $N(0, \sigma_e^2/h)$.

Summing (1) over i and dividing by k, we get

$\bar{y}_{.j} = \frac{1}{k} \sum_{i=1}^{k} y_{ij} = \mu + \beta_j + \bar{e}_{.j}, \forall\, j = 1,2,\cdots,h,$

(3)

and

$\bar{e}_{.j} = \frac{1}{k} \sum_{i=1}^{k} e_{ij}$ are *iid* random variables each distributed as $N(0, \sigma_e^2/k)$.

Summing (1) over i and j and dividing by $n = hk$, we get

$\bar{y}_{..} = \frac{1}{hk} \sum_{i=1}^{k} \sum_{j=1}^{h} y_{ij} = \mu + \bar{e}_{..},$

(4)

Where $\bar{e}_{..} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} e_{ij}$ are *iid* random variables each distributed as$N(0, \sigma_e^2/hk)$.

The null hypotheses to be tested are ***H₀₁***: The rows do not differ significantly or there is no additive effect due to different rows. In other words,

$\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$ and

H₀₂: The columns do not differ significantly or there is no additive effect due to different columns. In other words,

$\beta_1 = \beta_2 = \cdots = \beta_h = 0$.

Now the total variation in each observation is given by the total sum of squares as

T.S.S. $= \sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij} - \bar{y}_{..}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{h}(\bar{y}_{i.} - \bar{y}_{..} + \bar{y}_{.j} - \bar{y}_{..} + y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$

$= \sum_{i=1}^{k}\sum_{j=1}^{h}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{h}\left(\bar{y}_{.j} - \bar{y}_{..}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$

$= h\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2 + k\sum_{j=1}^{h}\left(\bar{y}_{.j} - \bar{y}_{..}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}\right)^2$.

Or *T.S.S. = S.S.R. + S.S.C. +S.S.E,*

where *T.S.S* = Total sum of squares = $\sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij} - \bar{y}_{..}\right)^2$. *S.S.R* = Sum of squares due to rows = $h\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2$. *S.S.C* = Sum of squares due to columns = $k\sum_{j=1}^{h}\left(\bar{y}_{.j} - \bar{y}_{..}\right)^2$ and *S.S.E* = Sum of squares due to error or residuals = $\sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}\right)^2$.

6.3.2 Degrees of freedom

T.S.S = $\sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij} - \bar{y}_{..}\right)^2$ is computed from *hk* quantities of the type $\left(y_{ij} - \bar{y}_{..}\right)$ with one restriction that $\sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij} - \bar{y}_{..}\right) = 0$. Hence, it has *hk* – 1 degrees of freedom.

S.S.R = $h\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2$ is computed from *k* quantities of the type $(\bar{y}_{i.} - \bar{y}_{..})$ with one restriction of the type $\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..}) = 0$. Therefore *S.S.R* has *k* – 1 degrees of freedom.

S.S.C = $k\sum_{j=1}^{h}\left(\bar{y}_{.j} - \bar{y}_{..}\right)^2$ is computed from *h* quantities of the type $\left(\bar{y}_{.j} - \bar{y}_{..}\right)$ with one restriction of the type $\sum_{j=1}^{h}\left(\bar{y}_{.j} - \bar{y}_{..}\right) = 0$. Therefore *S.S.R* has *h* – 1 degrees of freedom.

Finally, $S.S.E = \sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}\right)^2 = T.S.S - S.S.R. - S.S.C.$ Hence its degree of freedom is given by $hk - 1 - (k - 1) - (h - 1) = hk - k - h + 1 = (h - 1).(k - 1)$.

6.3.3 Least square estimates

In the mathematical model (1), , α_i and β_j , $i = 1,2,\cdots,k, j = 1,2,\cdots,h$ are the unknown parameters which have to be estimated by the principle of least squares. Hence, we consider the sum of squares due to errors, which is given by

$$S.S.E = \sum_{i=1}^{k}\sum_{j=1}^{h} e_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij} - \mu - \alpha_i - \beta_j\right)^2.$$

(5)

Differentiating (5) with respect to, α_i and β_j and equating to zero individually, we get

$$\frac{dS.S.E}{d\mu} = 0 \Rightarrow -2\sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij} - \mu - \alpha_i - \beta_j\right) = 0$$

$$\Rightarrow \sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij} - \mu - \alpha_i - \beta_j\right) = 0$$

$$\Rightarrow \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = hk\mu + h\sum_{i=1}^{k}\alpha_i + k\sum_{j=1}^{h}\beta_j = hk\mu\ .$$

Hence, the estimate of μ is given by

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \bar{y}_{..}.$$

$$\frac{dS.S.E}{d\alpha_i} = 0 \Rightarrow -2\sum_{j=1}^{h}\left(y_{ij} - \mu - \alpha_i - \beta_j\right) = 0,\ i = 1,2,\cdots,k.$$

$$\Rightarrow \sum_{j=1}^{h}\left(y_{ij} - \mu - \alpha_i - \beta_j\right) = 0$$

$$\Rightarrow \sum_{j=1}^{h} y_{ij} = h\mu + h\alpha_i + \sum_{j=1}^{h}\beta_j$$

$$\Rightarrow \hat{\alpha}_i = \frac{1}{h}\sum_{j=1}^{h} y_{ij} - \hat{\mu} = \bar{y}_{i.} - \bar{y}_{..}$$

$$\frac{dS.S.E}{d\beta_j} = 0 \Rightarrow -2\sum_{i=1}^{k}\left(y_{ij} - \mu - \alpha_i - \beta_j\right) = 0,\ j = 1,2,\cdots,h.$$

$\Rightarrow \sum_{i=1}^{k}\left(y_{ij} - \mu - \alpha_i - \beta_j\right) = 0$

$\Rightarrow \sum_{i=1}^{k} y_{ij} = k\mu + \sum_{i=1}^{k} \alpha_i + k\,\beta_j = k\mu + k\beta_j$

$\Rightarrow \hat{\beta}_j = \frac{1}{k}\sum_{i=1}^{k} y_{ij} - \hat{\mu} = \bar{y}_{.j} - \bar{y}_{..}$

6.3.4 Variance of the estimates

We have $\hat{\mu} = \bar{y}_{..}$, $\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$ and $\hat{\beta}_j = \bar{y}_{.j} - \bar{y}_{..}$

$V(\hat{\mu}\,) = E[\bar{y}_{..} - E(\bar{y}_{..})]^2 = E[\mu + \bar{e}_{..} - \mu]^2$

$= E\left[\,\bar{e}_{....}\right]^2 = V(\bar{e}_{..}) = \frac{\sigma_e^2}{hk}.$

Also we have $\hat{\alpha}_i - E(\hat{\alpha}_i) = \bar{y}_{i.} - \bar{y}_{..} - E(\bar{y}_{i.} - \bar{y}_{..}).$

Now $\bar{y}_{i.} - \bar{y}_{..} = \mu + \alpha_i + \bar{e}_{i.} - \mu - \bar{e}_{..} = \bar{e}_{i.} - \bar{e}_{..} + \alpha_i$

$E(\bar{y}_{i.} - \bar{y}_{..}) = \alpha_i$. Hence,

$\hat{\alpha}_i - E(\hat{\alpha}_i) = \bar{e}_{i.} - \bar{e}_{..} + \alpha_i - \alpha_i = \bar{e}_{i.} - \bar{e}_{..}$

Hence, $V(\hat{\alpha}_i\,) = E[\bar{e}_{i.} - \bar{e}_{..}]^2 = E[\bar{e}_{i.}^2 + \bar{e}_{..}^2 - 2\bar{e}_{i.}\bar{e}_{..}]$

$= E(\bar{e}_{i.}^2) + E(\bar{e}_{..}^2) - 2E(\bar{e}_{i.}\bar{e}_{..}).$

Now, $E(\bar{e}_{i.}\bar{e}_{..}) = E(\frac{1}{h}\sum_{j=1}^{h} e_{ij}\,\frac{1}{kh}\sum_{i=1}^{k}\sum_{j=1}^{h} e_{ij})$

$= \frac{1}{kh^2}E[e_{i1}^2 + e_{i2}^2 + \cdots + e_{ih}^2] + \frac{1}{kh^2}E\left[\sum_{j=1}^{h} e_{ij}\sum_{g\neq i=1}^{k}\left(e_{g1} + \cdots + e_{gh}\right)\right]$

$= \frac{1}{kh^2}E[e_{i1}^2 + e_{i2}^2 + \cdots + e_{ih}^2]$ since $E(e_{ij}e_{gj}) = 0$ for $g \neq i$;

$= \frac{1}{kh^2}\sum_{j=1}^{n_i} E\left(e_{ij}^2\right) = \frac{1}{kh^2}\sum_{j=1}^{h} V\left(e_{ij}\right) = \frac{1}{kh^2}h\sigma_e^2 = \frac{\sigma_e^2}{kh}.$

Hence, $V(\hat{\alpha}_i\,) = \frac{\sigma_e^2}{h} + \frac{\sigma_e^2}{kh} - 2\frac{\sigma_e^2}{kh} = \frac{\sigma_e^2}{h} - \frac{\sigma_e^2}{kh} = \frac{\sigma_e^2}{h}\frac{(k-1)}{k}.$

Similarly, for $V(\hat{\beta}_j)$, we have

$\hat{\beta}_j - E(\hat{\beta}_j) = \bar{y}_{.j} - \bar{y}_{..} - E(\bar{y}_{.j} - \bar{y}_{..})$

$\bar{y}_{.j} - \bar{y}_{..} = \mu + \beta_j + \bar{e}_{.j} - \mu - \bar{e}_{..} = \bar{e}_{.j} - \bar{e}_{..} + \beta_j$

$E(\bar{y}_{.j} - \bar{y}_{..}) = \beta_j$. Hence,

$V(\hat{\beta}_j) = E(\bar{e}_{.j} - \bar{e}_{..})^2 = E[\bar{e}_{.j}^2 + \bar{e}_{..}^2 - 2\bar{e}_{.j}\bar{e}_{..}]$

$= E(\bar{e}_{.j}^2) + E(\bar{e}_{..}^2) - 2E(\bar{e}_{.j}\bar{e}_{..})$

Now, $E(\bar{e}_{.j}\bar{e}_{..}) = E(\frac{1}{k}\sum_{i=1}^{k} e_{ij} \frac{1}{kh}\sum_{i=1}^{k}\sum_{j=1}^{h} e_{ij})$

$= \frac{1}{hk^2} E[e_{1j}^2 + e_{2j}^2 + \cdots + e_{kj}^2] + \frac{1}{kh^2} E[\sum_{i=1}^{k} e_{ij} \sum_{l \neq j=1}^{h} (e_{1l} + \cdots + e_{kl})]$

$= \frac{1}{hk^2} E[e_{1j}^2 + e_{2j}^2 + \cdots + e_{kj}^2]$ since $E(e_{ij}e_{il}) = 0$ for $l \neq j$;

$= \frac{1}{hk^2} \sum_{i=1}^{k} E(e_{ij}^2) = \frac{1}{hk^2}\sum_{i=1}^{k} V(e_{ij}) = \frac{1}{hk^2} k\sigma_e^2 = \frac{\sigma_e^2}{kh}$.

Hence, $V(\hat{\beta}_j) = \frac{\sigma_e^2}{k} + \frac{\sigma_e^2}{kh} - 2\frac{\sigma_e^2}{kh} = \frac{\sigma_e^2}{k} - \frac{\sigma_e^2}{kh} = \frac{\sigma_e^2}{k}\frac{(h-1)}{h}$

6.3.5 Expectation of Sum of Squares

We have $y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}; i = 1,2,\cdots,k; j = 1,2,\cdots,h$

$\bar{y}_{i.} = \frac{1}{h}\sum_{j=1}^{h} y_{ij} = \mu + \alpha_i + \bar{e}_{i.}, \forall\ i = 1,2,\cdots,k,$

$\bar{y}_{.j} = \frac{1}{k}\sum_{i=1}^{k} y_{ij} = \mu + \beta_j + \bar{e}_{.j}, \forall\ j = 1,2,\cdots,h,$ and

$\bar{y}_{..} = \frac{1}{hk}\sum_{i=1}^{k}\sum_{j=1}^{h} y_{ij} = \mu + \bar{e}_{..},$

where $\bar{e}_{i.} = \frac{1}{h}\sum_{j=1}^{h} e_{ij}$ are *iid* random variables each distributed as $N(0, \sigma_e^2/h)$, $\bar{e}_{.j}$ $= \frac{1}{k}\sum_{i=1}^{k} e_{ij}$ are *iid* random variables each distributed as $N(0, \sigma_e^2/k)$ and $\bar{e}_{..}$ $=$ $\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}$ are *iid* random variables each distributed as $N(0, \sigma_e^2/hk)$.

Then

$$E(y_{ij}^2) = E(\mu^2 + \alpha_i^2 + \beta_j^2 + e_{ij}^2 + 2\mu\alpha_i + 2\mu\beta_j + 2\mu e_{ij} + 2\alpha_i\beta_j + 2\alpha_i e_{ij} + 2\beta_j e_{ij})$$

$$= E(\mu^2) + E(\alpha_i^2) + E\left(\beta_j^2\right) + E\left(e_{ij}^2\right) + 2\mu E(\alpha_i) + 2\mu E\left(\beta_j\right) + 2\mu E(e_{ij}) + 2E(\alpha_i)E(\beta_j) + 2E\left(\alpha_i\right)E\left(e_{ij}\right) + 2E(\beta_j)E(e_{ij})$$

$$= \mu^2 + \alpha_i^2 + \beta_j^2 + 2\mu\alpha_i + 2\mu\beta_j + 2\alpha_i\beta_j + \sigma_e^2 .$$

$$E(\bar{y}_{i.}^2) = E(\mu^2 + \alpha_i^2 + \bar{e}_{i.}^2 + 2\mu\alpha_i + 2\mu\bar{e}_{i.} + 2\alpha_i\bar{e}_{i.})$$

$$= E(\mu^2) + E(\alpha_i^2) + E(\bar{e}_{i.}^2) + 2\mu E(\alpha_i) + 2\mu E(\bar{e}_{i.}) + 2E(\alpha_i)E(\bar{e}_{i.})$$

$$= \mu^2 + \alpha_i^2 + \frac{\sigma_e^2}{h} + 2\mu\alpha_i.$$

$$E(\bar{y}_{.j}^2) = E(\mu^2 + \beta_j^2 + \bar{e}_{.j}^2 + 2\mu\beta_j + 2\mu\bar{e}_{.j} + 2\beta_j\bar{e}_{.j})$$

$$= E(\mu^2) + E(\beta_j^2) + E\left(\bar{e}_{.j}^2\right) + 2\mu E(\beta_j) + 2\mu E(\bar{e}_{.j}) + 2E(\beta_j)E(\bar{e}_{.j})$$

$$= \mu^2 + \beta_j^2 + \frac{\sigma_e^2}{k} + 2\mu\beta_j.$$

$$E(\bar{y}_{..}^2) = E(\mu^2 + \bar{e}_{..}^2 + 2\mu\bar{e}_{..})$$

$$= E(\mu^2) + E(\bar{e}_{..}^2) + 2\mu E(\bar{e}_{..}) = \mu^2 + \frac{\sigma_e^2}{hk}.$$

$$E(S.S.R.) = E\{h\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2\}$$

$$= E\{h\sum_{i=1}^{k}\bar{y}_{i.}^2 - hk\bar{y}_{..}^2\}$$

$$= h\sum_{i=1}^{k}E(\bar{y}_{i.}^2) - hkE(\bar{y}_{..}^2)$$

$$= h\sum_{i=1}^{k}\left(\mu^2 + \alpha_i^2 + \frac{\sigma_e^2}{h} + 2\mu\alpha_i\right) - hk(\mu^2 + \frac{\sigma_e^2}{hk})$$

$$= hk\mu^2 + h\sum_{i=1}^{k}\alpha_i^2 + k\,\sigma_e^2 + 2\mu\sum_{i=1}^{k}\alpha_i - hk\mu^2 - \sigma_e^2$$

$$= h\sum_{i=1}^{k}\alpha_i^2 + (k-1)\sigma_e^2. \qquad \text{[Since } \sum_{i=1}^{k}\alpha_i = 0\text{].}$$

Or E (M.S.R.) = $E\left(\frac{S.S.R}{k-1}\right) = \frac{h}{k-1}\sum_{i=1}^{k}\alpha_i^2 + \sigma_e^2.$

E (S.S.C.) = $E\{k \sum_{j=1}^{h}\left(\bar{y}_{.j} - \bar{y}_{..}\right)^2\}$

$$= E\{k\sum_{j=1}^{h}\bar{y}_{.j}^2 - hk\bar{y}_{..}^2\}$$

$$= k\sum_{j=1}^{h}E\left(\bar{y}_{.j}^2\right) - hkE(\bar{y}_{..}^2)$$

$$= k\sum_{j=1}^{h}\left(\mu^2 + \beta_j^2 + \frac{\sigma_e^2}{k} + 2\mu\beta_j\right) - hk\left(\mu^2 + \frac{\sigma_e^2}{hk}\right)$$

$$= hk\mu^2 + k\sum_{j=1}^{h}\beta_j^2 + h\sigma_e^2 + 2\mu\sum_{j=1}^{h}\beta_j - hk\mu^2 - \sigma_e^2$$

$$= k\sum_{j=1}^{h}\beta_j^2 + (h-1)\sigma_e^2. \qquad \text{[since } \sum_{j=1}^{h}\beta_j = 0\text{].}$$

Or

E(M.S.C.) = $E\left(\frac{S.S.C}{h-1}\right) = \frac{k}{h-1}\sum_{j=1}^{h}\beta_j^2 + \sigma_e^2$

Now, E(S.S.E.) = $E\{\sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}\right)^2\}$

$$= E\{\sum_{i=1}^{k}\sum_{j=1}^{h}(y_{ij}^2 + \bar{y}_{i.}^2 + \bar{y}_{.j}^2 + \bar{y}_{..}^2 - 2y_{ij}\bar{y}_{i.} - 2y_{ij}\bar{y}_{.j} + 2y_{ij}\bar{y}_{..} + 2\bar{y}_{i.}\bar{y}_{.j} - 2\bar{y}_{i.}\bar{y}_{..} - 2\bar{y}_{.j}\bar{y}_{..})\}$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{h}E(y_{ij}^2) + h\sum_{i=1}^{k}E(\bar{y}_{i.}^2) + k\sum_{j=1}^{h}E\left(\bar{y}_{.j}^2\right) + hkE(\bar{y}_{..}^2) - 2E\{\sum_{i=1}^{k}\bar{y}_{i.}\sum_{j=1}^{h}y_{ij}\} -$$

$$2E\{\sum_{j=1}^{h}\bar{y}_{.j}\sum_{i=1}^{k}y_{ij}\} + 2E\{\bar{y}_{..}\sum_{i=1}^{k}\sum_{j=1}^{h}y_{ij}\} + 2E\{\sum_{i=1}^{k}\bar{y}_{i.}\sum_{j=1}^{h}\bar{y}_{.j}\} - 2E\{h\,\bar{y}_{..}\sum_{i=1}^{k}\bar{y}_{i.}\}$$

$$- 2E\{k\bar{y}_{..}\sum_{j=1}^{h}\bar{y}_{.j}\}$$

$= \sum_{i=1}^{k}\sum_{j=1}^{h}E(y_{ij}^2) + h\sum_{i=1}^{k}E(\bar{y}_{i.}^2) + k\sum_{j=1}^{h}E\left(\bar{y}_{.j}^2\right) + hkE(\bar{y}_{..}^2) - 2h\sum_{i=1}^{k}E(\bar{y}_{i.}^2) - 2k\sum_{j=1}^{h}E(\bar{y}_{.j}^2) + 2hkE(\bar{y}_{..}^2) + 2hkE(\bar{y}_{..}^2) - 2hkE(\bar{y}_{..}^2) - 2hkE(\bar{y}_{..}^2)$

$= \sum_{i=1}^{k}\sum_{j=1}^{h}E(y_{ij}^2) - h\sum_{i=1}^{k}E(\bar{y}_{i.}^2) - k\sum_{j=1}^{h}E\left(\bar{y}_{.j}^2\right) + hkE(\bar{y}_{..}^2)$

$= \sum_{i=1}^{k}\sum_{j=1}^{h}\left(\mu^2 + \alpha_i^2 + \beta_j^2 + 2\mu\alpha_i + 2\mu\beta_j + 2\alpha_i\beta_j + \sigma_e^2\right) - h\sum_{i=1}^{k}(\mu^2 + \alpha_i^2 + \frac{\sigma_e^2}{h} + 2\mu\alpha_i)$

$$-k\sum_{j=1}^{h}\left(\mu^2 + \beta_j^2 + \frac{\sigma_e^2}{k} + 2\mu\beta_j\right) + hk\left(\mu^2 + \frac{\sigma_e^2}{hk}\right)$$

$= hk\mu^2 + h\sum_{i=1}^{k}\alpha_i^2 + k\sum_{j=1}^{h}\beta_j^2 + hk\sigma_e^2 - hk\mu^2 - h\sum_{i=1}^{k}\alpha_i^2 - k\sigma_e^2 - hk\mu^2 - k\sum_{j=1}^{h}\beta_j^2 - h\sigma_e^2 + hk\mu^2 + \sigma_e^2$ [since $\sum_{i=1}^{k}\alpha_i = 0$ and $\sum_{j=1}^{h}\beta_j = 0$].

$= (hk - k - h + 1)\sigma_e^2 = (k-1)(h-1)\sigma_e^2$.

Or E (*M.S.E.*) = $E\left(\frac{S.S.E}{(k-1)(h-1)}\right) = \sigma_e^2$.

Thus under H01, $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0 \Rightarrow \sum_{i=1}^{k}\alpha_i^2 = 0$. Hence,

$E(M.S.R.) = \sigma_e^2 = E(M.S.E.)$.

Also under H01, *S.S.R.* $/\sigma_e^2$ follows a χ^2 distribution with k – 1 degrees of freedom and *S.S.E.* $/\sigma_e^2$ follows a χ^2 distribution with $(k-1)(h-1)$ degrees of freedom.

Hence, for testing ***H*₀₁**, the test statistic is given by $F_R = \frac{S.S.R/(k-1)}{S.S.E./(k-1)(h-1)} = \frac{M.S.R}{M.S.E}$.

Which will follow a central *F* distribution with *k* – 1 and $(k-1)(h-1)$ degrees of freedom.

Similarly, under H02, $\beta_1 = \beta_2 = \cdots = \beta_h = 0 \Rightarrow \sum_{j=1}^{h}\beta_j^2 = 0$. Hence,

$E(M.S.C.) = \sigma_e^2 = E(M.S.E.)$.

Also under H02, $S.S.C. /\sigma_e^2$ follows a χ^2 distribution with h – 1 degrees of freedom and $S.S.E. /\sigma_e^2$ follows a χ^2 distribution with $(k-1)(h-1)$ degrees of freedom.

Hence, for testing $\boldsymbol{H_0}$, the test statistic is given by FC = $\frac{S.S.C/(h-1)}{S.S.E./(k-1)(h-1)} = \frac{M.S.C}{M.S.E}$.

Which will follow a central F distribution with h – 1 and $(k-1)(h-1)$ degrees of freedom.

ANOVA Table for a two-way classified data with one observation per cell

Sources of Variation	**Degrees of freedom**	**Sum of Squares**	**Mean Sum of Squares**	**Variance ratio**
Rows	k – 1	S.S.R.= $h\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2$	M.S.R = $\frac{S.S.R.}{k-1}$	F$_R$ = $\frac{M.S.R.}{M.S.E}$
Columns	h – 1	S.S.C.= $k\ \sum_{j=1}^{h}\left(\bar{y}_{.j} - \bar{y}_{..}\right)^2$	M.S.C = $\frac{S.S.C.}{h-1}$	F$_C$ = $\frac{M.S.C.}{M.S.E.}$
Error	(k – 1)(h – 1)	S.S.E.=$\sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}\right)^2$	M.S.E. = $\frac{S.S.E.}{(k-1)(h-1)}$	
Total	kh – 1	T.S.S. =$\sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij} - \bar{y}_{..}\right)^2$		

6.3.6 For Practical calculations

We have $T.S.S = \sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij} - \bar{y}_{..}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij}^2 + \bar{y}_{..}^2 - 2y_{ij}\bar{y}_{..}\right)$

$$= \sum_{i=1}^{k}\sum_{j=1}^{h} y_{ij}^2 + kh\,\bar{y}_{..}^2 - 2\bar{y}_{..}\sum_{i=1}^{k}\sum_{j=1}^{h} y_{ij}$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{h} y_{ij}^2 + kh\,\bar{y}_{..}^2 - 2kh\bar{y}_{..}^2 = \sum_{i=1}^{k}\sum_{j=1}^{h} y_{ij}^2 - kh\,\bar{y}_{..}^2$$

$$= Raw\ Sum\ of\ Squares\ (RSS)\ \text{-}\ kh\left(\frac{T_{..}}{hk}\right)^2$$

$$= RSS - \frac{T_{..}^2}{hk} = RSS - Correction\ Factor(C.F.),$$

Where C.F. = $\frac{T_{..}^2}{hk}$.

Similarly, $S.S.R = h\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2 = h\sum_{i=1}^{k}\bar{y}_{i.}^2 - hk\bar{y}_{..}^2$

$$= h\sum_{i=1}^{k}\left(\frac{T_{i.}}{h}\right)^2 - C.F. = \frac{1}{h}\sum_{i=1}^{k}T_{i.}^2 - C.F.$$

$$S.S.C = k\ \sum_{j=1}^{h}\left(\bar{y}_{.j} - \bar{y}_{..}\right)^2 = k\ \sum_{j=1}^{h}\bar{y}_{.j}{}^2 - hk\bar{y}_{..}^2$$

$$= k\ \sum_{j=1}^{h}\left(\frac{T_{.j}}{k}\right)^2 - C.F = \frac{1}{k}\sum_{j=1}^{h}T_{.j}{}^2 - C.F\ .$$

$S.S.E = T.S.S - S.S.R. - S.S.C.$

6.3.7 Analysis of variance of a two-way classified data with equal number of Observations per cell

Suppose in a two-way layout with p rows and q columns there are equal number of observations, *i.e.* say m observations are present, and let y_{ijk} denote the k^{th} observation of $(i,j)^{th}$ cell, then the observation table is given by

Observation table

Columns **Rows**	**1**	⋯	***j***	⋯	***q***
1	y_{111} y_{112} ⋮ y_{11m}	⋯	y_{1j1} y_{1j2} ⋮ y_{1jm}	⋯	y_{1q1} y_{1q2} ⋮ y_{1qm}
⋮	⋮	⋮	⋮	⋮	⋮
i	y_{i11} y_{i12} ⋮ y_{i1m}	⋯	y_{ij1} y_{ij2} ⋮ y_{ijm}	⋯	y_{iq1} y_{iq2} ⋮ y_{iqm}

$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p	y_{p11} y_{p12} $\vdots$ y_{p1m}	$\cdots$	y_{pj1} y_{pj2} $\vdots$ y_{pjm}	$\cdots$	y_{pq1} y_{pq2} $\vdots$ y_{pqm}

Now let $T_{ij.} = \sum_{k=1}^{m} y_{ijk}$ be the total of $(i,j)^{\text{th}}$ cell for $i = 1,2,\cdots,p; j = 1,2,\cdots,q;$

$T_{i..} = \sum_{j=1}^{q} \sum_{k=1}^{m} y_{ijk} = \sum_{j=1}^{q} T_{ij.}$ be the total of i^{th} row for $i = 1,2,\cdots,p;$

$T_{.j.} = \sum_{i=1}^{p} \sum_{k=1}^{m} y_{ijk} = \sum_{i=1}^{p} T_{ij.}$ be the total of j^{th} column for $j = 1,2,\cdots,q$ and

$T_{...} = \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{k=1}^{m} y_{ijk} = \sum_{i=1}^{p} \sum_{j=1}^{q} T_{ij.} = \sum_{i=1}^{p} T_{i..} = \sum_{j=1}^{q} T_{.j.}$.

Then the table of totals is given by

Table of Totals

Columns / **Rows**	**1**	$\cdots$	***j***	$\cdots$	***q***	**Row Total**
1	$T_{11.} = \sum_{k=1}^{m} y_{11k}$	$\cdots$	$T_{1j.} = \sum_{k=1}^{m} y_{1jk}$	$\cdots$	$T_{1q.} = \sum_{k=1}^{m} y_{1qk}$	$T_{1..} = \sum_{j=1}^{q} \sum_{k=1}^{m} y_{1jk}$ $= \sum_{j=1}^{q} T_{1j.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	$T_{i1.} = \sum_{k=1}^{m} y_{i1k}$	$\cdots$	$T_{ij.} = \sum_{k=1}^{m} y_{ijk}$	$\cdots$	$T_{iq.} = \sum_{k=1}^{m} y_{iqk}$	$T_{i..} = \sum_{j=1}^{q} \sum_{k=1}^{m} y_{ijk}$ $= \sum_{j=1}^{q} T_{ij.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p	$T_{p1.} = \sum_{k=1}^{m} y_{p1k}$	$\cdots$	$T_{pj.} = \sum_{k=1}^{n_{pj}} y_{pjk}$	$\cdots$	$T_{pq.} = \sum_{k=1}^{m} y_{pqk}$	$T_{p..} = \sum_{j=1}^{q} \sum_{k=1}^{m} y_{pjk}$ $= \sum_{j=1}^{q} T_{pj.}$
Column total	$T_{.1.} =$ $\sum_{i=1}^{p} \sum_{k=1}^{m} y_{i1k} =$ $\sum_{i=1}^{p} T_{i1.}$	$\cdots$	$T_{.j.} =$ $\sum_{i=1}^{p} \sum_{k=1}^{m} y_{ijk} =$ $\sum_{i=1}^{p} T_{ij.}$	$\cdots$	$T_{.q.} =$ $\sum_{i=1}^{p} \sum_{k=1}^{m} y_{iqk} =$ $\sum_{i=1}^{p} T_{iq.}$	$T_{...} =$ $\sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{k=1}^{m} y_{ijk}$ $= \sum_{i=1}^{p} \sum_{j=1}^{q} T_{ij.} =$ $\sum_{i=1}^{p} T_{i..} = \sum_{j=1}^{q} T_{.j.}$

6.3.8 Statistical Analysis

By the basic assumptions of analysis of variance the observationsy_{ijk}'s are random sample observations from a normal distribution with mean μ_{ij} and variance σ^2, *i.e.* $y_{ijk} \sim N\left(\mu_{ij}, \sigma^2\right)$. Hence, the mathematical model is given by

$$y_{ijk} = \mu_{ij} + e_{ijk}, \qquad (1)$$

$i = 1,2,\cdots,p; j = 1,2,\cdots,q$ and$k = 1,2,\cdots,m$.

The unrestricted residual sum of squares is then given by

$$SSE = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m} e_{ijk}^2 = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(y_{ijk} - \mu_{ij}\right)^2.$$

The least square estimate of μ_{ij} is then obtained by differentiating this, equating to zero and solving for μ_{ij} as follows.

$$\frac{\partial SSE}{\partial \mu_{ij}} = 0 \Rightarrow -2\sum_{k=1}^{m}\left(y_{ijk} - \mu_{ij}\right) = 0 \text{ for all } i = 1,2,\cdots,p; j = 1,2,\cdots,q.$$

$$\Rightarrow \sum_{k=1}^{m}\left(y_{ijk} - \mu_{ij}\right) = 0 \text{ for all } i = 1,2,\cdots,p; j = 1,2,\cdots,q.$$

Hence, $\sum_{k=1}^{m} y_{ijk} = m\hat{\mu}_{ij} \Rightarrow \hat{\mu}_{ij} = \frac{1}{m}\sum_{k=1}^{m} y_{ijk} = \frac{T_{ij.}}{m} = \bar{y}_{ij.}$.

$$\therefore SSE = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(y_{ijk} - \bar{y}_{ij.}\right)^2 = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m} y_{ijk}^2 - \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{T_{ij.}^2}{m}.$$

and it will have *mpq* – *pq* = (*m* – 1)*pq* d.f.

Now the mean μ_{ij} can be split as

$$\begin{aligned}\mu_{ij} &= \mu + (\mu_{i.} - \mu) + \left(\mu_{.j} - \mu\right) + \left(\mu_{ij} - \mu_{i.} - \mu_{.j} + \mu\right)\\ &= \mu + \alpha_i + \beta_j + \gamma_{ij},\end{aligned}$$

Where $\mu = \frac{1}{p}\sum_{i=1}^{p}\mu_{i.} = \frac{1}{q}\sum_{j=1}^{q}\mu_{.j}$ is the general mean, $\alpha_i = (\mu_{i.} - \mu)$ is the additive effect due to *i*th row,

$\beta_j = \left(\mu_{.j} - \mu\right)$ is the additive effect due to *j*th column,

$\gamma_{ij} = \left(\mu_{ij} - \mu_{i.} - \mu_{.j} + \mu\right)$ is the additive effect due to the interaction between i^{th} row and j^{th} column.

Hence, the model (1) can be written as

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk},$$

(2)

$i = 1,2,\cdots,p; j = 1,2,\cdots,q$ and $k = 1,2,\cdots,m$ with side conditions as

$$\sum_{i=1}^{p} \alpha_i = \sum_{j=1}^{q} \beta_j = \sum_{i=1}^{p} \gamma_{ij} = \sum_{j=1}^{q} \gamma_{ij} = 0.$$

We now make the following null hypotheses.

H_{01}: There is no additive effect due to interaction between rows and columns, *i.e.*

$$\gamma_{ij} = 0 \text{ for all } i = 1,2,\cdots,p; j = 1,2,\cdots,q.$$

H_{02}: There is no additive effect due to rows, *i.e.* rows are homogeneous or

$$\alpha_i = 0 \text{ for all } i = 1,2,\cdots,p.$$

H_{03}: There is no additive effect due to columns, *i.e.* columns are homogeneous or

$$\beta_j = 0 \text{ for all } j = 1,2,\cdots,q.$$

To test these hypotheses, we shall first assume that the null hypothesis H_{01} to be true. Then under this restriction the model (2) reduces to

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk},$$

$$i = 1,2,\cdots,p; j = 1,2,\cdots,q \text{ and } k = 1,2,\cdots,m.$$

Hence, the restricted residual sum of squares under the restriction that H_{01} is true is given by

$$SSE^* = \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{k=1}^{m} e_{ijk}^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{k=1}^{m} \left(y_{ijk} - \mu - \alpha_i - \beta_j\right)^2.$$

The least square estimates of μ, α_i and β_j are obtained by differentiating SSE^* with respect to μ, α_i and β_j and equating to zero individually. Thus the normal equations are

(i) $\frac{\partial SSE^*}{\partial \mu} = 0 \Rightarrow \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j\right) = 0.$

Or $\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m} y_{ijk} = n\hat{\mu}.$

Or $\hat{\mu} = \frac{T_{\ldots}}{mpq} = \bar{y}_{\ldots}$

(ii)$\frac{\partial SSE^*}{\partial \alpha_i} = 0 \Rightarrow \sum_{j=1}^{q}\sum_{k=1}^{m}\left(y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j\right) = 0$ for all $i = 1,2,\cdots,p.$

Or $\sum_{j=1}^{q}\sum_{k=1}^{m} y_{ijk} = \sum_{j=1}^{q}\sum_{k=1}^{m}\hat{\mu} + \sum_{j=1}^{q}\sum_{k=1}^{m}\hat{\alpha}_i + \sum_{j=1}^{q}\sum_{k=1}^{m}\hat{\beta}_j.$

Or $T_{i..} = \sum_{j=1}^{q} m\hat{\mu} + \sum_{j=1}^{q} m\hat{\alpha}_i + \sum_{j=1}^{q} m\hat{\beta}_j.$

Or $T_{i..} = qm\hat{\mu} + qm\hat{\alpha}_i.$ [since $\sum_{j=1}^{q}\beta_j = 0$].

Or $\hat{\alpha}_i = \frac{T_{i..}}{qm} - \hat{\mu} = \bar{y}_{i..} - \bar{y}_{\ldots}.$ (3)

(iii) $\frac{\partial SSE^*}{\partial \beta_j} = 0 \Rightarrow \sum_{i=1}^{p}\sum_{k=1}^{m}\left(y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j\right) = 0$ for all $j = 1,2,\cdots,q.$

Or $\sum_{i=1}^{p}\sum_{k=1}^{m} y_{ijk} = \sum_{i=1}^{p}\sum_{k=1}^{m}\hat{\mu} + \sum_{i=1}^{p}\sum_{k=1}^{m}\hat{\alpha}_i + \sum_{i=1}^{p}\sum_{k=1}^{m}\hat{\beta}_j.$

Or $T_{.j.} = \sum_{i=1}^{p} m\hat{\mu} + \sum_{i=1}^{p} m\hat{\alpha}_i + \sum_{i=1}^{p} m\hat{\beta}_j.$

Or $T_{.j.} = pm\hat{\mu} + pm\hat{\beta}_j.$ [since $\sum_{i=1}^{p}\alpha_i = 0$].

Or $\hat{\beta}_j = \frac{T_{.j.}}{pm} - \hat{\mu} = \bar{y}_{.j.} - \bar{y}_{\ldots}.$

(4)

Now, the restricted residual sum of squares SSE^* under the restriction that the null hypothesis H_{01} is true is given by

$$SSE^* = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j\right)^2$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(y_{ijk} - \bar{y}_{\ldots} - \bar{y}_{i..} + \bar{y}_{\ldots} - \bar{y}_{.j.} + \bar{y}_{\ldots}\right)^2$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{\ldots}\right)^2$$

This will have $mpq-q-p+1$ degrees of freedom.

Hence, the sum of squares due to H_{01} or due to interaction is given by

$SSI = SSE^{*} - \ SSE$

$= \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(y_{ijk}-\bar{y}_{i..}-\bar{y}_{.j.}+\ \bar{y}_{...}\right)^{2}-\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(y_{ijk}-\ \bar{y}_{ij.}\right)^{2}$

$= \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(y_{ijk}-\bar{y}_{ij.}+\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\ \bar{y}_{...}\right)^{2}-\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}(y_{ijk}-\bar{y}_{ij.})^{2}$

$= \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(y_{ijk}-\bar{y}_{ij.}\right)^{2}+\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\ \bar{y}_{...}\right)^{2}-\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(y_{ijk}-\ \bar{y}_{ij.}\right)^{2}$ [Product terms vanish being sum of deviations from arithmetic mean.]

$= \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\ \bar{y}_{...}\right)^{2}$

This will have $mpq-q-p+1-mpq+\ pq\ = pq-q-p+1 = (p-1)(q-1)$ degrees of freedom.

Next, let us assume that the null hypothesis H_{02} is also true, *i.e.* $\alpha_i=0$ for all $i=1,2,\cdots,p,$ along withH_{01}. Then the model reduces to

$y_{ijk} = \mu+\beta_j+\ e_{ijk},$

$j=1,2,\cdots,q$ and $k=1,2,\cdots,m.$

Hence, the restricted residual sum of squares under the restriction that H_{02} is also true along with H_{01} is given by

$SSE^{**} = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}e_{ijk}^{2} = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(y_{ijk}-\ \mu-\beta_j\right)^{2}.$

The least square estimates of μ and β_j are obtained by differentiating SSE^{**} with respect to μ and β_j and equating to zero individually. Thus the normal equations are

$$\frac{\partial SSE^*}{\partial \mu} = 0 \Rightarrow \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(y_{ijk} - \hat{\mu} - \hat{\beta}_j\right) = 0.$$

Or $\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m} y_{ijk} = mpq\hat{\mu}.$

Or $\hat{\mu} = \frac{T_{\dots}}{mpq} = \bar{y}_{\dots}.$

$$\frac{\partial SSE^*}{\partial \beta_j} = 0 \Rightarrow \sum_{i=1}^{p}\sum_{k=1}^{n_{ij}}\left(y_{ijk} - \hat{\mu} - \hat{\beta}_j\right) = 0 \text{ for all } j = 1,2,\cdots,q.$$

Or $\sum_{i=1}^{p}\sum_{k=1}^{m} y_{ijk} = \sum_{i=1}^{p}\sum_{k=1}^{m}\left(\hat{\mu} + \hat{\beta}_j\right) = pm\left(\hat{\mu} + \hat{\beta}_j\right).$

Or $\left(\hat{\mu} + \hat{\beta}_j\right) = \frac{1}{pm}\sum_{i=1}^{p}\sum_{k=1}^{m} y_{ijk} = \frac{T_{.j.}}{pm} = \bar{y}_{.j.}.$

Hence,

$$SSE^{**} = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(y_{ijk} - \hat{\mu} - \hat{\beta}_j\right)^2$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(y_{ijk} - \bar{y}_{.j.}\right)^2$$

This will have *mpq* – *q* degrees of freedom.

Hence, the adjusted sum of squares due to the hypothesis H_{02}, *i.e.* the sum of squares due to rows is given by

$$SSR = SSE^{**} - SSE^*$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(y_{ijk} - \bar{y}_{.j.}\right)^2 - \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{\dots}\right)^2$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(\bar{y}_{i..} - \bar{y}_{\dots} + y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{\dots}\right)^2$$

$$- \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{\dots}\right)^2$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}(\bar{y}_{i..} - \bar{y}_{\dots})^2 + \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{\dots}\right)^2$$

$- \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{\dots}\right)^2$ [Product terms vanish being sum of deviations from arithmetic mean.]

$$= \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}(\bar{y}_{i..} - \bar{y}_{\dots})^2 = mq\sum_{i=1}^{p}(\bar{y}_{i..} - \bar{y}_{\dots})^2$$

This will have $mpq – q - mpq + q + p - 1 = p - 1$ degrees of freedom.

Similarly, the sum of squares due to columns can be obtained by interchanging the roles of *i* and *j* and that of *p* and *q* in the above derivation of SSR. Thus the sum of squares due to columns *SSC* is obtained as

$SSC = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(\bar{y}_{.j.} - \bar{y}_{...}\right)^2 = mp\sum_{j=1}^{q}\left(\bar{y}_{.j.} - \bar{y}_{...}\right)^2$ and it will have $q - 1$ degrees of freedom.

The total sum of squares (*TSS*) which has *mpq* – 1 degrees of freedom is given by

$$TSS = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(y_{ijk} - \bar{y}_{...}\right)^2$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left[(\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}) + (y_{ijk} - \bar{y}_{ij.})\right]^2$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}(\bar{y}_{i..} - \bar{y}_{...})^2 + \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(\bar{y}_{.j.} - \bar{y}_{...}\right)^2$$

$$+ \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}\right)^2 + \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(y_{ijk} - \bar{y}_{ij.}\right)^2$$

[Product terms vanish being sum of deviations from arithmetic mean.]

$= SSR + SSC + SSI + SSE.$

UnderH_{01}, $E(MSI) = E\left(\frac{SSI}{(p-1)(q-1)}\right) = \sigma^2 = E(MSE) = E\left(\frac{SSE}{(m-1)pq}\right)$. Similarly under$H_{02}$,

$E(MSR) = E\left(\frac{SSR}{(p-1)}\right) = \sigma^2 = E(MSE) = E\left(\frac{SSE}{(m-1)pq}\right)$and under H_{03}, $E(MSC) = E\left(\frac{SSC}{(q-1)}\right) = \sigma^2 = E(MSE) = E\left(\frac{SSE}{(m-1)pq}\right)$.

Hence, the test statistics for testing H_{01} is $F_I = \frac{M.S.I.}{M.S.E.}$ which will follow a central F distribution with $(p-1)(q-1)$and $(m – 1)pq$. Similarly, for testing H_{02} is $F_R = \frac{M.S.R.}{M.S.E.}$ and for testing H_{03} is $F_C = \frac{M.S.C.}{M.S.E.}$ will follow F distributions with $(p-1)$and $(m – 1)pq$ and $(q-1)$and $(m – 1)pq$ degrees of freedom respectively

Now, the ANOVA table is given by

ANOVA Table for a two way classification with equal number of observations per cell

Sources of Variation	**Degrees of freedom**	**Sum of Squares**	**Mean Sum of Squares**	**Variance ratio**
Rows	p – 1	S.S.R.= $mq\sum_{i=1}^{p}(\bar{y}_{i..} - \bar{y}_{...})^2$	M.S.R $= \frac{S.S.R.}{p-1}$	$\mathrm{F_R} = \frac{M.S.R.}{M.S.E}$
Columns	q – 1	S.S.C.= $mp\sum_{j=1}^{q}(\bar{y}_{.j.} - \bar{y}_{...})^2$	M.S.C = $\frac{S.S.C.}{q-1}$	$\mathrm{F_C} = \frac{M.S.C.}{M.S.E.}$
Interaction	$(p-1)(q-1)$	SSI =$m\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$	M.S.I. = $\frac{S.S.I.}{(p-1)(q-1)}$	$\mathrm{F_I} = \frac{M.S.I.}{M.S.E.}$
Error	$(m-1)pq$	S.S.E.=$\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}(y_{ijk} - \bar{y}_{ij.})^2$	M.S.E. = $\frac{S.S.E.}{(m-1)pq}$	
Total	mpq – 1	T.S.S. =$\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}(y_{ijk} - \bar{y}_{...})^2$		

6.3.9 For Practical calculations

We have *TSS* given by

$$TSS = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}(y_{ijk} - \bar{y}_{...})^2$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}{y_{ijk}}^2 + \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}{\bar{y}_{...}}^2 - 2\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}y_{ijk}\bar{y}_{...}$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}{y_{ijk}}^2 + mpq{\bar{y}_{...}}^2 - 2mpq\bar{y}_{...}\frac{1}{mpq}\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}y_{ijk}$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}{y_{ijk}}^2 + mpq{\bar{y}_{...}}^2 - 2mpq{\bar{y}_{...}}^2$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}{y_{ijk}}^2 - mpq{\bar{y}_{...}}^2 = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}{y_{ijk}}^2 - mpq\frac{{T_{...}}^2}{(mpq)^2}$$

$= \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}{y_{ijk}}^2 - \frac{{T_{...}}^2}{mpq}$ = Raw Sum of Squares(*RSS*) – Correction Factor(*CF*).

$$SSR = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}(\bar{y}_{i..} - \bar{y}_{...})^2 = mq\sum_{i=1}^{p}(\bar{y}_{i..} - \bar{y}_{...})^2$$

$$= mq \sum_{i=1}^{p} \bar{y}_{i..}{}^{2} + mq \sum_{i=1}^{p} \bar{y}_{...}{}^{2} - \ 2mq \sum_{i=1}^{p} \bar{y}_{i..}\bar{y}_{...}$$

$$= mq \sum_{i=1}^{p} \frac{T_{i..}{}^{2}}{(mq)^{2}} + mpq\bar{y}_{...}{}^{2} - 2mpq\bar{y}_{...} \frac{1}{p} \sum_{i=1}^{p} \bar{y}_{i..}$$

$$= \sum_{i=1}^{p} \frac{T_{i..}{}^{2}}{mq} + mpq\bar{y}_{...}{}^{2} - 2mpq\bar{y}_{...}{}^{2}$$

$$= \sum_{i=1}^{p} \frac{T_{i..}{}^{2}}{mq} - mpq\bar{y}_{...}{}^{2} = \sum_{i=1}^{p} \frac{T_{i..}{}^{2}}{mq} - \frac{T_{...}{}^{2}}{mpq} = \sum_{i=1}^{p} \frac{T_{i..}{}^{2}}{mq} - \ CF.$$

$$SSC = \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{k=1}^{m} \left(\bar{y}_{.j.} - \bar{y}_{...}\right)^{2} = mp \sum_{j=1}^{q} \left(\bar{y}_{.j.} - \bar{y}_{...}\right)^{2}$$

$$= mp \sum_{j=1}^{q} \bar{y}_{.j.}{}^{2} + mp \sum_{j=1}^{q} \bar{y}_{...}{}^{2} - \ 2mp \sum_{j=1}^{q} \bar{y}_{.j.}\bar{y}_{...}$$

$$= mp \sum_{j=1}^{q} \frac{T_{.j.}{}^{2}}{(mp)^{2}} + mpq\bar{y}_{...}{}^{2} - 2mpq\bar{y}_{...} \frac{1}{q} \sum_{j=1}^{q} \bar{y}_{.j.}$$

$$= \sum_{j=1}^{q} \frac{T_{.j.}{}^{2}}{mp} + mpq\bar{y}_{...}{}^{2} - 2mpq\bar{y}_{...}{}^{2}$$

$$= \sum_{j=1}^{q} \frac{T_{.j.}{}^{2}}{mp} - mpq\bar{y}_{...}{}^{2} = \sum_{j=1}^{q} \frac{T_{.j.}{}^{2}}{mp} - \frac{T_{...}{}^{2}}{mpq} = \sum_{j=1}^{q} \frac{T_{.j.}{}^{2}}{mp} - \ CF.$$

$$SSE = \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{k=1}^{m} \left(y_{ijk} - \ \bar{y}_{ij.}\right)^{2} = \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{k=1}^{m} y_{ijk}^{2} - \ \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{T_{ij.}^{2}}{m}$$

$$= RSS \ - \ \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{T_{ij.}^{2}}{m}$$

$$SSI = TSS - SSR - SSC - SSE.$$

6.4. ANOVA with missing observation

It happens many times in conducting the experiments that some observations are missed. This may happen due to several reasons. For example, in a clinical trial, suppose the readings of blood pressure are to be recorded after three days of giving the medicine to the patients. Suppose the medicine is given to 20 patients and one of the patients doesn't turn up for providing the blood pressure reading. Similarly, in an agricultural experiment, the seeds are sown and yields are to be recorded after few months. Suppose some cattle destroy the crop of any plot or the crop of any plot is destroyed due to storm, insects etc.

In such cases, one option is to somehow estimate the missing value on the basis of available data, replace it back in the data and make the data set complete.

Now conduct the statistical analysis on the basis of completed data set as if no value was missing by making necessary adjustments in the statistical tools to be applied. Such an area comes under the purview of “missing data models” and lot of development has taken place.

We discuss here the classical missing plot technique proposed by Yates which involve the following steps:

Estimate the missing observations by the values which make the error sum of squares to be minimum.

- Substitute the unknown values by the missing observations.

- Express the error sum of squares as a function of these unknown values.

- Minimize the error sum of squares using principle of maxima/minima, i.e., differentiating it with respect to the missing value and put it to zero and form a linear equation.

- Form as many linear equation as the number of unknown values (i.e., differentiate error sum of squares with respect to each unknown value).

- Solve all the linear equations simultaneously and solutions will provide the missing values.

- Impute the missing values with the estimated values and complete the data.

- Apply analysis of variance tools.

- The error sum of squares thus obtained is corrected but treatment sum of squares are not corrected.

• The number of degrees of freedom associated with the total sum of squares are subtracted by the number of missing values and adjusted in the error sum of squares. No change in the degrees of freedom of sum of squares due to treatment is needed.

6.4.1. Analysis of variance of an RBD with one missing observation

Let there be *t* treatments and r blocks in the given RBD. Let the missing observation correspond to the j^{th} treatment of rth block, *i.e.* we assume that the $(i,j)^{\text{th}}$ observation to be the missing one. Let *x* be the magnitude of the missing observation. Then the observation table is given by

Treatments **Blocks**	**1**	**2**	⋯	***j***	⋯	***t***	**Total**
1	y_{11}	y_{12}	⋯	y_{1j}	⋯	y_{1t}	$T_{1'.}=\sum_{j=1}^{t} y_{1j}$
2	y_{21}	y_{22}	⋯	y_{2j}	⋯	y_{2t}	$T_{2'.}=\sum_{j=1}^{t} y_{2j}$
⋮	⋮	⋮	⋯	⋮	⋯	⋮	⋮
i	y_{i1}	y_{i2}	⋯	$y_{ij}= x$	⋯	y_{it}	$T_{i.}+x=\sum_{j'(\neq j)=1}^{t} y_{ij'}+x$
⋮	⋮	⋮	⋯	⋮	⋯	⋮	⋮
r	y_{r1}	y_{r2}	⋯	y_{rj}	⋯	y_{rt}	$T_{r'.}=\sum_{j=1}^{t} y_{rj}$
Total	$T_{.1'}$ $=$ $\sum_{i=1}^{r} y_{i1}$	$T_{.2'}$ $=$ $\sum_{i=1}^{r} y_{i2}$	⋯	$T_{.j}+x=\sum_{i'(\neq i)=1}^{r} y_{i'j}+x$	⋯	$T_{.t'}$ $=$ $\sum_{i=1}^{r} y_{it}$	$T_{..}+\ x$ $=\sum_{i'(\neq i)=1}^{r} T_{i'.}+T_{i.}+x$ $=\sum_{j'(\neq j)=1}^{t} T_{.j'}+T_{.j}+x$

Let $T_{i'.}$ denote the block total of those blocks in which there is no missing observation and $T_{.j'}$ denote the treatment total of those treatments in which there is no missing observation. As the missing observation corresponds to the j^{th} treatment of i^{th} block, the block total and the treatment total of the corresponding block and treatment where the missing observation is occurring be denoted by $T_{i.}+x$ and $T_{.j}+x$ respectively. Also the overall total be denoted as $T_{..}+\ x$.

The mathematical model is then given by

$$y_{i'j'} = \mu + \alpha_{i'} + \beta_{j'} + e_{i'j'};$$

(1)

$[i', j'$denote that the model is for $rt - 1$ known observations] .

For the missing observation,

$$x = y_{ij} = \mu + \alpha_i + \beta_j + e_{ij},$$

(2)

where μ is the general mean, α_i is the additive effect due to i^{th} block, β_j is the additive effect due to j^{th} treatment and e_{ij}'s are random effects which are assumed to be *iid* random variables distributed according to $N(0, \sigma^2)$.

The unrestricted residual sum of squares is then given by

$$SSE = \sum_{i=1}^{r}\sum_{j=1}^{t} e_{ij}^2 = \sum_{i=1}^{r}\sum_{j=1}^{t}\left(y_{ij} - \mu - \alpha_i - \beta_j\right)^2.$$

Differentiating this with respect to μ, α_i and $\beta_j$individually, equating to zero and solving, we get the least square estimates of μ, α_i and β_j as

$$\hat{\mu} = \frac{T_{..} + x}{rt}, \hat{\alpha}_i = \frac{T_{i.} + x}{t} - \frac{T_{..} + x}{rt} \text{ and } \hat{\beta}_j = \frac{T_{.j} + x}{r} - \frac{T_{..} + x}{rt}.$$

Substituting this in the expression of x in (2) and assuming the estimate of e_{ij} to be zero, we get an estimate of the missing observation as

$$\hat{x} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j = \frac{T_{..} + \hat{x}}{rt} + \frac{T_{i.} + \hat{x}}{t} - \frac{T_{..} + \hat{x}}{rt} + \frac{T_{.j} + \hat{x}}{r} - \frac{T_{..} + \hat{x}}{rt}$$

$$= \frac{T_{i.} + \hat{x}}{t} + \frac{T_{.j} + \hat{x}}{r} - \frac{T_{..} + \hat{x}}{rt}.$$

Or $\hat{x}\left(1 - \frac{1}{t} - \frac{1}{r} + \frac{1}{rt}\right) = \frac{T_{i.}}{t} + \frac{T_{.j}}{r} - \frac{T_{..}}{rt}.$

Or $\hat{x}\left(\frac{rt - r - t + 1}{rt}\right) = \frac{rT_{i.} + tT_{.j} - T_{..}}{rt}$.

Or $\hat{x} = \frac{rT_{i.}+tT_{.j}-T_{..}}{(r-1)(t-1)}$.

Raw Sum of Squares (RSS) = $\sum_{i=1}^{r}\sum_{j=1}^{t} y_{ij}^2 = \sum_{i'=1}^{r}\sum_{j'=1}^{t} y_{i'\,j'}^2 + \hat{x}^2$.

Correction Factor (CF) = $\frac{\left(\sum_{i=1}^{r}\sum_{j=1}^{t} y_{ij}\right)^2}{rt} = \frac{(T_{..}+\hat{x})^2}{rt}$.

Total Sum of Squares (TSS) = RSS – CF = $\sum_{i'=1}^{r}\sum_{j'=1}^{t} y_{i'\,j'}^2 + \hat{x}^2 - \frac{(T_{..}+\hat{x})^2}{rt}$.

Sum of Squares due to Treatments (SST) (unadj.) = $\frac{\sum_{j'(\neq j)=1}^{t} T_{.j'}^2 + \left(T_{.j}+\hat{x}\right)^2}{r}$ – CF

$$= \frac{\sum_{j'(\neq j)=1}^{t} T_{.j'}^2 + \left(T_{.j}+\hat{x}\right)^2}{r} - \frac{(T_{..}+\hat{x})^2}{rt}.$$

Sum of Squares due to blocks (SSB) = $\frac{\sum_{i'(\neq i)=1}^{r} T_{i'.}^2 + (T_{i.}+\hat{x})^2}{t}$ – CF

$$= \frac{\sum_{i'(\neq i)=1}^{r} T_{i'.}^2 + (T_{i.}+\hat{x})^2}{t} - \frac{(T_{..}+\hat{x})^2}{rt}.$$

Sum of Squares due to error (SSE) = TSS – SST – SSB

$$= \sum_{i'=1}^{r}\sum_{j'=1}^{t} y_{i'\,j'}^2 + \hat{x}^2 - \frac{(T_{..}+\hat{x})^2}{rt} - \frac{\sum_{j'(\neq j)=1}^{t} T_{.j'}^2 + \left(T_{.j}+\hat{x}\right)^2}{r} + \frac{(T_{..}+\hat{x})^2}{rt}$$

$$- \frac{\sum_{i'(\neq i)=1}^{r} T_{i'.}^2 + (T_{i.}+\hat{x})^2}{t} + \frac{(T_{..}+\hat{x})^2}{rt}.$$

Or

SSE = $\sum_{i'=1}^{r}\sum_{j'=1}^{t} y_{i'\,j'}^2 + \hat{x}^2 - \frac{\sum_{j'(\neq j)=1}^{t} T_{.j'}^2 + \left(T_{.j}+\hat{x}\right)^2}{r} - \frac{\sum_{i'(\neq i)=1}^{r} T_{i'.}^2 + (T_{i.}+\hat{x})^2}{t} + \frac{(T_{..}+\hat{x})^2}{rt}$. (3)

Now we make the null hypothesis $\mathbf{H_0}$: The treatments do not differ significantly or the treatments are homogeneous. Then under this assumption there is no additive effect due to treatments, *i.e.* $\beta_j = 0 \;\forall\; j = 1,2,\cdots,t$.

Under the assumption that $\mathbf{H_0}$ is true the mathematical model will be

$y_{i'j'} = \mu + \alpha_{i'} + e_{i'j'}$;

(4)

[i', j'denote that the model is for rt – 1 known observations].

For the missing observation,

$x' = \mu + \alpha_i + e_{ij}$.

(5)

The restricted residual sum of squares under the restriction that $\mathbf{H_0}$ is true is then given by

$$SSE^* = \sum_{i=1}^{r} \sum_{j=1}^{t} e_{ij}^2 = \sum_{i=1}^{r} \sum_{j=1}^{t} \left(y_{ij} - \mu - \alpha_i\right)^2.$$

Differentiating this with respect to μ and $\alpha_i$individually, equating to zero and solving, we get the least square estimates of μ and α_i as

$$\hat{\mu} = \frac{T_{..} + x}{rt} \text{ and } \hat{\alpha}_i = \frac{T_{i.} + x}{t} - \frac{T_{..} + x}{rt}.$$

Substituting this in the expression of x in (2) and assuming the estimate of e_{ij} to be zero, we get an estimate of the missing observation as

$$\hat{x}' = \hat{\mu} + \hat{\alpha}_i = \frac{T_{..} + \hat{x}'}{rt} + \frac{T_{i.} + \hat{x}'}{t} - \frac{T_{..} + \hat{x}'}{rt} \quad = \frac{T_{i.} + \hat{x}'}{t} ‘$$

Or $\hat{x}' \left(1 - \frac{1}{t}\right) = \frac{T_{i.}}{t}$

Or $\hat{x}' \left(\frac{t-1}{t}\right) = \frac{T_{i.}}{t}$.

Or $\hat{x}' = \frac{T_{i.}}{(t-1)}$.

Raw Sum of Squares (RSS) = $\sum_{i=1}^{r} \sum_{j=1}^{t} y_{ij}^2 = \sum_{i'=1}^{r} \sum_{j'=1}^{t} y_{i'j'}^2 + \hat{x}'^2$.

Correction Factor (CF) = $\frac{\left(\sum_{i=1}^{r} \sum_{j=1}^{t} y_{ij}\right)^2}{rt} = \frac{\left(T_{..} + \hat{x}'\right)^2}{rt}$.

Total Sum of Squares (TSS) = RSS – CF = $\sum_{i'=1}^{r}\sum_{j'=1}^{t} y_{i'j'}^{2} + \hat{x}'^{2} - \frac{(T_{..}+\hat{x}')^{2}}{rt}$.

Sum of Squares due to blocks (SSB) = $\frac{\sum_{i'(\neq i)=1}^{r} T_{i'.}^{2} + (T_{i.}+\hat{x}')^{2}}{t}$ – CF

$$= \frac{\sum_{i'(\neq i)=1}^{r} T_{i'.}^{2} + (T_{i.}+\hat{x}')^{2}}{t} - \frac{(T_{..}+\hat{x}')^{2}}{rt}.$$

Sum of Squares due to error (SSE) = TSS – SSB

$$= \sum_{i'=1}^{r}\sum_{j'=1}^{t} y_{i'j'}^{2} + \hat{x}'^{2} - \frac{(T_{..}+\hat{x}')^{2}}{rt} \quad - \frac{\sum_{i'(\neq i)=1}^{r} T_{i'.}^{2} + (T_{i.}+\hat{x}')^{2}}{t} + \frac{(T_{..}+\hat{x}')^{2}}{rt}.$$

Or $SSE^{*} = \sum_{i'=1}^{r}\sum_{j'=1}^{t} y_{i'j'}^{2} + \hat{x}'^{2} - \frac{\sum_{i'(\neq i)=1}^{r} T_{i'.}^{2} + (T_{i.}+\hat{x}')^{2}}{t}$.

(6)

Hence, SST (adj.) = $SSE^{*} - SSE$

$$= \sum_{i'=1}^{r}\sum_{j'=1}^{t} y_{i'j'}^{2} + \hat{x}'^{2} - \frac{\sum_{i'(\neq i)=1}^{r} T_{i'.}^{2} + (T_{i.}+\hat{x}')^{2}}{t}$$

$$- \sum_{i'=1}^{r}\sum_{j'=1}^{t} y_{i'j'}^{2} - \hat{x}^{2} + \frac{\sum_{j'(\neq j)=1}^{t} T_{.j'}^{2} + (T_{.j}+\hat{x})^{2}}{r} + \frac{\sum_{i'(\neq i)=1}^{r} T_{i'.}^{2} + (T_{i.}+\hat{x})^{2}}{t} -$$

$$\frac{(T_{..}+\hat{x})^{2}}{rt}$$

$$= \left[\frac{\sum_{j'(\neq j)=1}^{t} T_{.j'}^{2} + (T_{.j}+\hat{x})^{2}}{r} - \frac{(T_{..}+\hat{x})^{2}}{rt}\right] - \left[\hat{x}^{2} - \hat{x}'^{2} + \frac{(T_{i.}+\hat{x}')^{2}}{t} - \frac{(T_{i.}+\hat{x})^{2}}{t}\right]$$

= SST(unadj.) – Bias ,

Where Bias $= \hat{x}^{2} - \hat{x}'^{2} + \frac{(T_{i.}+\hat{x}')^{2}}{t} - \frac{(T_{i.}+\hat{x})^{2}}{t}$

$$= \hat{x}^{2} - \hat{x}'^{2} + \frac{1}{t}\left[(T_{i.}+\hat{x}')^{2} - (T_{i.}+\hat{x})^{2}\right]$$

$$= \hat{x}^{2} - \hat{x}'^{2} + \frac{1}{t}\left[T_{i.}^{2} + \hat{x}'^{2} + 2T_{i.}\hat{x}' - T_{i.}^{2} - \hat{x}^{2} - 2T_{i.}\hat{x}\right]$$

= $\hat{x}^{2} - \hat{x}'^{2} + \frac{1}{t}\left[\hat{x}'^{2} + 2(t-1)\hat{x}'^{2} - \hat{x}^{2} - 2(t-1)\hat{x}'\hat{x}\right]$ [since $\hat{x}' = \frac{T_{i.}}{(t-1)}$]

$$= \frac{1}{t}\left[t\hat{x}^2 - \hat{x}^2 - t\hat{x}'^2 + \hat{x}'^2 + 2(t-1)\hat{x}'^2 - 2(t-1)\hat{x}'\hat{x}\right]$$

$$= \frac{1}{t}\left[(t-1)\hat{x}^2 - (t-1)\hat{x}'^2 + 2(t-1)\hat{x}'^2 - 2(t-1)\hat{x}'\hat{x}\right]$$

$$= \frac{(t-1)}{t}\left[\hat{x}^2 + \hat{x}'^2 - 2\hat{x}'\hat{x}\right] = \frac{(t-1)}{t}(\hat{x} - \hat{x}')^2 > 0.$$

Hence the bias is always positive.

ANOVA Table for RBD with one missing observation

Sources of Variation	Degrees of freedom	Sum of Squares	Mean Sum of Squares	Variance ratio
Treatments (adjusted)	t – 1	S.S.T.(adj.)= $t\sum_{i=1}^{r}(\bar{y}_{i.} - \bar{y}_{..})^2$- Bias	M.S.T = $\frac{S.S.T.}{t-1}$	$F_T = \frac{M.S.T.}{M.S.E}$
Blocks	r – 1	S.S.B.= $r\ \sum_{j=1}^{t}\left(\bar{y}_{.j} - \bar{y}_{..}\right)^2$	M.S.B = $\frac{S.S.B.}{r-1}$	$F_B = \frac{M.S.B.}{M.S.E.}$
Error	(r – 1)(t – 1) – 1	S.S.E.=$\sum_{i=1}^{r}\sum_{j=1}^{t}\left(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}\right)^2$	M.S.E. = $\frac{S.S.E.}{(t-1)(r-1)-1}$	
Total	rt – 2	T.S.S. =$\sum_{i=1}^{r}\sum_{j=1}^{t}\left(y_{ij} - \bar{y}_{..}\right)^2$		

6.4.2. Analysis of variance of an RBD with two missing observation

For Two missing values say x and y. let R_1 and R_2 be the totals of known observations in the row containing x and y respectively and C_1 and C_2 be the totals of known observations in the Columns containing x and y respectively. And let S be the total of all the known observations

Treatments / **Blocks**	**1**	**2**	⋯	***j***	⋯	***q***	***t***	**Total**
1								
2					x			R_1
⋮							⋮	⋮
i			⋯					
p								

⋮		y	…					R_2
r			…					
Total		C_2			C_1			

The error sum of square E= $x^2 + y^2 \frac{1}{t}[(R_1 + x)^2 + (R_2 + y)^2] - \frac{1}{r}[(C_1 + x)^2 + (C_2 + y)^2] + \frac{1}{rt}(S + x + y)^2$+ (terms independent of x and y)

For minimizing of E subject to variation in x and y, we must have

$$\frac{dE}{dx} = 0$$

=$x - \frac{1}{t}[(R_1 + x)] - \frac{1}{r}[(C_1 + x) + \frac{1}{rt}(S + x + y)$

$$\frac{dE}{dx} = 0$$

=$y - \frac{1}{t}[(R_2 + y)] - \frac{1}{r}[(C_2 + y) + \frac{1}{rt}(S + x + y)$

(r-1)(t-1)x=r$R_1 + tC_1 - S - y$

(r-1)(t-1)y=r$R_2 + tC_2 - S - x$

Solving these equations simultaneously we get the estimates of x and y

6.5. Analysis of covariance

The analysis of covariance (ANCOVA) is a technique that is occasionally useful for improving the precision of an experiment. Suppose that in an experiment with a response variable Y, there is another variable X, such that Y is linearly related to X. Furthermore, suppose that the researcher cannot control X but can observe it along with Y. Such a variable X is called a covariate or a concomitant variable. The basic idea underlying ANCOVA is that precision in detecting the effects of treatments on Y can be increased by adjusting the observed values of Y for the effect of the concomitant variable. If such adjustments are not performed, the concomitant variable X could inflate the error mean square and make true differences in the response due to treatments harder to detect. The concept is very similar to the use of blocks to reduce the experimental error. However, when the blocking variable is a continuous variable, the delimitation of the blocks can be very subjective. The ANCOVA uses information about in X in two ways:

1. Variation in Y that is associated with variation in X is removed from the error variance (MSE), resulting in more precise estimates and more powerful tests

2. Individual observations of Y are adjusted to correspond to a common value of X, thereby producing group means that are not biased by X, as well as equitable group comparisons.

A sort of hybrid of ANOVA and linear regression analysis, ANCOVA is a method of adjusting for the effects of an uncontrollable nuisance variable.

Any scientific experiment is performed to know something that is unknown about a group of treatments and to test certain hypothesis about the corresponding treatment effect. When variability of experimental units is small relative to the treatment differences and the experimenter do not wishes to use experimental design, then just take large number of observations on each treatment effect and compute its mean. The variation around mean can be made as small as desired by taking more observations. When there is considerable variation among observations on the same treatment and it is not possible to take an unlimited number of observations, the techniques used for reducing the variation are

(i) use of proper experimental design

(ii) Use of concomitant variables.

The use of concomitant variables is accomplished through the technique of analysis of covariance. If both the techniques fail to control the experimental variability then the number of replications of different treatments (in other words, the number of experimental units) is needed to be increased to a point where adequate control of variability is attained.

6.6. Summary

Analysis of variance (ANOVA) is used for determine whether or not the means of more than two populations are equal. Analysis of variance is large area of application of statistics developed by Prof. R.A. Fisher. The term “analysis of variance” is used because the total variability in the set of data can be broken into the sum of variability among the sample means and the variability within the samples. This procedure is based on the question- is there significantly more variation among the group means than there is within the groups. The pooled variation within groups is used as a standard of comparison, because it measures the inherent observational variabilityin the data. Usually the objective or hypotheses of the experiments are to make comparison among the effects of several treatments. If the number of treatments is more than two, then analysis of variance technique is mainly adopted for the analysis and introduction of the observations, collected from an experiment, to make comparisons among the effect of the treatments.

6.7. Terminal questions

Four groups of patients were subjected to the different treatments for a particular disease. At the end of a specified time period each was given a test to measure treatment effectiveness. the scores were obtained. Conduct appropriate test to find, if there is any significant difference in the treatments. Also give the ANOVA table.

A	B	C	D
60	76	58	95
88	70	74	90
70	80	66	80
80	90	60	87
76	75	82	88
70	79	75	85

Conduct appropriate test to find, if there is any significant
difference in thetreatment. (F = -2.6, F_a = 4.26)

2. The following table gives the gains in weights of four different
types of pigs fed on three different ratios. Test whether the

rations or the pig types differ in their on mean weight.

Types of pigs

Types of Rations		I	II	III	IV
	I	7	16	10	11
	II	15	14	15	14
	III	8	16	7	11

6.8. Further suggested readings

1. Alok Dey (1986): Theory of Block Designs Wiley Eastern.
2. Das M.N. and Giri N. (1979): Design and analysis of Experiments.
3. Joshi D.D. (1987): Linear Estimation and Design of Experiments Wiley Eastern.
4. Goon, Gupta & Dasgupta : Fundamentals of Statistics Vol. I and Vol. IIThe World Press Pvt. Ltd., Kolkata.

Rajarshi Tandon Open

University, Prayagraj

Numerical
and
Statistical Computing

Block- 3

Introduction to Computer

Unit-7

Computer Basics

Unit-8

Hardware and Software

Unit-9

MS Office

Rajarshi Tandon Open

University, Prayagraj

Numerical and Statistical Computing

Course Design Committee

Prof. Ashutosh Gupta **Chairman**
School of Science, UPRTOU, Prayagraj

Dr. Uma Rani Agarwal **Member**
Rtd. Professor, Department of Botany
CMP Degree College, Prayagraj

Dr. Ayodhaya Prasad Verma **Member**
Red. Professor, Department of Botany
B.S.N.V. P.G. College, Lucknow

Dr. Sudhir Kumar Singh **Member**
Assistant Professor
K. Banerjee Centre for Atmospheric and Ocean Studies
University of Allahabad, Prayagraj

Dr. Ravindra Pratap Singh **Member**
Assistant Professor (Biochemistry)
School of Science, UPRTOU, Prayagraj

Dr. Dharmveer Singh **Course Coordinator**
Assistant Professor (Biochemistry)
School of Science, UPRTOU, Prayagraj

Course Preparation Committee

Dr. Anuj Kumar Singh **Author** **Block-1** (Unit: 1)
Assistant Prof. (Statistics)
School of Sciences, UPRTOU, Prayagraj

Dr. Upasana Singh **Author** **Block-1&2** (Unit: 2-6)
Assistant Professor-Zoology
Prof. Rajendra Singh Rajju Bhaiya
University, Prayagraj

Dr. Jaspal Singh **Author** **Block-1&3,4** (Unit: 1,7,8,9,10,11)
Assistant Professor
Department of Environmental Science,
Bareilly College, Bareilly

Dr. Nishtha Seth **Author** **(All blocks and units)**
Associate Professor
Department of Environmental Science,
Bareilly College, Bareilly

Dr. Dharmveer Singh
(Course and SLM Coordinator)
School of Sciences, UPRTOU, Prayagraj

Introduction

This third block of numerical and statistical computing, this consists of following three units:

Unit-7: This unit covers the features of computers, computer generation, and using computers to solve categorization problems.

Unit-8: The central processing unit (CPU), memory organization, input and output devices, system software, file commands, and editing commands are all covered in this unit.

Unit-9: This unit covers the word processing software like microsoft office word and spread sheet software like microsoft office excel and also presentation software such as microsoft office-power point, excel as data base software.

Unit-7: Computers Basics

Contents

7.1. Introduction

Everybody knows how to calculate in their daily lives. For computations, we use numerous formulas and mathematical operations like addition, subtraction, multiplication, and so forth. While calculations that are simpler require less time, those that are more complex take much longer. Furthermore, we occasionally are unable to complete all of the calculations by hand. The correctness of the results is another consideration that emerges from these laborious computations. Because necessity is the mother of invention, man set out to create a machine in antiquity that could complete these kinds of arithmetic operations more quickly and accurately. As a result, the "computer" as a machine or equipment was created.

Although the original goal of inventing the computer was to create a fast computing machine, the computer we see now is very different from the one created in the beginning. The number of computer applications has grown, as have the speed and accuracy with which calculations are performed. The majority of work done by computers today is non-numerical. Computer applications include, for example, airline and railway ticket reservations, telephone and power bill payments, commercial data processing, medical diagnostics, weather forecasting, and others. A computer can be described more precisely as a device that processes information or data. The data might have various sizes and shapes depending on the application. People started calling it a 'data processor' because of its ability to process data.

Objectives

After the study of this unit you will be in a position to

- Computer Definition and Characteristics
- Understanding computer origin and evolution.
- Identifying computer's speed and accuracy capabilities.
- Appreciating computer evolution through five generations.
- Classifying computers based on performance.

7.2. Computer overview

Computer is a machine capable of solving problems and manipulating data. It accepts data, processes the data by doing some mathematical and logical operations and gives us the desired output. However, the computer is the

- Electronic device capable of faster and accurate arithmetic calculations.
- Comparable to a magic box serving different purposes.
- Machine capable of solving problems and manipulating data.
- Accepts data, processes it through mathematical and logical operations.
- Transforms data into desired outputs.
- Data can include student marks, income, savings, investments, etc.

Thus, a computer might be defined as a device that modifies data. Any type of information can be considered data, including a state's income, savings, investments, and test results for individual students. Consequently, a computer data processing overview is

- Accepting data
- Storing data
- Processing data as needed
- Retrieving stored data
- Printing results in desired format

7.3. Characteristics of Computer

A computer is simply a machine that executes instructions. It cannot accomplish anything without the user's instructions. It executes instructions quickly and accurately. Thus, the user must pick what he wants to perform and with what level of accuracy. In other words, the machine cannot make its own decisions like we can. The major characteristics of computer are as follows:

Speed: Computers are capable of carrying out billions of operations per second at astonishing rates. They can effectively perform complicated calculations and operations thanks to their quick processing speed. For instance, the daily weather forecasts we watch on TV are the outcome of computers compiling and analyzing enormous amounts of data on pressure, temperature, humidity, and other variables from different locations.

Accuracy: Computers are extremely accurate at executing instructions and calculations, provided that the instructions are properly programmed. They can complete repetitive operations without errors, making them dependable for precision-based activities. For example, if we want our output to be accurate to 12 decimal places, it will be tough to do manually, but simple for a machine. The accuracy level is determined by the design of the computer. Computer errors are caused by human error or faulty data.

Diligence: Computers work with digital data in binary format (0s and 1s). They handle this data with algorithms and logical processes, allowing them to efficiently complete a wide range of computing tasks. It's fascinating that the computer is devoid of lethargy, loss of attention, and fatigue. It can function for hours or days without producing any errors. If billions of computations need to be performed, a computer will perform them all with the same accuracy.

Versatility: Computers can execute a broad variety of activities, including simple arithmetic calculations, complicated simulations, and data processing. They can be designed to adapt to different requirements, making them useful tools for a wide range of applications. It refers to the ability to undertake completely diverse forms of job. As we explained in the introduction, we may utilize our computers for a variety of tasks.

Storage: Computers can store large amounts of data in a variety of formats, including text, photos, videos, and more. This information can be immediately accessed and retrieved when needed. The computer contains built-in memory that can store vast amounts of data. We can also store data in secondary storage devices such as floppies, which can be retained outside of the computer and transported to another computer. The computer has the ability to store any amount of information or data. Any information can be saved and recalled for as long as we need it, whether it's a few days or several years.

Automation: Computers can automate operations, eliminating the need for human intervention in repetitive procedures. This automated potential increases efficiency and production in a variety of sectors.

Connectivity: Computers can link to networks, allowing many devices and people to communicate and share data. This connectivity makes it easier to collaborate, share resources, and access information from faraway locations.

Scalability: Computers may increase or decrease processing power, storage capacity, and other resources to meet changing needs. This scalability enables flexible and efficient use of computational **resources.**

Multitasking: Multitasking is a feature of modern computers that allows them to run numerous processes or programs at once. This feature allows users to accomplish multiple things simultaneously, increasing productivity and usefulness.

Reliability: Computers are highly reliable machines that can operate constantly for long periods of time without failure if they are properly maintained. However, as with any sophisticated system, they require regular maintenance and updates to ensure peak performance and security.

7.4. Historical Evaluation of Computer

The history of computers is a fascinating journey punctuated by key milestones and dramatic advances. The origins of computers can be traced back to man's search for a device capable of performing fast computations with high accuracy. We shall quickly explore several ground-breaking inventions in the world of computing devices.

Pre-computer Era (pre-20th century): The concept of computing stretches back to ancient cultures, when devices such as the abacus were utilized to perform arithmetic computations. Other early computational machines include the Antikythera mechanism, which was an ancient Greek analog computer used for astronomical calculations.

Mechanical Computers (17th to 19th centuries): Mechanical calculators like Blaise Pascal's Pascaline (17th century) and Charles Babbage's Analytical Engine (19th century) paved the way for modern computing. Babbage's innovations, in particular, are regarded as the earliest antecedents of programmable computers.

Early Electronic Computers (20th century):

- **First generation (1940s-1950s):** Early electronic computers from the 20th century include the first generation. The development of electronic computers began with machines such as the ENIAC (Electronic Numerical Integrator and Computer) and the Colossus, which were primarily employed for military and scientific applications. These computers were big, heavy, and relied on vacuum tubes for processing.
- **Second Generation (1950s-1960s):** The arrival of transistors signaled the second generation of computers, which resulted in smaller, quicker, and more dependable machines. Examples include the IBM 1401 and UNIVAC 1107.

- **Third Generation (1960s-1970s):** Integrated circuits transformed computing, allowing for the creation of smaller, more powerful, and less expensive computers. During this time, the use of mainframes and minicomputers increased.

Fourth Generation (1970s–present): The microprocessor, introduced in the early 1970s, laid the groundwork for the fourth generation of computers. This period saw the proliferation of personal computers (PCs) and the democratization of computing power. Companies such as Apple and IBM played important roles in popularizing computers.

Modern Era:

- **Calculating Machines:** It took years for early humans to create mechanical devices for counting big numbers. The Egyptians and Chinese built the earliest calculating instrument, known as ABACUS. The name ABACUS refers to a calculating board. It contains a number of horizontal bars, each containing ten beads. Horizontal bars signify units (tens, hundreds, etc.). Nepier's bones: In 1617 AD, English mathematician John Napier created a mechanical apparatus for multiplication.
- **Slide Rule:** Edmund Gunter, an English mathematician, devised the slide rule. This machine could carry out operations such as addition, subtraction, multiplication, and division. It was commonly employed throughout Europe in the sixteenth century.
- **Pascal’s Adding and Subtracting Machine**: Basic Pascal created a machine that could add and subtract. The machine consisted of wheels, gears, and cylinders.
- **Leibniz's Multiplication and Division Machine:** In 1673, German mathematician Gottfried Leibniz created a mechanical device capable of multiplication and division.
- **Babbage's Analytical Engine**: Charles Babbage is widely regarded as the pioneer of computing. In 1823, he created a mechanical mechanism for doing complex mathematical calculations. It was named the Difference Engine. Following that, he created a general-purpose calculating machine known as the analytical engine.
- **Mechanical and electrical calculators:** It was invented in the early nineteenth century to conduct various mathematical calculations, and they were frequently used

until the 1960s. Later, the spinning portion was replaced with an electric motor. After that, it was referred to as the electrical calculator.

- **Modern Electronic Calculator:** The electronic calculator used in the 1960s was powered by electron tubes, which were rather large. Later, it was replaced with transistors, resulting in calculators that were too small. Modern electronic calculators can compute a wide range of mathematical computations and functions. It can also be used to save data permanently. Some calculators include built-in programs for performing complex calculations.
- **Mobile Computing:** The widespread use of smartphones and tablets has altered computing, making it more accessible and pervasive.
- **Cloud Computing**: Cloud computing has altered the way computing resources are provisioned, allowing for scalable and cost-effective solutions for organizations and individuals.
- **Artificial Intelligence and Machine Learning**: Recent advances in AI and machine learning have opened up new opportunities for data analysis, automation, and decision making.

7.5. Computer Generations

The modern computer has undergone remarkable change during the last five decades. This period of computer evolution is separated into five distinct phases known as Computer Generations. Each generation of computer is distinguished by significant technological advancements that fundamentally alter the way computers function, resulting in increasingly smaller, cheaper, more powerful, efficient, and dependable machines. Here we cover the many generations and the developments that lead to the current devices.

A. First Generation Computers [1940-1956]

The first generation of computers appeared in the 1940s, with the introduction of electronic computers such as the ENIAC and the Colossus. The first computers utilized vacuum tubes for circuitry and magnetic drums for memory. These computers were vast in size, and programming on them was challenging. They were extremely expensive to run and, in addition to consuming a large amount of electricity, produced a lot of heat, which was frequently the source of failures. First-generation computers used machine language to accomplish operations and were limited to solving one problem at a time. Input was done via punched cards and paper tape, while output was displayed on printouts. Despite their massive size and limited capabilities, they marked a significant advancement in computing technology. Some computers from this generation were:

ENIAC (Electronic Numerical Integrator and Calculator)

John Mauchly and J. Presper Eckert constructed the Electronic Numerical Integrator and Computer (ENIAC), the world's first general-purpose electronic digital computer, during World War II. ENIAC, completed in 1945, weighed 30 tons and took up an entire room. It used more than 17,000 vacuum tubes to do calculations at extraordinary rates, which aided military calculations such as artillery firing tables. ENIAC could conduct a wide range of numerical computations and was configurable for many purposes, establishing the groundwork for modern computing. Despite its size and complexity, ENIAC was a significant step forward in computing technology, paving the way for following advances.

EDV AC: (Electronic Discrete Variable Automatic Computer)

The Electronic Discrete Variable Automatic Computer (EDVAC) was one of the first stored-program computers, created in the late 1940s by John von Neumann and his associates. Following the construction of the ENIAC, von Neumann introduced the idea of a computer with a stored program, in which both data and instructions are kept in the machine's memory. This principle was reflected by EDVAC's design, which allowed instructions to be stored alongside data in memory, resulting in more flexible and efficient computing. Although EDVAC was not completed until after the ENIAC, its design greatly influenced succeeding computer architectures, laying the groundwork for contemporary computing techniques.

EDSAC (Electronic Delay Storage Automatic Computer)

The Electronic Delay Storage Automatic Computer (EDSAC) was a pioneering and significant computer developed at the University of Cambridge in the late 1940s and early 1950s. EDSAC, designed by Maurice Wilkes and his team, was the world's first practical stored-program computer, capable of storing both program instructions and data in memory. EDSAC used mercury delay lines for memory, with sound waves used to store binary data. It was largely used for scientific calculations and paved the way for future advances in computer design and programming. The success of EDSAC stimulated subsequent computer research and improvements around the world.

UNIVAC-I: Eckert and Mauchly created it in 1951 with the Universal Accounting Computer configuration. Power consumption was really high. It required a huge area for installation and an air conditioning chamber. The programming capabilities were relatively limited, and it was not portable.

B. Second Generation Computers [1956-1963]

The second generation of computers saw the advent of transistors, which replaced vacuum tubes. Transistors were smaller, more dependable, and energy-efficient than vacuum tubes, resulting in major increases in computer size, speed, and reliability. During this time, mainframe computers such as the IBM 1401 and the UNIVAC 1107 became increasingly common, making it easier to process business data and conduct scientific research. The introduction of magnetic core memory increased storage capacity and speed. Second-generation computers transitioned from binary machine language to symbolic or assembly languages, allowing programmers to define commands in words. At the same time, high-level programming languages such as early COBOL and FORTRAN were being developed. These were also the first computers to store their instructions in memory, transitioning from magnetic drum to magnetic core technology. The second generation introduced the Central Processing Unit (CPU), memory, programming language, and input and output modules. Some computers from the second generation were

- **IBM 1620:** The IBM 1620 was primarily utilized for scientific applications and had a smaller size than First Generation computers.

- **IBM 1401:** IBM 1401 was a small or medium-sized computer designed for corporate use.
- **CDC 3600:** CDC 3600 was utilized for scientific reasons.

The major advantages of the computer of this generation were:

1. Smaller in size than the first generation computers
2. Less heat generated.
3. Faster than the first generation computers
4. Less prone to hardware failures, etc.

Drawbacks:

1. Air conditioned rooms were required.
2. Frequent maintenance required
3. Commercial production was difficult and costly.

C. Third Generation [1964-1971]

The third generation of computers was distinguished by the development of integrated circuits, which merged many transistors onto a single semiconductor chip. This discovery, pioneered by businesses like as Texas Instruments and Fairchild Semiconductor, resulted in further reductions in size and cost while increasing computing capability dramatically. Mainframes and minicomputers became more economical and accessible, allowing for widespread use across industries and organizations. During this time, time-sharing systems were popular, allowing numerous people to access a single computer at the same time.

The IBM-360, ICL-1900, IBM-370, and V AX-750 were among the computers developed during this time period. This decade saw the development of higher-level languages such as BASIC (Basic All Purpose Symbolic Instruction Code). This generation's computers were tiny in size, affordable in cost, had plenty of memory space, and had a highly fast processor. This generation's computers had the following main advantages: Smaller in size than the computers of previous generations.

1. Computational timings were further reduced in comparison generation computers.
2. Easily portable
3. Commercial production was also easier and cheaper
4. There were no major drawbacks of the computers of these generations except that sophisticated technology was required for the manufacturing of IC chips.

D. Fourth Generation [1971-Present]

The fourth generation of computers began in the early 1970s with the development of the microprocessor, which combined the central processing unit (CPU) on a single chip. This innovation, spearheaded by businesses such as Intel and Motorola, transformed computing by enabling the creation of smaller, more powerful, and less expensive computers. Personal computers (PCs) became widely available, democratizing computing power and propelling the software industry forward. Graphical user interfaces (GUIs) and operating systems such as MS-DOS and Apple's Macintosh OS improved computer usability and accessibility for non-technical users.

IBM debuted its first home computer in 1981, followed by Apple's Macintosh in 1984. Microprocessors have also spread beyond desktop computers and into many other aspects of life, as more and more ordinary devices use them. As these little computers got more powerful, they could be joined together to establish networks, resulting in the creation of the Internet.

The major advantages of the computers of this generation were:

- **Miniaturization:** Smaller, more compact computers for portability and space efficiency.
- **Increased Processing Power**: Microprocessors boost computing power, enabling faster, complex tasks.
- **Cost-Effectiveness:** Reduced manufacturing costs make computers more affordable and accessible.
- **Energy Efficiency:** Lower power consumption, reducing energy costs and environmental impact.

- **Versatility:** Can run a wide range of software applications, suitable for business, scientific research, and personal use.
- **User-Friendly Interfaces**: Graphical user interfaces enhance usability and user interaction
- **Connectivity:** Networking capabilities facilitate internet connectivity and information sharing.
- **Integration:** Can integrate with other technologies, enhancing productivity.
- **Multitasking**: Can run multiple programs simultaneously, increasing efficiency.
- **Scalability:** Can be easily upgraded for flexibility and longevity.

E. Fifth Generation [Present and Beyond]

Fifth-generation computing devices, based on artificial intelligence, are still in development phase, with applications like voice recognition already in use. These devices are expected to have high speed and can perform parallel processing. Quantum computation and molecular and nanotechnology are expected to revolutionize computing in the future. The goal is to develop devices that respond to natural language input, learn, and self-organize, making artificial intelligence a reality.

Advances in artificial intelligence, quantum computing, and bioinformatics are transforming computer processing and analysis. Artificial intelligence (AI) technologies such as machine learning and neural networks enable applications in natural language processing, computer vision, and autonomous systems. Quantum computing provides quicker computation, whereas quantum mechanics principles promise exponentially faster processing. Bioinformatics, which combines biology and computer science, can provide fresh insights into genetics, drug discovery, and customized medicine.

In conclusion, the evolution of computers has been characterized by consecutive generations of technological progress, each building on the achievements of the previous one. Computers have evolved from room-sized machines with limited capabilities to handheld devices with enormous processing power, becoming crucial instruments that shape practically every area of modern life.

7.6. Classification of Computers

The various types of computers available today can be classified based on their size, efficiency, memory, and number of users. They can be broadly categorized into the following categories.

A. **Based on Size**:

- **Supercomputers:** Supercomputers are extremely powerful computers meant to execute complex computations and simulations. They are utilized for scientific research, weather forecasting, and other high-performance computing applications. These are the quickest and most expensive devices utilized to meet the needs of any enterprise with a large volume of data processing. They have a higher processing speed than other computers. They also use multiprocessing techniques. These computers' applications include forecasting, biological research, remote sensing, and aircraft design. Supercomputers include CRA Y YMP, CRA Y2, NEC SX-3, CRA Y XMP, and PARAM from India.
- **Mainframe computers:** Mainframe computers are powerful machines that can handle large amounts of data and serve several users at once. They are utilized in business-critical applications like banking, finance, and large-scale data processing. These computers function at high speeds, have a vast storage capacity, and can handle the workload of multiple users at once. These microprocessors are typically 32-bit and are utilized in centralized databases. They also serve as controlling nodes in Wide Area Networks (WAN). Examples of mainframes include the DEC, ICL, and IBM 3000 series.
- **Minicomputers:** Minicomputers are mid-sized computers with moderate processing capability and the ability to serve many users simultaneously. They are widely utilized in small and medium-sized enterprises for duties such as database management and network administration. This computer is superior to a microcomputer. It is utilized in a multi-user system, which allows multiple people to operate simultaneously. It has a large storage capacity and performs faster. This sort of computer is commonly used to process enormous amounts of data in a variety of companies, including scientific and technology institutions. They can also be utilized as servers in local area networks (LANs).

- **Microcomputers:** Also known as personal computers (PCs), these are small, single-user computers designed for individual use. They come in various forms such as desktops, laptops, tablets, and smartphones. Microcomputers are small and lightweight, making them convenient to transport and utilize in various settings. Laptops and tablets are highly portable, allowing users to work or access information on the go. Microcomputers are often less expensive than bigger computer systems such as mainframes and minicomputers, making them available to a diverse variety of users, including people, small enterprises, and educational institutions.

B. Based on Purpose:

- **General-Purpose Computers**: Versatile computers capable of performing a wide range of tasks and running various software applications. Personal computers fall into this category.
- Special-Purpose Computers: Designed for specific applications or tasks. Examples include embedded systems in devices like digital cameras, ATMs, and control systems for industrial machinery.

C. Based on Architecture:

- **RISC (Reduced Instruction Set Computer**): Computers with a simplified instruction set and optimized for executing a small set of instructions quickly. They are commonly used in mobile devices and high-performance computing.
- **CISC (Complex Instruction Set Computer):** Computers with a complex instruction set capable of executing a wide variety of instructions. They are often found in desktop and server environments.

D. Based on Processing Capabilities:

- **Analog Computers:** Process continuous data and perform calculations based on physical quantities such as voltage, current, and temperature. They are used in applications such as signal processing and control systems.
- **Digital Computers:** Process discrete data represented in binary form (0s and 1s). They are the most common type of computers and are used in virtually all modern computing applications.

- **Hybrid Computers**: Combine the features of analog and digital computers to leverage the strengths of both. They are used in applications such as real-time control systems and scientific simulations.

7.7. Number System

Any system for representing numerical values or quantities uses a number of digits, such as the decimal system, which uses ten digits from 0 to 9. These digits can be grouped in groups, with each digit's contribution determined by its value and position in the group.

Decimal Notation: A method of representing numbers in which successive digit locations are represented by powers of 10.

Examples: 6235 means

1000s(10^3)	100s(10^2)	10s(10^1)	1s(10^0)	
6	2	3	5	
=1000×6	+100×2	+10×3	+1×5	
=6000	+200	+30	+5	=6235

Binary Notation A positional notation scheme for representing integers with a base or radix of two. In this method, numbers are represented by two digits, 0 and 1, with each digit representing a power of two.

Examples: 1010_2 is 10 in decimal system as shown below:

8s(2^3)	4s(2^2)	2s(2^1)	1s(2^0)	
1	0	1	0	
=8×1	+4×0	+2×1	0+1×0	
=8	+0	+2	+0	=10

Octal Notation: The number system using 8 as base or radix. This system uses the digits from 0 to 7 and each digit position represents a power of 8.

Example: 232_8 is 154 in decimal system as shown below:

8s(8^2)	8s(8^1)	1s(8^0)	
2	3	2	
+64×2	+8×3	+1×2	
=128	+24	+2	=154

Hexadecimal Notation: A notation for numbers using 16 as the base or radix. Ten decimal digits ranging from 0 to 9 are used, together with six additional characters a, b, c, d, e, and f, to represent ten, twelve, thirteen, fourteen, and fifteen as single characters. Each digit position indicates a power of sixteen.

Example: 21316 means:

256s(16^2)	16s(16^1)	1s(16^0)	
2	1	3	
+256×2	+16×3	+1×3	
+512	+16	+3	=531_{10}

Converting from one Number System to Another

1. Any number value in one number system can be expressed in another number system. There are several techniques for converting numbers from one base to another.
2. **Converting to Decimal from another Base**

Following steps are used to convert to a base 10 value from any other number system:

Step 1: Determine the positional value for each digit. (This is dependent on the position of the digit and the base of the number system.)

Step 2: Multiply the column values from step 1 by the digits in the corresponding place.

Step 3: Sum the goods calculated in Step 2. The amount obtained is the equivalent value in decimal.

Binary to Decimal

This conversion can be done by assigning the values to each position and then adding these values together.

Example: $(1101)_2$ is to be converted to its decimal equivalent

According to the steps given above:

Step 1 & 2:

Column Number (from right to left)	Column Value (step 1)	Digit Column Value (step 2)
1	$2^0=1$	$1\times1=1$
2	$2^1=2$	$0\times2=0$
3	$2^2=4$	$1\times4=4$
4	$2^3=8$	$1\times8=8$
	=13	

Step 3: Sum of the products

So, $(1101)_2 = (13)_{10}$

$(1101)_2 = (13)_{10}$ can also be represented as:

2^3	2^2	2^1	2^0
1	1	0	1
8×1	$+4\times1$	$+2\times0$	$1\times1=1$
			$=13_{10}$

i. Hexadecimal to Decimal

Example: $(IAC)_{16}$ is to be converted to its decimal equivalent.

Step 1 & 2:

Column Number	Column Value	Digit Column Value
(from right to left)	(step 1)	(step 2)
1	$16^0=1C\times1$	$12\times1=12$
2	$16^1=16$	$1\times16=10\times16=160$
3	$16^2=256$	$1\times256=256$
		$=428$

Step 3: Sum of the products

So, $(IAC)_{16} = (428)_{10}$

Similarly, $(F5)_{16} = (245)_{10}$ can be represented as

$16^1 \times 15 + 16^0 \times 5$

$=16\times15+1\times5$

$=240+5$

$=(245)_{10}$

a. Octal to Decimal

Example: $(4706)_8$ is to be converted to its decimal equivalent.

Step 1 & 2:

Column Number	Column Value	Digit Column Value
(from right to left)	(step 1)	(step 2)

1	$8^0=1$	6×1=1
2	$8^1=8$	0×8=0
3	$8^2=64$	7×64=448
4	$8^3=512$	4×512=2048

Step 3: Sum of the products = 2502

So, $(4706)_8 = (2502)_{10}$

Similarly, (356)8 = $(238)_{10}$ can be represented as

$8^2 \times 3+8^1 \times 5+8^0 \times 6$

$=64\times 3+8\times 5+1\times 6$

$=192+40+6$

$=(238)_{10}$

1. **Converting from Base 10 to a New Base**

Following steps are used to convert a number from base 10 to a new base:

Step 1: Divide the decimal number to be converted by the value of the new base.

Step 2: Write the leftover from step 1 on the right side of the new base number.

Step 3: Divide the quotient from the previous division by the new base.

Step 4: Assign the remainder from step 3 as the next digit (to the left) of the new base number. Repeat steps 3 and 4, noting remainders from right to left, until the quotient equals zero in step. The last leftover will be the base number's leftmost digit.

Decimal to Binary

Example: $(62)_{10} = (111110)_2$ can be represented as:

Divisor	Quotient	Remainder
2	62	
2	31	0
2	15	1
2	7	1
2	3	1
2	1	1
	0	1

i. Decimal to Octal

Example: $(428)_{10}$ can be converted to its Octal number as follows:

Divisor	Quotient	Remainder
8	952	
8	119	0
8	14	7
8	1	6
8	0	1

Hence $(428)_{10}$= $(1670)_8$

i. Decimal to Hexadecimal

Example: $(428)_{10}$ can be converted to its equivalent hexadecimal number.

Divisor	Quotient	Remainder in hexadecimal
16	428	
16	26	12=C
16	1	10=A
	0	1

Summary

In this session, we spoke about the main aspects of computers. Speed, accuracy, memory, and versatility are some of the characteristics associated with computers. However, the computer we see today did not emerge overnight. It took millennia of human labor to create the computer in its current form. There are five generations of computers. Over the years, the physical size of computers has shrunk, but their processing speed has increased dramatically. We also talked about the different types of computers accessible today. Computers process and store data in binary code, which is composed of 0s and 1s. These binary digits, or bits, are combined into bytes that represent letters, numbers, and other sorts of data. Computers can link to networks like the internet or local area networks (LANs) to interact and share resources with other devices. Networking allows for activities such as web browsing, email communication, file sharing, and online gaming. Computers employ a variety of storage devices to store data, either permanently or temporarily. This covers HDDs, SSDs, CDs, DVDs, and USB flash devices. Software includes programs and apps that direct the computer on what duties to execute. This includes the operating system (e.g., Windows, macOS, Linux), which manages hardware resources and serves as a user interface, as well as programs like as word processors, web browsers, and games. Computer security is critical for protecting against dangers such as viruses, malware, hackers, and data breaches. This includes putting in place security measures like antivirus software, firewalls, encryption, and frequent software updates.

Terminal question

Q. 1. What is a computer? Why is it known as data processor? Write the important characteristics of a computer.

Answer:--

Q. 2. Explain in brief the various generations in computer technology.

Answer:--

Q. 3. Write a short note on Fifth Generations of computer. What makes it different from Fourth Generation computer?

Answer:--

Q. 4. What is the first mathematical device built and when was it built?

Answer:--

Q. 5. Discuss the number system in computer

Answer:--

7.10. Further Readings

1. Rajaraman, V.; Fundamentals of Computers 5th ed., Prentic-Hall of India, New Delhi, (2007).

2. Hennessy, J.L. and Patterson, D. A.; Computer Organization and Design: The Hardware/Software Interfaces, Morgan Kauffman Publishers, San Mateo, CA, (1994

3. Chauhan Sunil, Saxena Akash and Gupta Kratika; Fundamentals of Computer, Laxmi Publication, (2006).

Unit-8: Hardware and Software

Contents

8.1. Introduction

A computer is an electronic device used for processing data that has the ability to read, write, compute, compare, store, and analyze massive amounts of data quickly, accurately, and reliably. Without human intervention, it retains the orders it has been given and swiftly and automatically carries them out. It utilizes the idea of stored programs. It reads the instructions and performs them to create outcomes after feeding it the data and instruction set. Thus, a computer is made up of a device known as "hardware" that operates with the aid of software, which is a collection of instructions.

Hardware are the physical components of a computer system that can be touched and seen. These physical components include the central processing unit (CPU), memory (RAM), storage devices (hard disk drive, solid-state drive), input devices (keyboard, mouse, touch screen), output devices (monitor, printer, speakers), and networking devices (network interface cards, routers, modems). Hardware components collaborate to execute instructions, process data, and complete tasks based on software instructions. On the other hand, software describes the data, instructions, and intangible programs that tell the hardware what to do. It consists of system software, which controls hardware resources and offers an interface for interacting with the computer. Examples of this software include the operating system (Windows, macOS, Linux, etc.). Software also includes programs or apps that carry out particular duties or operations, such word processors, web browsers, games, and multimedia applications. Hardware carries out software instructions to allow users to do a wide range of jobs, from simple computations to intricate simulations and calculations.

Software is made up of the programs and instructions that manage and coordinate the functions of the hardware in a computer system, whereas hardware is made up of the actual components of the system. On computers and other electronic devices, hardware and software cooperate to allow users to carry out a variety of jobs and activities. However, Computers are made up of various components that work together to process data and complete tasks. Computer components include CPU, RAM, storage, input/output devices, motherboard, GPU, and PSU.

Objectives

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

- Main components of computer
- Functions of different components of computers
- Learn about software and its functioning

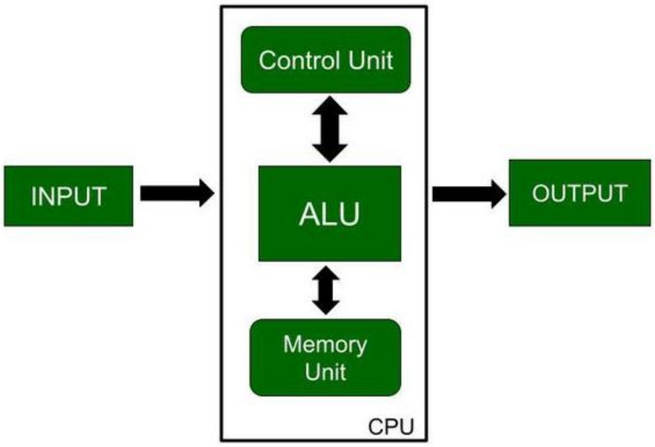
8.2. Central Processing Unit (CPU)

A computer system's most significant component is the central processing unit (CPU). It functions as the computer's brain. Its primary functions are to do calculations and other logical operations. A CPU is the hardware that handles data input/output, processing, and storage in a computer system. A CPU can be put in a CPU socket. These sockets are often found on the motherboard. The CPU can execute a variety of data processing functions. The CPU can store data, instructions, programs, and interim results. This unit will process the data provided by the input unit. Since 1823, when Baron Jons Jakob Berzelius discovered silicon, which is still the key component used in the production of CPUs today, the CPU's history has seen multiple significant turning moments. The first transistor was invented by John Bardeen, Walter Brattain, and William Shockley in December 1947. In 1958, Robert Noyce and Jack Kilby created the first functioning integrated circuit. The CPU contains three components.

(i) The Control Unit

(ii) The Arithmetic and Logic Unit

(iii) The Memory Unit



(i) Control Unit: As the name implies, a control unit monitors the operations of all computer components but does not perform any data processing. It consists of electronic circuits. It oversees the overall operation of the computer system. To execute previously stored instructions, it sends electrical signals to the computer system. It is regarded as the heart of the computer system. It supervises all other units, ensures that they function effectively, and arranges a variety of operations. Design the input device to provide data and instructions to main memory, which is subsequently routed to the Arithmetic and Logical Unit (ALU). The processed results are then sent from the ALU to the memory unit for storage before being shown graphically. The control unit coordinates the many components of the computer system, including the Arithmetic and Logic Unit, memory unit, and peripheral units. The following are some of the control unit's key functions:

- Control Unit in Computers
- Manages all computer units.
- Obtains and interprets input instructions from memory unit.
- Directs computer operation based on these instructions.
- Communicates with input and output devices for data transfer.
- Not responsible for data processing or storage.

The instruction register receives the instructions to be executed in the proper sequence one by one. The instructions' operation code is then transmitted to the Arithmetic and Logic Unit, which performs the operation. The address register allows data in the position provided in the command to be moved to a specific accumulator for the arithmetic and logic unit.

(ii) Arithmetic and Logic Unit

It is composed of electronic circuits. It works at incredible speeds, executing millions of instructions per second (MIPS). This unit performs two types of operations: arithmetic processing and logical processing. Arithmetic processing includes all mathematical operations such as addition, subtraction, multiplication, and division. In logical processing, it executes relation and logical operation operations such as comparing larger or smaller values, true or false statements, and so on.
The CPU's Arithmetic and Logic Unit contains a variety of sub-units and special-purpose circuits such as registers, counters, and adders.

(a) **Register:** Registers are high-speed storage circuitry sections that serve as a "work area" for the temporary storage of instructions and data while the control, arithmetic, and logic units are in operation. The number, function, and capacity of registers and other elements in a CPU are determined by the internal design of each individual computer. In addition to general purpose registers, other registers may be named based on their functions.

(i) **Storage Register:** It temporarily stores data or instructions that are transferred from or sent to primary storage.

(j) **(ii) Address Register:** It may hold the address of the storage location of data, or the address of an input/output device or a control function.

(iii) **Instruction Register:** It contains the instructions being executed by the

(iv) **Accumulator:** It is a register which accumulates the result of arithmetic or logic operations.

(v) **The multiplier-quotient register:** It holds either a multiplier or a quotient.

(vi) **Floating point register:** It is used for floating point arithmetic operations.

(b) **Counter:** The counter and the register are closely related. It is a gadget whose contents can be increased or decreased by a set quantity. The instruction counter, commonly known as the instruction address register, carries the storage location address for the computer instruction being executed. The index register is a counter designated for altering the component of an instruction that provides the address of the data to be changed. This results in a procedure known as "indexing" in which the CPU automatically repeats the execution of the identical instructions until all of the data covered by the instruction is processed.

(c) **Adders:** Adders are sub-units that perform the arithmetic operations of the arithmetic and logic unit receive data from two or more sources, perform the specific arithmetic operation desired and then convey the result to a receiving register such as the accumulator.

(iii) Memory Unit

This unit has the ability to store instructions, data, and intermediate results, as its name implies. When necessary, information must be transferred from the memory unit to other computer units. It's also known as major or main memory. It is composed of ultrafast memories, such as semiconductor or magnetic memory. Words, bytes, and bits are used to store the data and instructions in this unit, which subsequently transfers them to the ALU for processing. Likewise, the results of processing are either delivered to the output unit or kept once more for use in future calculations. Multiple static and dynamic memory cells make up the primary storage. It is possible to save part of the information or instructions somewhere else for future retrieval. Whereas later-needed data and instructions are externally saved, only the data that is being processed and the instructions needed to process it are stored inside. The terms primary storage and main memory are frequently used to refer to the internal storage. Typically, this is restricted. The infinite storage is found in the external storage, often known as secondary or supplemental storage. The memory serves as the microprocessor's "Processor," storing and supplying binary instructions and data as needed. Results are occasionally also kept in memory. Its size has an impact on performance, power, and speed. The computer has two different kinds of memory: primary memory and secondary memory. The following is a list of some of memory units' primary uses:

- Memory units hold instructions and data and are necessary for processing.
- it saves the interim outcomes of any computation or job while it is being completed.
- Before being delivered to an output device so that the user may see the findings, the processing's final results are kept in memory units.
- The memory unit transmits a wide range of inputs and outputs.

8.3. Memory Organization

Primary Storage/ Internal Storage

The main memory, also called random access memory (RAM) is the work area of the computer. It stores program instructions or part of data for immediate needs.

(a) Magnetic Core Memory

In the past magnetic core memory was used as internal memory. It was a non volatile memory i.e., its contents were not lost if the power supply was interrupted. However, the necessity to store after reading was technological disadvantage of core storage.

(b) Semiconductor Memory

These days, internal memory consists of extremely small bit storage circuits (flip-flops) etched on a silicon chip. All the electronic elements to store a bit are placed in such a small area of the chip that a single chip can store millions of bits. The individual chips are arranged in groups to form a memory module.

Types of Semi-conductor Memory

(i) **Random Access Memory (RAM):** Any information can be read from and written into a RAM. It is a read/write memory. It is a volatile memory i.e. its contents are lost if the power supply is interrupted or turned off. The main memory of the computer is RAM

(ii) **Read Only Memory (ROM):** Rom is permanently programmed with information during manufacture, by implementing the appropriate pattern of two state values. It cannot be changed subsequently by a normal write operation. It is thus completely non-volatile. It is mainly used to hold those programs which are required permanently.

(iii) **Programmable Read only Memory (PROM):** This can be programmed to record information using special electronic equipments known as a PROM programmer. However, it cannot be changed subsequently.

(iv) **Erasable Programmable Read Only Memory (EPROM):** EPROM is a PROM which can be reversed by exposing it to an ultraviolet light source. The device can be re-erased and re-programmed again and again.

(v) **Cache Memory:** It is a small capacity high speed memory used to make processing faster. The main memory can process information very fast, but it takes much longer to transfer data to and from the input/output devices. The cache memory compensates for this mismatch in operating speeds. It holds those parts of data and the active program which are most frequently used. Thus the performance rate of the CPU improves. However, cache memory is very expensive as compared to the main memory, so its size is normally much smaller than the size of the main memory.

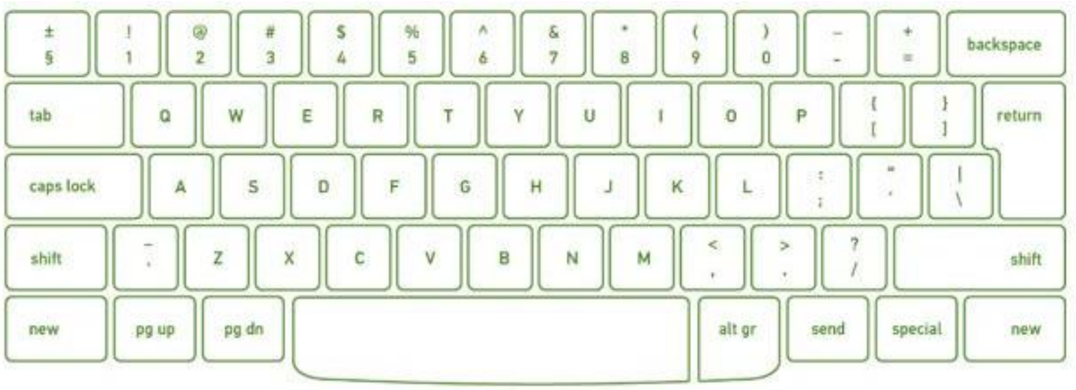
8.4. Input-Output Devices

Any computer system must include input and output devices in order to enable communication between users and the computer. The computer system is made to carry out duties assigned by the user and generate outcomes effectively. It receives instructions (in the form of inputs), processes information (also known as computing tasks), and outputs the results (also known as outputs). With the aid of input, processing, and output units three major computer components the computer's software is built to accomplish this task. To receive instructions and receive the corresponding outcomes of the operations carried out, the computer conducts tasks using a combination of input and output devices.

One of the most used input devices is the keyboard. The information is stored on a substance known as the media. In a similar vein, the computer system requires an input device in order to provide the user with the information it has processed. A few examples of common output device kinds are printers and monitors. However, some devices, such as disk drives, tape drives, and floppy drives, can be used as input or output devices.

Keyboard

A computer keyboard is an electromechanical device that, when a key is pressed, generates unique, standardized electrical codes. The codes are sent through the cable that joins the keyboard to the terminal or computer system unit. The most common and extensively utilized input device for entering data into a computer is the keyboard. The keyboard layout is comparable to that of a standard typewriter, notwithstanding the addition of a few more keys for different functions. Keyboards are typically available in two sizes: 84 keys or 101/102 keys; however, there are also keyboards with 104 or 108 keys that are compatible with Windows and the Internet. At the terminal, the incoming code is decoded and transformed into the relevant computer code. While keyboards vary widely in size and form, they all share the following features such as standard type writer keys, function keys, special purpose keys, cursor movement keys and numeric keys etc.


± § ! 1 @ 2 # 3 $ 4 % 5 ^ 6 & 7 * 8 (9) 0 – - + = backspace
tab Q W E R T Y U I O P { [}] return
caps lock A S D F G H J K L : ; " ' | \
shift ~ ` Z X C V B N M < , > . ? / shift
new pg up pg dn alt gr send special new

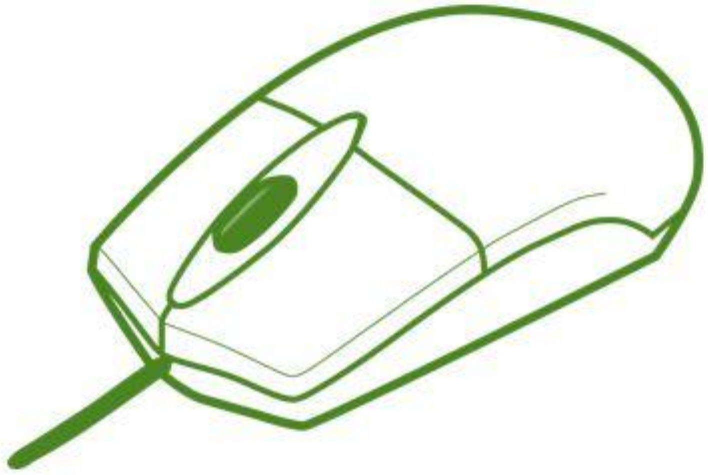
Types of Keys

- **Numeric Keys:** It is used to enter numeric data or move the cursor. It usually consists of a set of 17 keys.
- **Typing Keys:** The letter keys (A-Z) and number keys (09) are among these keys.
- **Control Keys:** These keys control the pointer and the screen. There are four directional arrow keys on it. Home, End, Insert, Alternate(Alt), Delete, Control(Ctrl), etc., and Escape are all control keys (Esc).
- **Special Keys:** Enter, Shift, Caps Lock, NumLk, Tab, etc., and Print Screen are among the special function keys on the keyboard.
- **Function Keys:** The 12 keys from F1 to F12 are on the topmost row of the keyboard.

Mouse

The mouse is the most widely used pointing device. A small cursor is moved across the screen by clicking and dragging with the mouse. You cannot release the mouse to move the pointer. The mouse must be moved by you in order for the computer to move; it cannot move itself. It is therefore an input device. These days, the mouse is a commonly used input device for working with GUIs (Graphic User Interfaces) and visuals. It's about the size of an audio cassette, with two or more buttons on top, and it glides on a rubber ball. The screen cursor moves in the same direction as the mouse when it is slid across a flat surface. The specified position can be sent to the system by simply clicking on the button.

With a mouse, you may control the coordinates and movement of the on-screen pointer or cursor by moving the mouse on a flat surface. While the right mouse button, when clicked, brings up additional options, the left mouse button can be used to move or select objects.



MICR

Devices for character recognition using magnetic ink were created to support the banking sector. Cheque processing is one usage for it. The E13B font, which has four special characters in addition to the numbers 0 through 9, is the character set that MICR devices utilize the most frequently. These assist in classifying checks and drafts according to the code printed in the E13B font on the check. MICR speeds up processing, but its primary drawback is that it can only handle 10 numbers and 4 characters.

Scanners

In essence, these are input devices with character or mark recognition capabilities. They are employed in the direct input of data into computers. An input device that works similarly to a photocopier is called a scanner. It is used when data that needs to be moved from paper to the computer's hard drive for further processing needs to be done. Images are gathered from the source by the scanner, which then transforms them into a digital format that can be stored on a disc. These graphics can be altered before they are printed. The need for duplicate human labor to enter data into a computer is eliminated by scanners. Data accuracy is increased when human interaction is reduced. Scanners require high-quality documents because they are direct data entry devices.

TYPES:

1. **OCR**

   These scanners work by comparing the forms to patterns that are stored inside to identify alphabetic and numeric characters. These are pricey and only employed in high volume processing applications, such as those provided by credit card companies. CR optically scans the text, character by character turns it into a machine-readable code, and saves it to the system memory.

2. **OMR**
3. One tool that's commonly used in educational institutions to verify answers to objective exams is the optical mark reader. Pen and pencil marks are recognized by it. These scanners are able to identify a pencil mark of a predetermined type. These are typically used to validate input documents and assess response sheets for objective assessments.
4. **Bar Code Reader**

   Bar codes are patterns of data that alternate between light and dark lines or bars. The retail industry is one that uses bar codes primarily for product labeling. A tool for reading data with bar codes is a bar code reader. A laser-bean scanner that is connected to a computer reads bar codes.
5. **Desk Scanning**

The scanner exposes your image to light throughout the scanning process. The image's light is reflected back into the scanner's optics, which interprets the image's various light levels. After then, a computer reconstruction of your image appears on your screen. With scanning software, you can modify the data from an image captured by a camera and save it on your computer for use in any kind of application.

Magnetic Tape

Throughout the world, tape is a highly common sequential access storage medium. On one side of a plastic tape, magnetizable substance coatings tiny areas on which data is stored. The tape's coated side is separated into horizontal rows called tracks and vertical columns called frames.A 9-track tape is used with an 8-bit coding. The parity bit is recorded on the ninth track. When a bit from a string of eight bits is lost during data input or output operations, an error can be detected using a parity bit, also known as a check bit or check bit. To ensure that there is always an even number of one bit, an extra one bit is added to the check bit location if the basic code for a character demands an odd number of one bit. An illustration of even parity is this. Likewise, with odd parity, an odd number of one bits is always produced by using the check bit. In other words, the check bit will be 1 in the event that the total number of one bits used to represent a certain character is even, and it will be 0 in the other case.

Because magnetic tape is a continuous-length medium, blank spaces known as Inter-Record Gaps (IRGs) may exist between tracks that are stored on the tape. IRGs typically have a length of 0.5 inches. Records on tape can have different durations. A lot of tape is wasted if there are a lot of short records on the tape and if an IRG separates each record. A tape block can be created by combining many records to prevent this inefficient scenario. The amount of records in each block is fixed, and the blocking factor controls the number of records in data blocks. Two blocks are separated by an interblock gap, or IBG. The tape drive can recognize the end of the block, accelerate to the necessary speed for reading, and decelerate after reading thanks to the interblock gap. A tape is referred to as a multifile reel if it holds several files, while a file is referred to as a multireel file if it is stored on multiple tape reels.

The amount of frames per inch of tape is known as the tape density. Bits per inch, or BPI, is used to measure it. Densities of 1600 BPI and 6250 BPI are typical. On a tape, each file has a header label at the beginning and a trailer label at the end. Header label: A header label is a block of text that contains information that uniquely identifies a file, such as its name, creation date, and retention duration.

A trailer label is a block of text that appears as the final record of a file. It has the same details as the header label, but it also includes a block count that indicates how many data blocks are in the file.

Cartridges Tape

It's a plastic ribbon with a magnetizable iron-oxide coating on one side. Its length ranges from 140 to 450 feet, and its width is 1/4". It is enclosed in a dust-resistant jacket with a pocket size of 3 W'/5 1;4". utilized in personal computers and minicomputers. used to take hard drive backups. The process of copying a hard disk's contents takes five to twenty minutes. The capacities of tape cartridges range from 60 MB to 32 GB.

Floppy Diskette

Even though it has a far smaller capacity than a hard disk, the floppy diskette is a direct access storage device. The plastic used to make the diskette is flexible. A recording substance consisting of iron oxide is applied to this base. Bits of magnetic spots are used to record data. Like a hard disk, the surface is separated into sectors and tracks. A diskette's recording density determines how many tracks it can hold. The size of each sector is fixed (512 bytes). Data is saved on both sides of the diskette. For reading and writing data on the diskette, the floppy drive has one head per surface. To keep it free from dust and scratches, the round plastic disk is housed in a square jacket with a smooth interior. These days, the common sizes are 3.5 and 5.25 inches. The following are their capacities:

1. 5.25 inch

(a) DSDD	48TPI	360KB
(b) DSHD	96TPI	1.2MB

2. 3.5 inch

(a) 120	TPI	1,44MB
(b) 240	TPI	2.88MB

Salient Points

1. The index hole marks the beginning of the first sector.
2. The outermost track is labeled 0
3. The write protect notch can be covered to disable writing.

Advantages

1. Low cost.
2. Convenient to transport.
3. Compatible between computers.

Disadvantages

1. Floppy diskettes are prone to frequent errors due to mishandling.

Magnetic Disk

The most widely used INPUT/OUTPUT device for Direct Access Storage is a magnetic disk. These are metal plates having a tiny layer of magnetic material applied to both sides. A disk pack is created by fixing a series of these magnetic material plates, one below the other, to a spindle. The disk drive, which has a motor to rotate the disk pack about its axis, holds the disk pack in place. A series of magnetic heads positioned on arms that radically move in and out are part of the disk drive. There is extremely little space between the head and the plate surface because the head can wear out and destroy the data if it comes into touch with it. It would cause the head to wear out. A predetermined number of bytes can be stored in each sector, which is further divided into tracks, which are several invisible concentric rings that store data on both surfaces.Today's common diskettes have capacities ranging from 360KB to 1.44 MB. The hard drive size of a microcomputer might be anywhere from 10MB to 1GB.
The following are the features of hard disks.

They are rigid metal platters connected to a central spindle.

1. The disks and read/write heads of the complete disk unit are housed in a container that is hermetically sealed.
2. To avoid contamination, air passing through the container is filtered.
3. The disks rotate at an extremely fast rate (about 3,600 RPM; floppy disks rotate at roughly 300 RPM). These disk drives come in sealed units with four or more disk platters. In the majority of these disk units, also known as Winchester disk drives. The disk surface is never in contact with the read/write heads. Rather, their intended function is to levitate between 0.5 and 1.25 millionths of an inch above the disk surface.

Removable Disk Packs

In big computer systems, hard disks are sometimes included in removable packs. This means that they can be removed from the computer and replaced whenever necessary. Disk packs typically accommodate 6 to 12 platters measuring 14 inches in diameter. In disk packs, all tracks with the same number are arranged one above the other. All tracks include heads, allowing both sides of the disk to be read. The read/write heads travel together and are always on the same cylinder at the same time. Data that requires more space than one track is transferred to the same track on another disk, eliminating the need for the read/write heads to travel. (Only one read/write head is active at a time, although they operate quickly). When all of the tracks in a cylinder are full, the read/write heads shift to the next cylinder; the computer operating system uses cylinder numbers to calculate data addresses. The capacity of removable disk packs varies by manufacturer, ranging from 150 to 250 megabytes. The total storage capacity might be significantly expanded by having a dozen or so extra packs that can be swapped out with the packs in the disk drive. These are now obsolete.

Fixed Disks

The capacity of fixed units has been raised to 16GB. The common capacity of fixed disk drives for desktops is 4 GB, however a typical server will have many disk drives of 8 GB apiece. The average access time for such drives is in the order of a few milliseconds, with data transmission rates of a few million bits per second.

Access Time

This is defined as the time taken to locate and then transfer data from the disk into internal storage. It includes there elements namely.

1. **Seek Time:** This is the amount of time needed to move a read write head to the desired track from its current position.
2. **Latency Time:** Also known as Rotational Delay time. This is the time require for the portion of the track to be read or written to come beneath the read-write head.
3. **Transfer Time:** The time taken to transfer data from disk to internal storage.

Optical Disk

Optical technology can circumvent some of the constraints encountered by consumers when employing magnetic storage technology. In optical disks, a laser beam is utilized to

read and write data. The information is stored in the form of minuscule dots. There are approximately 54000 spiral tracks on a disk. The track density is around 16,000 TPI. Optical disks can hold video, text, music, and graphics.
A layer of plastic protects an optical disk from physical handling. Data is written on the disk by blasting minuscule pits on its surface using a laser beam. The pits are darker than the shiny disk's background. The data is read by sending a less intense beam across the surface, and the variations in reflectivity indicate a 1 or 0 bit.

Advantages

1. Not as fragile as floppy disks.
2. Longer lasting.
3. More storage.
4. No heads required for reading/writing.

The main limitation of optical storage is that data once written can’t be erased.

NOTE

New technologies are being developed for erasable and rewritable optical media. CD-ROM optical disks, composed of aluminum and plastic, offer high storage density and quick access. A CDROM is only 1.2 mm thick, but it is robust. Not only is a CD-ROM more resistant to harm, but there is no danger of accidently overwriting any data or infecting it with a virus. A normal 12 cm diameter CD-ROM can store up to 680 MB of data. Because of its enormous storage capacity, CDROM is currently an economical medium for distributing moderately advanced multimedia content. CD-ROM addressing is done by measuring the time and number of data blocks read. Minutes, seconds, and blocks provide information for locating a piece of information. A track commencing halfway through the CD-ROM, for example, is addressed at 29:29:37 (minutes: seconds: blocks).

A CD-ROM player reads information optically using laser light.

WORM

Write-once, read-many systems use writable optical storage devices. A laser recording device records one bit by deforming a thin sensitive layer of material on the disk surface. Unmodified sections contain 0 bits. The distorted WORM disk cannot be recovered to its previous state, therefore writing is permanent. Reading recorded data occurs when a low-power laser beam passes over the disk and detects changes in reflections from 0 and I-bits.

Digital Versatile Disc

Digital versatile disks are optical disks with the same overall dimensions as CDs but significantly larger capacity. These can hold at least seven times as much data as D ROM. These drives support the MPEG-2 standard for data compression. High compression of DV D films requires either a fast Pentium II processor or an MPEG Card for decoding. Dual layer DVD disks have an 8.5GB capacity on one side and a total capacity of 17GB when both sides are used. Large storage capacities on DVD expand its multimedia capabilities.
These discs are backwards compatible and can read CD-ROM, CD-RW, and CD-audio storage media. One key feature is that a DVD drive does not require a particular interface.

Visual Display Terminal

The most often utilized I-a device for interactive processing these days is this one. Data is entered into a processor via a keyboard, and processed information and messages are received from the computer and shown on a monitor, a type of video display unit. VDTs belong to a category.

Dumb terminals: These are simple devices that immediately transmit each keyed data character to the processor.

Intelligent Tenninals: These combine VDT hardware with built in microprocessor, They can process small jobs without the need to interact with the main computer.

Cathode Ray Tube

The most widely used softcopy output device is likely the cathode-ray tube, which is also used as a monitor for microcomputer systems and terminals connected to larger computer systems. The operator can observe data entering and computer output on this kind of video display panel. The term "alphanumeric monitors" refers to terminals that show only letters and special characters like $, *, and?. With 24 lines viewable at once, they resemble television displays and show 80 characters per line. Graphic monitors, often known as graphic terminals, are displays that include graphics in addition to alphanumeric data.

The tiny visual components that make up the CRT's screen display are referred to as pixels. Resolution, or the amount of points that can be illuminated on the screen, is better for smaller pixels than for visual clarity. The sharpness of the image that is presented on screen is referred to as resolution. Resolution is determined by three factors: band width, raster scan rate, and lines of resolution (horizontal and vertical). The number of times per second that the image on the screen can be refreshed and illuminated again is referred to as the raster scan rate. The electron beam must constantly sweep across the screen to avoid making the phosphors it hits shine for very long, which can be quite taxing on the eyes.

The visual quality and eyestrain are positively correlated with the greater the raster scan rate.The pace at which data may be delivered to the electron gun to regulate its movement, position, and firing is referred to as bandwidth. The electron cannon may be directed to work more quickly the higher the bandwidth.

Monochrome and Color Monitors: First, the number of electron guns on a monochrome monitors (a monitor that can only display a single color image) and a ROB color monitor (ROB stands for red, green, and blue) differs. There is only one electron gun on a monochrome monitor.

Printers

The main output device used to create lasting papers for human use is a printer. Printers fall into one of two categories:

1. **Impact Printers**

   These operate like typewriter, pressing a typeface against paper and linked ribbon E.G. daisy wheel printer, dot-matrix printer.

(i) **Letter Quality Printer:** Letter quality printers also called character printers or serial printers because they print one character at a time, produce a very high quality print image (one that is very clear and precise) because the entire character is formed with a single impact.

(a) **DMP (Dot Matrix Printer):** Because they are serial printers, each character is printed one at a time. Every character appears as a dot pattern on paper. The print head is made up of a matrix of small needles that hammers out characters in the form of patterns of tiny dots. The matrix is usually 9 rows by 7 columns. Despite having somewhat lower letter quality than Daisy Wheel printers, these printers are speedier. Another benefit of DMPs is that, unlike line printers, they do not have a fixed character font set, which allows them to print a variety of shapes, including graphs, charts, and diagrams.

(b) **Daisy Wheel Printer:** With a set of print characters representing the outside tips of the flat spokes, the Daisy wheel printers include a print "wheel." The wheel is rotated until the relevant spoke, or petal, is in line with the print hammer in order to print a particular character. After that, the print hammer is fired, pressing the print character firmly enough to leave a clean, distinct impression on the paper and ribbon.

(ii) Line Printers

These are high-speed printers that can handle big volumes of output for large computer businesses. Chain printers, band printers, and drum printers are examples of line printers, which use impact mechanisms to generate one line of printed output at a time.

Depending on the printer, it is possible to print 300 to 3000 lines per minute. These printers use multiple copies of each printable character on a drum, bell, or print chain, with a separate print hammer for each position across the width of the paper guide. As the drum, bell, or print chain turns, the hammers are actuated when the relevant character passes in front of them. This type of printer can print 200 to 3,000 lines per minute (LPM). This printer's main benefit is its speed. The primary disadvantages are noise and poor image quality.

Non-Impact Printers

These are thermal, electrostatic chemical and inkjet technologies.

(i) **Thermal Printers:** They use heat to create a picture on specialized paper. The print mechanism, which functions similarly to a dot-matrix print head, is intended to heat the surface of chemically treated paper in order to form a dot based on the chemical's reaction with heat. There's no ribbon or ink involved. Thermal printers are the best option for users looking for high-quality desktop color printing. They are, however, pricey, and special paper is required.

(ii) **Ink Jet Printer:** The ink jet printer sends a constant stream of ink droplets towards the printed paper. The drops are selectively rejected by electrostatic attraction, leaving only those required to make the desired sign. Those that aren't needed are collected in a small gutter and filtered to remove contaminants. They are then recalculated using the drop-generating method.

(iii) **Laser Printer Technology:** This is far less mechanical than impact printing (no print heads move, no print hammers strike), resulting in significantly faster and quieter operation. The procedure is similar to how a photocopier works. A laser beam is directed across the surface of a light-sensitive drum and adjusted as needed to create a picture in the shape of a pattern of small dots. The image is then transferred to paper, one page at a time, in the same manner as a copy machine, using a specific toner.

The main advantages of laser printer are:

1. Very high speed.
2. Low noise level.
3. Low maintenance requirements.
4. Very high image quality.
5. Excellent graphics capabilities.
6. A variety of type sizes and styles.

7. On large high speed laser printers, form can be printed at the same time data is recorded in them.

PLOTTERS

These are line-drawing devices which move a pen under computer control in such a way that continuous lines and curves can be drawn. These are used for drawing maps, engineering drawing etc.

COM (Computer Output Microfilm)

COM Technology is utilized to capture computer output data as microscopic film images. Thus, COM is essentially an output device that stores data on a roll of microfilm. The COM recording technique comprises of a microfilm recorder that receives information. The recorder, in turn, projects the output characters onto a CRT screen. A high-speed camera built into the system captures images of the displayed information. The COM recording technology generates characters that are approximately 50 times smaller than those produced by traditional printers. The recorded information is viewed using a specific instrument called as a MICROFILM READER. A COM system is appropriate for applications requiring a huge volume of data to be retained.

8.5. System Software

System software is a set of programs that control and ease the operation of a computer system and its hardware components. Unlike application software, which is created for specific tasks or user demands, system software includes key functionalities that allow programs to execute and interact with hardware properly. It is a collection of general programs used to govern the operation of a computer system. System software is required to execute application bundles. Examples include control and processing programs, among others.

1. **Control Programs**

They handle all system activities including as input, output, scheduling, and interruptions. These include programs such as operating systems, job control programs, and I/O management applications. **Operating System (O.S.)**

The operating system is the most important system software component, managing hardware resources and providing key services to other software applications. It enables communication between hardware components, manages system memory, schedules tasks, handles input and output processes, and serves as a user interface for interacting with the computer. Popular operating systems include Windows, macOS, Linux, and other Unix variants. On the basis of functioning and facilities provided by them, Operating system can be classified as follows:

(i) Single User Operating System

These Operating Systems allow only one user to work on a computer at a time. **Example:** MS-DOS, CP/M.

(ii) Multi User Operating System

This operating system lets multiple users to work on the computer at the same time. These operating systems allocate memory so that multiple users can operate concurrently without interfering with one another. It also distributes processing time so that each user receives a rapid answer from the machine. These are also known as time-sharing operating systems. **Example:** UNIX, XENIX, VMS, Windows NT.

Various function of Operating System:

- Memory Management
- Processor Management
- Device Management
- File Management

(i) Memory Management Functions

The operating system manages the system's primary memory. It allocates memory based on the request of the process that is currently running. It also keeps track of how many bytes of RAM are being used at any given time and by which process. It also keeps track of which parts of it are free. In the case of a multi-user system, it determines which user has access to memory and when. The quantity used is determined by the requirements.

(ii) Processor Management Functions

The operating system also takes care of the processor. It allocates the processor to the user. In a multi-user system, processor time is allocated to different users as needed, ensuring that each user has the shortest possible wait time.

(iii) Device Management Functions

It keeps track of all the peripherals connected to the computer, such as I/O devices. When necessary, it arranges the devices in such a way that each can be used efficiently. It initiates I/O operations and assigns them to the user with other devices.

(iv) File Management Functions

An operating system's function is to write and retrieve information to and from a secondary storage medium. It adheres to a comprehensive approach for file maintenance, ensuring that multiple sets of information are not jumbled up and that the user receives the exact set of information required.

2. Processing: Programs

- Processing programs are programs that manage the computer's application programs. These programs operate under the supervision of control programs. They assist the application program to perform the real data processing: These are primarily of two types:
- **Language translators** convert user-created high-level programs to machine language. Examples include compilers and assemblers.
- **-Service programs,** often known as utility programs, are a collection of programs that perform common data processing tasks. Those are routine actions that all computer users expect their machines to accomplish on occasion (saving, copying, etc.). Examples include sort/merge programs that arrange unsequenced data into a defined sequence, debugging tools that assist the user in locating and correcting logical problems in the program, and so on.

- **Some Processing Programmes**

Linker

A program that combines the separately translated modules to create an absolute load module that can be run as a whole.

Loader

A program for loading the absolute load module into main memory

Interpreter

A program that translates and executes each instruction in a high-level language before moving on to the next instruction.

8.6. File Commands

i) HELP

PURPOSE: This command displays information about the DOS commands. One can seek information in two ways.

FORMATE: a) HELP [command]

A screenful of descriptive information is displayed about the named command.

b) Command

ii) DATE

PURPOSE: To set the system date. The change date shall now be used for the date stamping of files. Format of date can be changed.

FORMATE: DATE [dd-mm-yy]

iii) TIME

PURPOSE: To set the system time. The changed time shall now be used for the time stamping of files.

FORMATE: TIME [HH: MM: SS: cc]

iv) DIR

PURPOSE: To list all or specified files of the connected area on the specified device.

FORMATE: DIR [drive:] [pathname] [/P] [lW] [lA:x] [/B] [/L] [/O:z] [IS]

/P To see the page-wise listing of directory. In this case if the output of the command is more than one page, it shall pause after each screen full. On pressing any key, generally the space bar, next screen is shown.

/W To see the width-wise listing of directory. It displays only file name an extension. Each line contains five file names. The directory names are enclosed in square Brackets.

/A:a Displays files having certain file attributes, where attribute(a) is one of the following:

h	hidden files	-d	files only (no directory names)
-h	non-hidden files	a	files that have been archived
s	system files	-a	files that are not archived
-s	non-system files	r	read only files
d	directory names only	-r	files that are not read only

v) SET DIRCMD

PURPOSE: The DIR command parameters can be preset using this command. It can be entered directly from the DOS prompt.

FORMATE: you can use any valid combination of DIR parameters and switches with the SET DIRCMD command, including the location and name of a file.

i) ATTRIB

PURPOSE: To set or show file attributes.

FORMATE: ATTRIB [+R] [-R] [+A] [+S] [-S] [drive:] [path] <filename> [IS]

+	Sets an attribute
-	Clear an attribute
R	Read only file attribute
A	Archive file attribute
S	System file attribute
H	Hidden file attribute
/S	Process files in all subdirectories in the specified path.

ii) DOSKEY

PURPOSE: This command saves all DOS commands typed from the DOS prompt to a memory buffer. Use the up and down arrow keys to remember commands. By default, only one command is stored in the buffer. It remains in the buffer as long as no additional commands are run. The DOSKEY command allows us to save multiple commands in the buffer, which can then be recalled.

FORMATE: DOSKEY [/REINSTALL] [BUFSIZE=n] [/HISTORY] [/INSERT] [LOVER TRIKE]

/INSERT Puts DOSKEY into insert mode. This allows you to insert text into a display command. To briefly engage overstrike mode, hit the Ins key.

/OVERSTRIKE Puts DOSKEY in the insert mode. This lets you insert text within a display command. To temporarily activate the insert mode, Ins key can be pressed.

/RINSTALL: Clears the buffer.

/HISTORY Displays all commands presently in the DOSKEY buffer.

iii) CLS

PURPOSE: Clears the display screen and DOS prompt appears on the top left comer of the screen.

FORMATE: CLS

iv) TYPE

PURPOSE: Display the contents of specified file.

FORMATE: TYPE [drive:] [path] <filename>

v) COPY

PURPOSE: Copies one or more files to specified or files on specified disk.

FORMATE: Copy <source-file-spec> <target-file>[Iv]

/V Causes DOS to verify that the sectors written on the target diskette are recorded properly.

vi) MOVE

PURPOSE: Move one or more files to the location you specify. The MOVE command can also used to rename directories.

FORMATE: MOVE [/Y I I-Y] [drive:] [path] [filename] [drive] [path] filename […] destination

/Y Indicates that you want MOVE to replace exiting files(s) without prompting you for confirmation. By default, if you specify an existing file. MOVE will ask you if you want to overwrite the existing file.

/-Y Indicates that you want MOVE to prompt you for confirmation when replacing an existing file

vii) REPLACE

PURPOSE: Used to selectively replace files on the target disk with files having the same name on the source disk.

FORMATE: REPLACE [DRIVE 1:] [PATH 1] FILENAME [DRIVE 2:] [PATH 2] [/A] [IP] [/R] [/W]

REPLACE [DRIVE 1:] [PATH 1] FILENAME [DRIVE 2:] [PATH 2] [IP] [/R] [IS] [IU] [1W]

/A Copies specified files that are not present on the target disk. Cannot use with /S or /U switches

/P Prompts you as each file is encountered on the target drive.

/R Also replaces read-only files on the target drive.

/U Searches all directories on the target drive for filenames that match those on the source drive.

/W Waits for you to insert a diskette before beginning.

viii) RENAME

PURPOSE: Changing the name of a file.

FORMATE: RENAME [drive:] [path] <old name> <new name>

ix) PRINT

PURPOSE: Prints a queue(list) of data on the printer.

FORMATE: PRINT [/D:device] [IQ:n] [IT] [drive: [path] [filename[….]][/C][IP]

/D:device Specifies a print device. The valid values for parallel ports are LPT1, LPT2 and LPT3. Valid values for serial parts are COM1, COM2, COM3 and COM4. The values LPT1 and PRN refer to the same parallel port which is also the default.

/Q:n By default 10 files are allowed in the print queue, otherwise the range is 4-32 /Q:n switch is used to specify the maximum number of files you want in the print queue. This should be used before giving any print command.

/T Terminates print queue i.e. removes all files from print queue.

/C Cancels printing of the preceding filename and subsequent filenames.

/P Adds the preceding filename and subsequent filenames to the print queue.

x) PRINT

PURPOSE: To delete specified files from specified diskette.

FORMATE: [drive:] [path] filename [IP]

/P Prompts you before the deletion actually occurs.

xi) MEN

PURPOSE: Displays the amount of used and free memory.

FORMATE: MEN

8.7. Editing Commands

1. How to Start Edit

To invoke full screen editor the following command is used.

EDIT

PURPOSE: Edit is the full screen editor program available with DOS which allows us to create, change and display program and text files. It can be used to:

- Create new files and save them on disk.
- Update existing files and save both the updated and original files.
- Delete, edit, insert and display lines in files.
- Search, delete or replace txt within one or more lines in a file.

FORMATE: EDIT [drive:] [path] filename]

The EDIT program, which is a convenient full screen editor was introduced as a standard feature with the release of DOS 5.0. EDIT does not operate without the presence of vertically and horizontally.

2. USING PULL DOWN MENUS

Once in the full-screen editor, hit the FI key to display help for the current action. The Edit program features four pull down menus, which can be accessed by hitting the

AL T key. Once the AL T key is struck, use the left or right arrow keys to access file, Edit, Search, and Options. To pull down the menu, either highlight the desired option using the left and right arrow keys and then press the return key, or simply write the initial letter of the menu name, such as F, E, S, or O. Escape to exit the menu operations.

FILE: This menu perform all operations required to open and save files or to exit the EDIT program. The various options are:

NEW Clear out the existing papers before creating a new one. If any modifications have been made after the last save, it will prompt you to save before clearing it.

OPEN Open a document; it will request you for a file name and offer a list of files in the current directory with the TXT extension. To open a file, input the file name and press return, or press tab to navigate to it using an arrow key and then press return. Wildcards can be used to list certain filenames. One can even use the tab key to navigate to a different directory or drive.

SAVE Save the current file using the existing file name.

SAVE AS Save the current file with a new file name.

PRINT Print either the complete document or a selected part of text. The part is selected by highlighting it using shift and arrow keys.

EXIT Exit the editing session. If the open file has been modified then a chance is given to save it before quitting.

i) **EDIT:** This menu performs all operations like cut, copy, paste or delete, selected text. Text is selected using the shift and arrow keys.

CUT Removes selected text from the screen and puts it on clipboard. The shift-Del key combination.

COPY Places the selected text on the clipboard without cutting it from the screen. The short-cut is to press Ctrl-Ins key combination.

PASTE Insert text from the clipboard to the present cursor location. The shortcut is to press Shift-Ins key combination.

CLEAR Delete select text without putting in on the clipboard. The shortcut is to press Del key when the cursor is in the selection.

ii) SEARCH: This menu perform all operations like locating a specified text string within the current document, replacing it with another text string.

FIND Finds a specified string.

REPEAT Finds the next match in the current document.

LASTFIND The shortcut is to press Ctrl-L or F3 key.

CHANGE Finds a specified string and replaces it with another. One can verify the change before replacement and advance to the next occurrence of the search string.

iii) OPTIONS: It can be used to change. Display attributes. This menu performs all functions like setting up display colors, turning the scroll bar on or off, changing the tab stop setting the file path for the EDIT help file.

DISPLAY Picks foreground and background color from a list, turn the scroll bars on or off, and set tab stops.

HELP PATH Picks the file path in which the EDIT. HELP file is located.

iv) HELP: It can be used to get help on usage of EDIT.

8.8. Summary

The physical components of a computer system are referred to as hardware, and include the CPU, memory, storage devices, and input/output devices such as keyboards and monitors. Software refers to the programs, applications, and data that direct the hardware on what tasks to accomplish. This includes both the operating system, which manages hardware resources and serves as a platform for application execution, and application software specialized for specific tasks like as word processing, gaming, or graphic design. Hardware and software work together to make a computer function, with hardware executing software instructions to execute various computing tasks. The different components of computers classified as central processor unit input and output devices. The central processor unit is the most significant portion of a computer since it performs mathematical and logical calculations as well as processing data received from the input unit. The CPU is further divided into three components: (i) the control unit, (ii) the arithmetic logic unit, and (iii) the memory unit. The control unit is made up of an instruction register, decoder, address, register, and instruction counter. The arithmetic and logic unit conducts numerous operations based on the computer's instructions. This lesson also includes a full overview of system software, as well as various instructions that must be entered to perform a specific function. System software serves as the backbone of a computer system, allowing it to start up, manage hardware resources, operate applications, and offer critical services to users and other software programs. It is critical to assuring the overall stability, security, and functionality of the computing environment.

8.9. Terminal questions

Q.1. What is a computer?

Answer: --

Q.2. State the essential components of a compute and give function of each of them in brief.

Answer: --

Q.3. Outside CPU, which unit supplements the main memory of a computer?

Answer: --

Q.4. What is different media and related storage devices?

Answer: --

Q.5. What is the purpose of having several mass storage devices with a computer?

Answer: --

Q.6. What are different types of software?

Answer: --

Q.7. What are the functions of an operating system?

Answer: --

8.10. Further Readings

1. Rajaraman, V.; Fundamentals of Computers 5th ed., Prentic-Hall of India, New Delhi, (2007).
2. Rajaraman, V. and Radhakrishnan T.; Digital Logic and Computer Organisatin, 1st ed., Prentic-Hall of India, New Delhi, (2006).

3. Hennessy, J.L. and Patterson, D. A.; Computer Oranization and Design: The Hardware/Software Interfaces, Morgan Kauffman Publishers, San Mateo, CA, (1994).

4. Chauhan Sunil, Saxena Akash and Gupta Kratika; Fundamentals of Computer, Laxmi Publication, (2006).

Unit-9: MS Office

9.1. Introduction

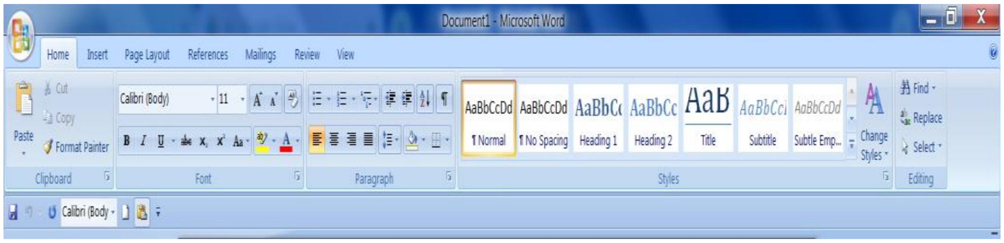
Microsoft Office is a complete suite of productivity software created by Microsoft that addresses a wide range of personal, academic, and professional purposes. Since its debut, Microsoft Office has developed tremendously, adding new capabilities, updating existing ones, and adapting to shifting technological landscapes. The suite includes numerous basic programs, each of which serves a specialized role, as well as supplementary tools and services to improve productivity and collaboration. Microsoft Word, a word processing product that transformed document creation and editing, is at the heart of the suite. Word has numerous options for formatting text, inserting graphics, tables, and objects, and managing document structure. Word's spell-checking, grammar-checking, and auto-correction features enable users to easily create polished texts. Its collaborative features, such as track changes and comments, allow for seamless collaborating on shared documents. Word documents are adaptable and widely compatible, making them useful for a variety of tasks, from writing letters and reports to creating resumes and brochures. Excel is a robust spreadsheet tool for data analysis, calculation, and visualization. It features a grid interface for arranging data into rows and columns, as well as built-in functions and formulas for mathematical operations and statistical analysis. Excel's charting tools allow users to build visually appealing graphs and charts that effectively communicate data trends and patterns. Excel is useful in a wide range of industries where data manipulation and analysis are critical, including budgeting, financial analysis, inventory management, and scientific research. PowerPoint is a presentation program that lets users construct dynamic slideshows to visually communicate information. It provides a variety of slide layouts, themes, and transitions to help improve the visual attractiveness of presentations. Users can use multimedia features such as photographs, movies, and audio to successfully engage their audience. PowerPoint's animation and slide timing capabilities allow presenters to communicate information in an engaging and disciplined manner, whether in boardrooms, classrooms, or conference halls. Outlook functions as an email client and personal information manager, including capabilities for email communication, calendar management, task organization, and contact monitoring. Its user-friendly layout and robust features make it popular among both consumers and corporations. Outlook's interface with Exchange Server allows for seamless synchronization of emails, calendars, and contacts across numerous devices, helping users remain productive and organized wherever they go. Furthermore, Outlook's advanced capabilities, such as rules and filters, assist users in managing their inbox more efficiently, minimizing clutter and increasing email productivity. Microsoft Office is a complete suite of productivity software designed to meet a wide range of personal, academic, and professional purposes. Microsoft Office, with its core applications such as Word, Excel, PowerPoint, Outlook, and OneNote, as well as additional tools and services, enables users to effectively create, communicate, collaborate, and manage information, increasing productivity and driving success in today's digital world.

Objectives:

After this reading, the learner will able to know

- The Microsoft office and their features
- the word processing software microsoft office word
- Spread sheet software like microsoft office excel
- Presentation software like power point
- Excel as data base software

9.2. Word Processing Software



Microsoft Office Word, sometimes known as Word, is powerful word processing software developed by Microsoft. It is part of the Microsoft Office suite, which includes other productivity applications like Excel, PowerPoint, and Outlook. Word is routinely used to generate, edit, and format documents for a wide range of applications, from simple letters and memos to detailed reports and academic papers. This is a 600-word summary of Microsoft Office Word. Former Xerox programmers Charles Simonyi and Richard Brodie developed Microsoft Word, which was released in 1983 for MS-DOS operating systems. It has evolved significantly over the years, with numerous updates and revisions, from a fully text-based interface to a graphical user interface (GUI) with advanced formatting and editing capabilities. The most recent version as of my last update is Microsoft Office Word 2021. The main features of Microsoft Office are:

- **Word Processing:** Word facilitates document creation, editing, and formatting. Users can add text, images, tables, charts, and other items to their documents.

- **Formatting Tools:** Offers various formatting options, such as font styles, sizes, colors, alignments, spacing, and indentation. Users can also employ styles, themes, and templates to enhance the appearance of their documents..

- **Spell Check and Grammar Check:** Word contains built-in capabilities for correcting spelling and grammar problems in documents. It can also provide recommendations for better writing.

- **Collaboration:** Track changes, comments, and real-time collaboration enable several users to work on a document simultaneously, ideal for team projects and reviews.

- **Templates:** Word offers pre-designed templates for various document formats, such as resumes, newsletters, brochures, and invoices, to help users get started quickly.

- **Integration:** It seamlessly integrates with other Microsoft Office programs, including Excel and PowerPoint, for easy content entry and editing.

- **Mail Merge:** Use Word's mail merge function to create personalized documents such as letters and envelopes by merging data from Excel spreadsheets or Outlook contacts lists

- **Compatibility:** Word supports multiple file formats, including.docx and older formats like.doc and.rtf. It also allows for the import and export of PDF files, making it ideal for document interchange.

- **Accessibility:** Word's accessibility features, including as screen readers and text-to-speech, improve accessibility for people with disabilities.

9.3. Interface of Microsoft words

The Microsoft Word interface provides a user-friendly environment for document creation, editing, formatting, and review. An overview of the interface's key components is provided below:

1. **Title Bar:** The title bar at the top of the window displays the current document name and the application name, "Microsoft Word."

2. **Ribbon:** The Ribbon is a tabbed toolbar located below the title bar. It organizes commands and tools into tabs, including Home, Insert, Design, Layout, References, Mailings, Review, and View. Each tab is further broken into sets of similar instructions. To access the commands on any tab, simply click on it.

3. **Quick Access Toolbar:** Located next to or above the Ribbon, it provides shortcuts to frequently used functions including Save, Undo, and Redo. This toolbar can be customized by adding or removing commands as desired.

4. **Document Area:** The document area is the main window area for creating, editing, and viewing document material. This part displays your document's actual content, which includes text, photos, tables, and other features.

5. **Status Bar:** The status bar at the bottom of the window displays document information, including page number, word count, and viewing options. It also has zoom in/out capabilities for the document.

6. **Vertical and Horizontal Scroll Bars:** If your content goes beyond the visible area, scroll bars will appear on the right and bottom edges of the window. You can use the scroll bars to navigate the document.

7. **View Options:** Microsoft Word offers customizable view options, including Print Layout, Full Screen Reading, Web Layout, Outline, and Draft. You can switch between these views by clicking the buttons in the bottom-right corner of the window or selecting the relevant view from the Ribbon's View tab.

8. **Dialog Boxes and Panes:** Microsoft Word may display additional dialogue boxes and panels when formatting text, adding objects, or reviewing changes. These dialogue boxes and panes provide additional options and settings for the current job.

Some of the most frequently used tabs are in the Ribbon section which have variety of

9.3.1. File

The "File" option on Microsoft Word's Ribbon lets you manage your documents and execute activities including saving, printing, sharing, and accessing document properties. Here's an overview of the primary choices available under the "File" tab:

1. New: You can start a new document from scratch or use one of the many pre-designed templates accessible online or on your device.

2. Open: Accesses an existing document on your PC, One Drive, or other connected storage sites. You can also get recent documents from here.

3. **Save / Save As:** Allows you to save the current document under its current name or in a new name or location. "Save As"

4. **Print:** Displays the Print dialog box, where you can select printing choices and print the document.

5. **Share:** Offers options for sending the document to others by email, OneDrive, or other sharing sites.

6. **Export:** You can export the document in a variety of formats, including PDF, XPS, and other file types.

7. **Close:** Closes the current document.

8. **Account:** Provides information about your Microsoft account, including your profile, account settings, and options for signing in and out.

9. **Options:** Opens the Word Options dialog box, where you can customize settings related to the Word application, such as display, proofing, and language options.

10. **Feedback:** Allows you to provide feedback to Microsoft about the Word application.

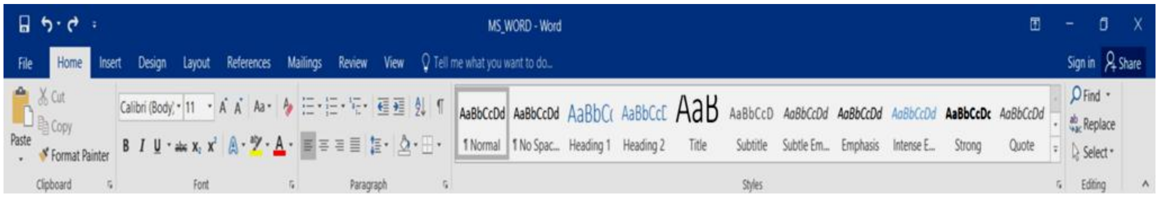
11. **Privacy:** Provides options for managing privacy settings and data collection in Word.

12. **Exit:** Closes the Word application.

These are the main options you'll find under the "File" tab in Microsoft Word's Ribbon. It serves as a central hub for managing your documents and accessing various file-related commands and settings.

9.3.2. Home

The "Home" tab in Microsoft Word's Ribbon is one of the most often utilized, giving you access to a variety of basic formatting and editing options. It includes instructions for formatting text, applying styles, changing paragraphs, and editing document content. The "Home" tab contains the following primary groups and commands:



1. Clipboard:

- Cut: (*use shortcuts* **Ctrl+x in Windows or Cmd + x in Mac OS)** Removes the selected content and stores it on the Clipboard.

- Copy: (*use shortcuts* **Ctrl+c in Windows or Cmd + c in Mac OS)** Copies the selected content to the Clipboard.

- Paste: (*use shortcuts* **Ctrl+p in Windows or Cmd + p in Mac OS)** Inserts the content from the Clipboard into the document.

- Format Painter: shortcut (**Alt + Ctrl + c**) Copies the formatting from one selection and applies it to another *using* shortcut (**Alt + Ctrl + v**).

2. Font:

- Font: Allows you to choose the font face for selected text.

- Font Size: Lets you adjust the size of the selected text.

- Bold, Italic, Underline: Applies bold, italic, or underline formatting to the selected text.

- Text Highlight Color: Applies a background color to the selected text.

- Font Color: Changes the color of the selected text.

4. Styles:

- Quick Styles: Provides quick access to predefined styles for text formatting.

- Change Styles: Allows you to apply different styles to your document.

5. Editing:

- Find, Replace: Helps you find specific text or replace it with different text.
- Select: Provides options for selecting specific elements in the document.
- Clear Formatting: Removes all formatting from the selected text.

6. Formatting:

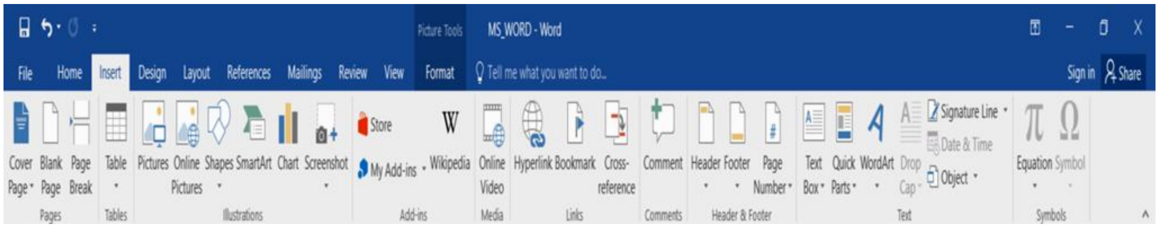
- Format Painter: Copies the formatting from one selection and applies it to another.
- Borders: Adds borders to selected paragraphs or text.
- Shading: Applies background shading to selected paragraphs or text.

7. View:

Enables you to transition between various document views, including Print Layout, Full Screen Reading, and Web Layout. The "Home" tab of the Ribbon in Microsoft Word contains the following primary groups and commands. It includes key tools for formatting text, paragraphs, and documents, as well as basic editing functions.

9.3.3. Insert

The "Insert" tab on Microsoft Word's Ribbon has a variety of tools and commands for adding different types of information to your document. The "Insert" tab contains the following key groups and commands:



1. Pages:

- Blank Page: Inserts a blank page at the cursor's position.

- Cover Page: Provides pre-designed cover page templates that you can insert into your document.

- Page Break: Inserts a manual page break, forcing content after the break onto the next page.

2. Tables:

- Table: Allows you to create a table by specifying the number of rows and columns.

- Excel Spreadsheet: Embeds an Excel worksheet into your Word document.

3. Illustrations:

- Pictures: Inserts an image file from your computer or device.

- Online Pictures: Allows you to search for and insert images from online sources like Bing Image Search.

- Shapes: Offers a variety of shapes that you can insert into your document.

- Icons: Provides access to a library of vector icons that you can insert.

- 3D Models: Allows you to insert and manipulate 3D models.

- Smart Art: Inserts pre-designed diagrams and graphic elements to visually represent information.

- Chart: Inserts a chart based on data from an Excel spreadsheet or other data sources.

4. Links:

- Hyperlink: Inserts a hyperlink to a webpage, file, or email address.

- Bookmark: Creates a bookmark within your document that you can link to.

5. Comments:

- Comment: Adds a comment at the cursor's position for collaboration and feedback purposes.

6. Text:

- Text Box: Inserts a text box that can contain text or graphics.

- Drop Cap: Applies a large initial letter at the beginning of a paragraph for decorative or stylistic purposes.

- Signature Line: Inserts a placeholder for a signature in the document.

7. Header & Footer:

- Header: Inserts a header at the top of the page.

- Footer: Inserts footer at the bottom of the page.

8. Symbols:

- Equation: Allows you to insert mathematical equations and symbols.

- Symbol: Inserts special characters and symbols into your document.

9. Page Number:

- Page Number: Inserts page numbers into the document footer or header.

10. Text:

- Date & Time: Inserts the current date and time into the document.

- Object: Embeds objects from other programs, such as Excel worksheets or PDF files.

The "Insert" tab of the Ribbon in Microsoft Word contains the following primary groups and commands. It offers a variety of choices for adding and enhancing material in your papers.

9.3.4. Design

The "Design" tab on Microsoft Word's Ribbon offers a variety of tools and instructions for formatting and decorating documents. It provides choices for customizing the overall appearance and feel of your document, such as themes, colors, fonts, and paragraph spacing. Here's a summary of the main categories and commands in the "Design" tab:

1. Document Formatting:

- Document Formatting: Provides access to different document themes that change the overall appearance of your document, including font styles, colors, and affects.

- Colors: Allows you to choose from different color schemes for your document.

- Fonts: Offers a selection of font combinations for headings and body text.

- Paragraph Spacing: Provides options to adjust the spacing between paragraphs.

2. Page Background:

- Page Color: Allows you to change the background color of the entire page.

- Watermark: Inserts a watermark, such as "Confidential" or "Draft," in the background of the document.

3. Page Borders:

- Page Borders: Opens a dialog box where you can add custom borders to the pages of your document.

4. Themes:

- Themes: Allows you to browse and apply different document themes, which include coordinated sets of fonts, colors, and effects.

5. Document Formatting:

- Document Formatting: Provides access to different document themes that change the overall appearance of your document, including font styles, colors, and affects.

6. Customize:

- Customize: Opens the "Modify Style" dialog box, allowing you to customize the formatting of individual styles used in your document.

7. Change Styles:

- Change Styles: Provides quick access to various style sets for different elements of your document, such as headings, titles, and body text.

8. Table Styles:

- Table Styles: Offers a selection of pre-designed styles for formatting tables in your document.

9. Page Setup:

- Page Setup: Opens the Page Setup dialog box, where you can adjust settings such as margins, orientation, and paper size.

10. Columns:

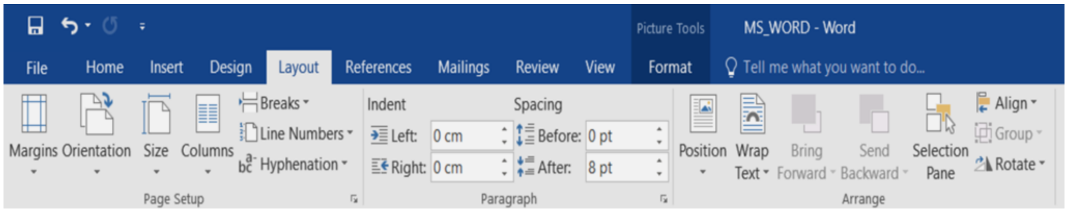
- Columns: Allows you to divide your document into multiple columns.

11. Breaks:

Uses several forms of breaks, like as page breaks, section breaks, and column breaks, to manage the layout of your content. The "Design" tab of the Ribbon in Microsoft Word contains the following primary groups and commands. It offers numerous possibilities for altering the design and layout of your document in order to obtain the desired visual style.

9.3.5. Layout

The "Layout" tab provides choices for adjusting page margins, orientation, size, and spacing, as well as managing sections and arranging items. The "Layout" tab contains the following key groups and commands:



Left: 0 cm

Right: 0 cm

Before: 0 pt

After: 8 pt

1. Page Setup:

- Margins: Allows you to set the margins for the entire document or selected sections.

- Orientation: Lets you choose between portrait (vertical) and landscape (horizontal) page orientation.

- Size: Allows you to change the paper size of your document.

- Columns: Divides the document into multiple columns.

- Breaks: Inserts different types of breaks, such as page breaks, section breaks, or column breaks.

2. Page Background:

- Watermark: Inserts a watermark, such as "Confidential" or "Draft," in the background of the document.

- Page Color: Allows you to change the background color of the entire page.

3. Paragraph:

- Indent: Adjusts the indentation of paragraphs.

- Spacing: Sets the spacing between paragraphs.

- Align: Aligns text to the left, center, right, or justified.

- Line Numbers: Adds line numbers to the document.

4. Arrange:

- Position: Allows you to position objects, such as images or shapes, relative to the text.

- Wrap Text: Controls how text flows around objects.

- Bring Forward/Send Backward: Changes the stacking order of objects.

- Rotate: Rotates selected objects.

5. Size:

- Height: Specifies the height of selected objects.

- Width: Specifies the width of selected objects.

6. Table:

- Insert: Allows you to insert tables into your document.

- Draw Table: Lets you draw a custom table shape.

- Convert to Text: Converts selected tables to text.

- Alignment: Aligns the contents of the selected cells within the table.

7. Table of Contents:

- Table of Contents: Inserts a table of contents based on the headings in your document.

- Update Table: Refreshes the table of contents to reflect any changes in the document.

8. Footnote:

Insert Footnote: Adds a footnote at the cursor's position. Show Footnote/Endnote Separator: This option allows you to show or hide the line that separates footnotes from the main text. The "Layout" tab of the Ribbon in Microsoft Word contains the following primary groups and commands. It offers numerous choices for tailoring the layout, formatting, and structure of your document to your exact requirements.

9.3.6. References

The "References" tab on the Microsoft Word Ribbon includes tools and commands for handling citations, bibliographies, and other reference-related aspects in your text. It allows you to input and format citations, create and manage a bibliography, include footnotes and endnotes, and generate a table of contents. The "References" tab contains the following main groups and commands:

1. Table of Contents:

- Table of Contents: Inserts a table of contents based on the headings in your document.

- Update Table: Refreshes the table of contents to reflect any changes in the document.

2. Footnotes:

- Insert Footnote: Inserts a footnote at the cursor's position.

- Show Footnote/Endnote Separator: Displays or hides the separator line between footnotes and the main text.

3. Citations & Bibliography:

- Insert Citation: Allows you to insert citations from your bibliography or add new sources to your document.

- Manage Sources: Opens the Source Manager, where you can view, edit, and add sources to your bibliography.

- Style: Lets you choose the citation style (e.g., APA, MLA) for your document.

- Bibliography: Inserts a bibliography or works cited list based on the sources in your document.

- Cross-reference: Allows you to insert references to captions, headings, footnotes, and other elements in your document.

4. Captions:

- Insert Caption: Inserts a caption below a selected object, such as a table or figure.

- Cross-reference: Allows you to insert references to captions, headings, footnotes, and other elements in your document.

5. Indexes:

- Mark Entry: Marks selected text for inclusion in an index.

- Insert Index: Inserts an index based on the marked entries in your document.

6. Table of Authorities:

- Mark Citation: Marks selected text as a citation for inclusion in a table of authorities.

- Insert Table of Authorities: Inserts a table of authorities based on the marked citations in your document.

7. Mailings:

- Mail Merge: Launches the Mail Merge Wizard to create personalized documents, such as letters or envelopes, for multiple recipients.

8. Review:

- Spelling & Grammar: Checks the spelling and grammar in your document.

- Word Count: Counts the number of words, characters, and pages in your document.

The "References" tab of the Ribbon in Microsoft Word contains the following primary groups and commands. It includes numerous tools for handling references, citations, and other aspects required for academic and professional documents.

9.3.7. Mailings

The "Mailings" tab on the Microsoft Word Ribbon includes tools and commands for executing mail merge operations, managing data sources, and creating tailored documents for mass dissemination. It includes options for connecting to data sources, inserting merge fields, evaluating merged documents, and completing the merge process. The "Mailings" tab contains the following primary groups and commands:

1. Start Mail Merge:

- Start Mail Merge: Launches the Mail Merge Wizard, guiding you through the process of creating personalized documents.

- Select Recipients: Allows you to choose a data source, such as an Excel spreadsheet or Outlook contacts, to use for the mail merge.

2. Write & Insert Fields:

- Insert Merge Field: Inserts merge fields into your document to pull data from the selected data source.
- Address Block: Inserts a predefined address block containing fields for recipient information.
- Greeting Line: Inserts a predefined greeting line based on the recipient's name.
- Insert Merge Field: Provides access to a list of merge fields available in the selected data source.

3. Preview Results:

- Preview Results: Allows you to preview how the merged document will appear with actual data.

- Highlight Merge Fields: Highlights merge fields in your document for easy identification.

4. Finish:

- Finish & Merge: Completes the mail merge process and allows you to choose how to finalize the merged documents.

- Edit Individual Documents: Generates separate documents for each record in the data source, allowing you to review and edit them individually.

- Send Email: Sends the merged documents as email messages directly from Word.

5. Write & Insert Fields:

- Insert Merge Field: Inserts merge fields into your document to pull data from the selected data source.

- Rules: Allows you to define rules for conditional mail merge operations.

6. Preview Results:

- Check for Errors: Checks for errors in the merge process, such as missing data or formatting issues.

- Preview Results: Allows you to preview how the merged document will appear with actual data.

7. Finish:

- Finish & Merge: Completes the mail merge process and allows you to choose how to finalize the merged documents.

- Edit Individual Documents: Generates separate documents for each record in the data source, allowing you to review and edit them individually.

- Send Email Messages: Sends the merged documents as email messages directly from Word.

These are the main groups and commands you'll find in the "Mailings" tab of the Ribbon in Microsoft Word. It provides extensive features for performing mail merge operations and creating personalized documents for mass mailing or distribution.

9.3.8. Review

The "Review" tab on the Microsoft Word Ribbon includes tools and instructions for collaborative document review and modification. It includes capabilities for checking spelling and grammar, logging changes to the text, adding comments, and protecting the document's information. The "Review" tab contains the following primary groups and commands:

1. Proofing:

- Spelling & Grammar: Checks the spelling and grammar in your document, highlighting potential errors and offering suggestions for correction.
- Thesaurus: Opens the Thesaurus tool, allowing you to find synonyms and antonyms for selected words.
- Word Count: Counts the number of words, characters, and pages in your document.
- Translate: Translates selected text or the entire document into another language using Microsoft Translator.

2. Comments:

- New Comment: Inserts a comment at the cursor's position for collaboration and feedback purposes.
- Delete: Deletes selected comments from the document.
- Previous/Next: Navigates between comments in the document.
- Show Comments: Displays all comments in the document for review.
- Reviewing Pane: Opens the Reviewing Pane to view comments and changes made to the document.

3. Tracking:

- Track Changes: Turns on Track Changes mode, which tracks all edits and revisions made to the document.
- Simple Markup: Displays tracked changes and comments as simple indicators in the document.

- All Markup: Displays tracked changes and comments with markup for detailed review.
- No Markup: Hides tracked changes and comments from view, displaying the final version of the document.
- Accept/Reject: Allows you to accept or reject changes made to the document.

4. Changes:

- Accept: Accepts the selected change in the document.
- Reject: Rejects the selected change in the document.
- Previous/Next: Navigates between tracked changes in the document.

5. Compare:

- Compare: Compares the current document with another document or a previous version to identify differences.
- Combine: Combines multiple documents into one, highlighting differences between them.

6. Protect:

- Restrict Editing: Restricts editing permissions for the document, allowing you to control who can make changes.
- Protect Document: Protects the document with a password to prevent unauthorized access or modifications.

The "Review" tab of the Ribbon in Microsoft Word contains the following primary groups and commands. It has many capabilities for reading and amending documents collaboratively, assuring accuracy, clarity, and security.

9.3.9. View

The "View" tab on Microsoft Word's Ribbon contains tools and commands for changing how your document is displayed and modified on screen. It allows you to change the layout of the document, zoom in and out, make select elements visible, and switch between viewing modes. Here's an overview of the major groups and commands you'll find under the "View" tab:

1. Document Views:

- Print Layout: Displays the document as it would appear when printed, including margins, headers, and footers.
- Read Mode: Optimizes the document for reading, with features like narrow margins and full-screen view.
- Web Layout: Displays the document as it would appear in a web browser, with text and images flowing dynamically.
- Outline: Shows the document's structure in outline form, allowing you to view and navigate through headings and subheadings.
- Draft: Displays the document without any formatting, providing a simplified view for editing.

2. Show:

- Ruler: Shows or hides the horizontal and vertical rulers at the top and left side of the document window.
- Gridlines: Displays or hides gridlines to help align objects and text.
- Navigation Pane: Opens the Navigation Pane, allowing you to search for and navigate through headings, pages, and other document elements.
- Thumbnails: Displays thumbnails of each page in the document for quick navigation.

3. Zoom:

- Zoom: Allows you to adjust the zoom level of the document to make the text and layout larger or smaller.
- One Page: Displays the entire page in the document window.
- Two Pages: Displays two pages side by side in the document window.
- Page Width: Adjusts the zoom level to fit the width of the page in the document window.

4. Window:

- New Window: Opens a new window with a separate view of the same document, allowing you to work on different parts of the document simultaneously.
- Arrange All: Arranges all open document windows side by side for easy comparison and editing.
- Split: Splits the document window into two panes, allowing you to view different parts of the document simultaneously.

5. Macros:

- Macros: Allows you to record and run macros, which are sequences of actions that can be automated to perform repetitive tasks.

6. Show/Hide:

- Show/Hide Paragraph Marks: Displays or hides formatting marks, such as paragraph marks, spaces, and tab characters.
- Show/Hide All Nonprinting Characters: Displays or hides all nonprinting characters, including formatting marks, page breaks, and hidden text.

These are the main groups and commands you'll find in the "View" tab of the Ribbon in Microsoft Word. It provides extensive features for customizing the document's appearance and layout to suit your editing and viewing preferences.

9.3.10. Find and Replace

Certainly! Here's a more detailed step-by-step guide to finding and replacing text in Microsoft Word:

1. Open Your Document: Launch Microsoft Word and open the document in which you want to find and replace text.

2. Access the Find and Replace Tool:

- Press `Ctrl + H` on your keyboard. This shortcut directly opens the Find and Replace dialog box.
- Alternatively, you can go to the "Home" tab on the ribbon.
- In the Editing group, locate and click on the "Replace" option. This will also open the Find and Replace dialog box.

3. Enter the Text to Find:

- In the Find what: field, type the text you want to find in your document.

4. Enter the Text to Replace:

- In the Replace with: field, type the text you want to replace the found text with.

5. Review Additional Options (Optional):

- Click on the "More >>" button if you want to see additional options.
- Here you can specify formatting options, such as font style or size, if you only want to find and replace text with certain formatting.

6. Start Finding and Replacing:

- Click on the "Find Next" button to locate the first instance of the text you entered in the Find what: field.
- Review the instance to confirm if you want to replace it.
- To replace this instance, click on the "Replace" button. If you want to replace all instances throughout the document, click on "Replace All".
- If you don't want to replace the instance found, click "Find Next" to continue searching.

7. Review Changes:

- Word will display each change it makes, allowing you to review and confirm them.

8. Close the Find and Replace Dialog Box:

- Once you've finished finding and replacing text, close the Find and Replace dialog box by clicking the "Close" button.

9. Save Your Document:

After making all desired changes, remember to save your document by clicking on the Save button or using the keyboard shortcut `Ctrl + S. Following these steps will help you efficiently find and replace text in your Microsoft Word document. In Microsoft Word, you can access the Thesaurus to find synonyms (words with similar meanings) and antonyms (words with opposite meanings) for the selected word. Here's how to use the Thesaurus feature:

1. **Select the Word**:

   Click on the word for which you want to find synonyms or antonyms. You can either double-click on the word or click and drag to select it.

2. Access the Thesaurus:

- Right-click on the selected word.
- From the context menu that appears, select "Synonyms" to view a list of synonyms for the word.
- If you want to explore more options or view antonyms as well, select "Thesaurus" from the context menu.

3. Review Synonyms and Antonyms:

- The Thesaurus pane will appear on the right side of the screen, displaying a list of synonyms for the selected word.
- To view antonyms, click on the "Antonyms" tab at the top of the Thesaurus pane.

4. Choose a Synonym or Antonym:

- Click on any synonym or antonym in the list to replace the selected word with it.

- Word will automatically replace the selected word with your chosen synonym or antonym.

5. Close the Thesaurus:

- After finding suitable synonyms or antonyms, you can close the Thesaurus pane by clicking the "X" button in the upper-right corner of the pane or by clicking anywhere outside the pane.

By following these steps, you can easily use the Thesaurus feature in Microsoft Word to find synonyms and antonyms for words in your document.

Mail- merge

Mail merge in Microsoft Word is a powerful feature that allows you to create personalized documents, such as letters, envelopes, labels, and emails, by combining a main document with a data source, such as an Excel spreadsheet or Outlook contacts list. Here's a basic guide on how to perform a mail merge in Word:

1. Prepare Your Data Source:

- Ensure your data source (e.g., Excel spreadsheet, Outlook contacts) is properly organized with the information you want to include in your merged document.
- Each column in your data source should represent a specific piece of information, such as names, addresses, or other personalized data.

2. Create Your Main Document:

- Open Microsoft Word and create a new document.
- In this document, you'll design the layout and content that will remain constant across all merged documents.
- Insert placeholders, called merge fields, where you want the personalized information from your data source to appear. To insert a merge field, go to the "Mailings" tab, click on "Insert Merge Field," and select the field name from your data source.

3. Start the Mail Merge Process:

- Once your main document is ready, go to the "Mailings" tab on the Word ribbon.
- Click on "Start Mail Merge" and select the type of document you want to create (e.g., Letters, Email Messages, Envelopes, Labels).
- Choose the option "Step-by-Step Mail Merge Wizard" or "Mail Merge Wizard" to guide you through the process.

4. Select Your Data Source:

- In the Mail Merge pane that appears on the right, click on "Select Recipients" and choose the option that corresponds to your data source.
- Browse to locate and select your data source file.

5. Insert Merge Fields:

- After selecting your data source, click on "Insert Merge Field" to insert merge fields into your main document.
- Choose the fields from your data source that you want to merge into your document.

6. Preview Your Merge:

- Click on "Preview Results" to see how your merged document will look with actual data from your data source.
- Use the navigation buttons in the Mail Merge pane to navigate through each record.

7. Complete the Merge:

- Once you're satisfied with the preview, click on "Finish & Merge" and choose the appropriate option:
- "Edit Individual Documents" to merge the documents directly to a new document.
- "Send Email Messages" if you're creating email messages.
- "Print Documents" if you're creating printed documents.

8. Save or Send Your Merged Documents:

- Save the merged document or send the emails, depending on the options you selected.
- By following these steps, you can perform a mail merge in Microsoft Word to create personalized documents efficiently.

9.4. Table and Chart

In Microsoft Word, you can easily create tables and charts to organize and visualize your data. Here's how to do it:

Creating Tables:

1. Insert a Table:

- Place your cursor in the document where you want to insert the table.
- Go to the "Insert" tab on the ribbon.
- Click on the "Table" button.
- Select the number of rows and columns for your table from the grid that appears.

2. Enter Data**:**

- Once the table is inserted, click into any cell and start typing to enter data.
- You can navigate between cells using the arrow keys or by clicking with your mouse.

3. Modify Table Properties:

- To modify the properties of the table (e.g., borders, shading), click anywhere in the table to activate the "Table Design" and "Table Layout" tabs on the ribbon.
- Here, you can change the table's appearance, alignment, and other properties.

Creating Charts:

1. Prepare Your Data:

- If you're creating a chart, you'll need to organize your data in a way that makes sense for the type of chart you want to create.
- Typically, your data should be organized into rows and columns with labels in the first row or column.

2. Select Your Data:

- Highlight the data you want to include in your chart.

3. Insert a Chart:

- With your data selected, go to the "Insert" tab on the ribbon.
- Click on the type of chart you want to insert from the "Chart" group (e.g., Column, Line, Pie).

4. Customize Your Chart:

- After inserting the chart, you can customize it further.
- Use the "Chart Design" and "Chart Format" tabs that appear on the ribbon when the chart is selected to change colors, styles, labels, and more.

5. Modify Data:

- If you need to change the data that's included in the chart, you can do so by editing the data directly in the Excel spreadsheet that's embedded in the Word document.
- Double-click on the chart to open the linked Excel spreadsheet.

6. Resize and Move Your Chart:

- Click and drag the chart to move it to a different location in your document.
- Use the sizing handles on the corners and sides to resize the chart as needed.

By following these steps, you can easily create and customize tables and charts in Microsoft Word to effectively present your data.

Handling graphics

Handling graphics in Microsoft Word allows you to add images, shapes, Smart Art, screenshots, and other visual elements to your documents. Here's a guide on how to work with graphics in Word:

9.4.1. Inserting Graphics:

1. Insert Pictures:

- Place your cursor where you want to insert the picture.
- Go to the "Insert" tab on the ribbon.
- Click on the "Pictures" button to select an image file from your computer. Alternatively, choose "Online Pictures" to search for images online, or "Screenshot" to capture a screenshot of your screen.

2. Insert Shapes:

- To insert shapes like rectangles, circles, arrows, etc.:
- Go to the "Insert" tab.
- Click on the "Shapes" button in the Illustrations group.
- Select the shape you want to insert, then click and drag on the document to draw it.

3. Insert Smart Art:

- Smart art allows you to create diagrams and graphics with predefined layouts and formatting.
- Go to the "Insert" tab.
- Click on the "smart art" button in the illustrations group.
- Choose a smart art graphic, and then click "ok" to insert it into your document.

Editing Graphics:

1. Resizing and Moving:

- Click on the graphic to select it.
- Use the sizing handles at the corners and sides to resize the graphic.

- Click and drag the graphic to move it to a different location in your document.

2. Formatting:

- With the graphic selected, you can format it using options in the "Format" tab that appears on the ribbon when the graphic is selected.
- You can change the fill color, line color, line style, shadows, and other properties.

3. Arranging and Grouping:

- Use the "Bring to Front," "Send to Back," "Group," and "Ungroup" options in the "Format" tab to arrange and group multiple graphics.

4. Crop and Rotate:

- To crop or rotate a picture:
- Click on the picture to select it.
- Go to the "Format" tab.
- Use the "Crop" or "Rotate" options to adjust as needed.

Captions and Labels:

1. Add Captions:

- Right-click on the graphic.
- Select "Insert Caption" from the context menu.
- Enter a caption in the dialog box that appears.

2. Add Labels:

- For charts, tables, and Smart Art, you can add data labels or axis titles by clicking on the respective elements and using the chart tools that appear.

Saving and Exporting Graphics:

1. Save Embedded Graphics:

- If you insert a picture from your computer, it will be embedded in the Word document by default.
- To save the document with the embedded pictures, simply save the Word document.

2. Save Linked Graphics:

- If you insert a picture that's linked to an external file, you'll need to keep the linked file in the same location if you plan to share the document.
- When saving the document, Word may prompt you to update the links or embed the pictures.
- By following these steps, you can effectively handle graphics in Microsoft Word to enhance the visual appeal and clarity of your documents.

Printing and formatting a document

Printing and formatting a document in Microsoft Word is a fundamental skill. Here’s a basic guide:

❖ **Formatting a Document:**

1. Font Styles and Sizes:

- Select the text you want to format.
- Choose a font style, size, and color from the Font group on the Home tab.

2. Paragraph Formatting:

- Adjust alignment (left, center, and right, justified) from the Paragraph group on the Home tab.
- Set line spacing, indentation, and spacing before/after paragraphs using the options in the Paragraph group.

3. Headers and Footers:

- Go to the Insert tab and click on Header or Footer to add them to your document.
- Customize headers and footers with text, page numbers, and date/time.

4. Page Layout:

- Adjust margins, orientation (portrait or landscape), and paper size from the Layout tab.
- Set up page breaks and columns if needed.

5. Styles:

- Use predefined styles from the Styles group on the Home tab to quickly format text consistently.

9.4.2. Printing a Document

1. Preview:

- Click on File > Print to see a preview of your document and adjust settings before printing.

2. Printer Settings:

- Select the printer you want to use.
- Choose the number of copies, page range, and other settings as needed.

3. Print Options:

- Customize print options like color/black and white, duplex printing, and paper tray selection.

4. Page Setup:

- Ensure that the page setup matches your document layout, including paper size and orientation.

5. Print:

- Click the Print button to send the document to the printer.

❖ **Tips:**

- Save Ink:

- Use the Print Preview feature to check how the document will look when printed.
- Print in draft mode or grayscale to save ink.

- Page Breaks:

- Use page breaks to control where pages end. Insert them from the Insert tab.
- Headers and Footers:
- Customize headers and footers to include important information like page numbers, document titles, and author names.

- Spell Check:

- Run a spell check before printing to catch any typos or errors.

- Print to PDF:

- If you don’t have a physical printer, you can print your document to PDF format using a virtual PDF printer.

Following these steps should help you effectively format and print your document in Microsoft Word.

Saving documents as TXT and rich text

Converting a Word document into various formats like text (Plain Text) and rich text format (RTF) can be done easily within Microsoft Word. Here's how:

- **Converting to Plain Text:**
  1. Open your Word document.
  2. Click on "File" in the top-left corner of the screen.
  3. Select "Save As" from the menu.
  4. Choose a location to save your file and enter a name for it.
  5. In the "Save as type" dropdown menu, select "Plain Text (*.txt)".
  6. Click "Save".
  7. A dialog box will appear. Choose the appropriate encoding (usually UTF-8) and click "OK".
- **Converting to Rich Text Format (RTF):**

1. Open your Word document.
2. Click on "File" in the top-left corner of the screen.
3. Select "Save As" from the menu.
4. Choose a location to save your file and enter a name for it.
5. In the "Save as type" dropdown menu, select "Rich Text Format (*.rtf)".
6. Click "Save".

Your document will now be saved in the selected format at the specified location.

- **Additional Tips:**

- Preserving Formatting: RTF format preserves most formatting, while plain text will remove formatting altogether.
- Reviewing Conversion: After saving in the desired format, open the file in a text editor to ensure the conversion was successful and the formatting meets your needs.
- File Size: Plain text files are usually smaller in size compared to RTF or Word documents because they do not contain any formatting information.

Following these steps should allow you to easily convert your Word document into plain text or RTF format. Let me know if you need further assistance!

9.4.3. Shortcuts key

Keyboard shortcuts in Microsoft Word can significantly improve your efficiency when working with documents. Here are some common keyboard shortcuts:

- **Basic Text Editing:**

- Ctrl + C: Copy selected text.
- Ctrl + X: Cut selected text.
- Ctrl + V: Paste copied or cut text.
- Ctrl + Z: Undo your last action.
- Ctrl + Y: Redo your last undone action.
- Ctrl + A: Select all text in the document.
- Ctrl + B: Bold selected text.
- Ctrl + I: Italicize selected text.

- Ctrl + U: Underline selected text.
- Ctrl + S: Save the document.

▪ **Navigation and Selection:**

- Arrow Keys: Move the cursor in the direction of the arrow.
- Ctrl + Left/Right Arrow: Move the cursor one word to the left or right.
- Ctrl + Up/Down Arrow: Move the cursor to the beginning or end of the paragraph.
- Shift + Arrow Keys: Select text in the direction of the arrow.
- Ctrl + Shift + Arrow Keys: Select text one word at a time.

▪ **Formatting:**

- Ctrl + E: Align text center.
- Ctrl + L: Align text left.
- Ctrl + R: Align text right.
- Ctrl + J: Justify text.
- Ctrl + 1: Single line spacing.
- Ctrl + 2: Double line spacing.
- Ctrl + 5: 1.5 line spacing.

▪ **Document Navigation:**

- Ctrl + Home: Move to the beginning of the document.
- Ctrl + End: Move to the end of the document.
- Ctrl + Page Up: Move up one page.
- Ctrl + Page Down: Move down one page.
- Ctrl + G: Go to a specific page, line, or bookmark.

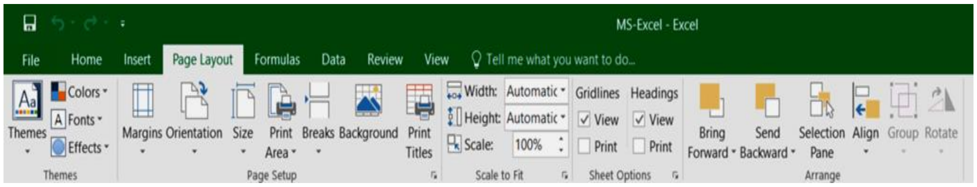
▪ **Miscellaneous:**

- Ctrl + P: Print the document
- Ctrl + F: Open the Find dialog box
- Ctrl + H: Open the Replace dialog box
- Ctrl + K: Insert a hyperlink.
- Ctrl + N: Create a new document

Remember, these are only a few of the numerous keyboard shortcuts accessible in Microsoft Word. Learning and implementing them can significantly boost your productivity. In Word's options, you may also see a full list of accessible shortcuts and change them.

9.5. Microsoft Excel

Microsoft Excel is a cornerstone of spreadsheet software, providing users with a sophisticated platform for data management, analysis, and visualization. Excel is based on a sequence of worksheets within workbooks, which provide a grid-like interface for entering, manipulating, and interpreting data. This grid is made up of cells, each of which can hold a variety of data kinds, including simple text and numbers, dates, formulae, and functions. In terms of functions, Excel has a large library of pre-built functions that serve to a wide range of computing needs, including fundamental arithmetic, statistical analysis, and advanced logical processes. Users can easily do calculations and alter data with functions such as SUM, AVERAGE, VLOOKUP, and IF, which facilitate operations ranging from financial modeling to inventory management.



Furthermore, Excel's capabilities go beyond raw data processing; it also provides advanced tools for data visualization and analysis. Users can use beautiful charts, graphs, and pivot tables to transform complex statistics into meaningful graphics, allowing for better decision-making and transmission of insights. Conditional formatting improves data display by dynamically emphasizing trends, outliers, and patterns based on defined criteria. Excel also promotes collaboration and data sharing by allowing several users to work on a shared workbook at the same time, as well as easy connection with cloud storage services such as OneDrive. Excel is a vital tool for financial analysis, project management, and scientific research, allowing users to maximize the value of their data and make educated decisions in a dynamic and interconnected environment. The main functions and components of Microsoft Excel:

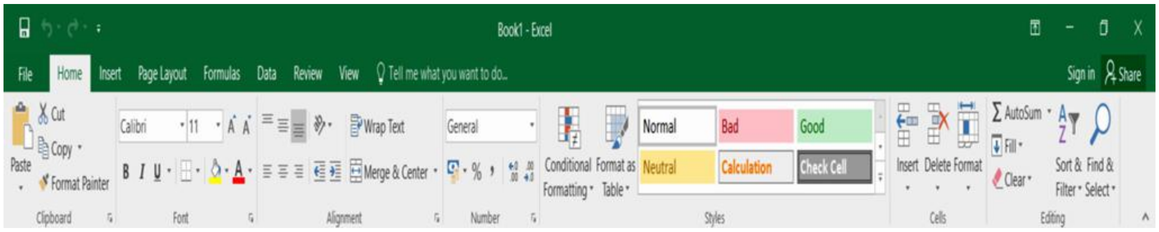
1. **Worksheets and Workbooks:** Excel documents are divided into workbooks, which each include one or more worksheets (also called spreadsheets). Each worksheet is made up of a grid of cells arranged in rows and columns, into which you may insert and alter data.
2. **Cells, Rows, and Columns:** Cells are distinct components in a worksheet that allow you to enter data, make calculations, and show results. Cells are organized horizontally into rows, and vertically into columns.
3. **Formulas and Functions:** Excel allows you to use formulae and functions to calculate your data. Formulas are expressions that execute mathematical operations or change data, whereas functions are established formulas that do precise calculations. Excel has a large number of built-in functions for diverse purposes, including SUM, AVERAGE, IF, VLOOKUP, and many more.
4. **Charts and Graphs**: Excel has capabilities for creating a variety of charts and graphs to visually portray your data. You can customize bar charts, line graphs, pie charts, scatter plots, and other graphics to meet your specific requirements.
5. **Data Analysis Tools**: Excel includes built-in data analysis and manipulation tools like sorting, filtering, and pivot tables. These tools let you organize and summarize huge datasets fast and efficiently.
6. **Data Validation:** Excel allows you to create data validation rules that restrict the kind and format of data entered into cells. This ensures data correctness and uniformity in your spreadsheets.
7. **Conditional Formatting:** Excel's conditional formatting function lets you apply formatting styles to cells depending on certain criteria. This makes it simple to visually highlight key data or patterns in your spreadsheets.
8. **Collaboration and Sharing:** Excel provides collaboration and sharing tools that let many users to work on the same workbook at the same time. You can also share your workbooks with others by email, OneDrive, or other file sharing services. In Microsoft Excel, the Ribbon is a collection of tabs that each contains sets of related instructions. Here's an overview of the default Ribbon tabs in Excel, including their key functions:

9. 1. File Tab:

- Contains commands related to file management, such as opening, saving, printing, and sharing workbooks.
- Provides access to options for customizing Excel settings and preferences.

2. Home Tab:

- Includes commonly used commands for formatting, styling, and editing data.
- Contains groups like Clipboard (cut, copy, paste), Font (font style, size, color), Alignment (text alignment), and Number (number formatting).



3. Insert Tab:

- Contains commands for inserting various elements into the worksheet, such as tables, charts, shapes, and images.
- Includes groups like Tables, Charts, Spark lines, and Illustrations.

4. Page Layout Tab:

- Provides options for controlling the layout and appearance of the printed worksheet.
- Includes commands for setting page orientation, margins, and print titles, as well as themes and page setup options.

5. Formulas Tab:

- Contains functions and commands for working with formulas and mathematical operations.
- Includes functions organized into categories like Financial, Logical, Text, Date & Time, and more.
- Also includes tools for formula auditing, calculation, and defining names.

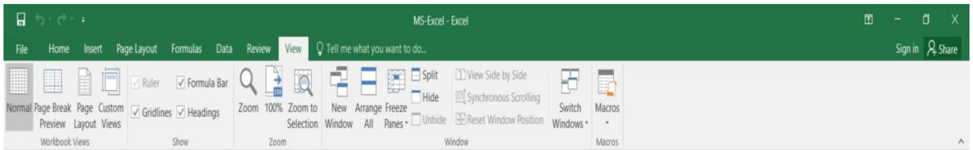
6. Data Tab:

- Contains commands for importing, sorting, filtering, and analyzing data.
- Includes tools for data validation, text-to-columns, subtotaling, and removing duplicates.
- Also includes connections to external data sources and tools for managing data models.

7. Review Tab:

- Contains tools for proofing, reviewing, and collaborating on workbooks.
- Includes commands for spell checking, comments, track changes, and protecting sheets and workbooks.

8. View Tab:



- Provides options for controlling the visual display and layout of the worksheet.
- Includes commands for zooming, arranging windows, and changing view modes like Normal, Page Layout, and Page Break Preview.
- Also includes options for gridlines, headings, and macros.

9. Developer Tab:

- This tab is hidden by default and needs to be enabled manually.
- Provides advanced tools for working with macros, add-ins, form controls, and ActiveX controls.
- Includes commands for Visual Basic Editor, macros, add-ins, and controls.
- These Ribbon tabs and their associated groups contain a wide range of tools and commands to help users perform various tasks and operations within Microsoft Excel.

In Microsoft Excel, the "File" tab, also known as the "Backstage view," is located at the top-left corner of the Excel window. Clicking on the "File" tab opens the File menu, providing access to various commands related to file management, settings, and options. Here's what you'll typically find in the File tab:

1. **New:** This option allows you to create a new workbook or choose from a variety of templates to start with, such as Blank workbook, recent templates, or templates from Office.com.
2. **Open:** You can open existing Excel workbooks from your computer or cloud storage services like OneDrive or SharePoint. It also displays a list of recent files for quick access.
3. **Save and Save As**: Save allows you to save changes made to the current workbook, while Save As allows you to save the workbook with a different name or location.
4. **Print:** Access printing options to print the current workbook. You can set print settings, select the printer, and preview how the workbook will appear when printed.
5. **Share:** Share the workbook with others by sending it as an email attachment, saving it to OneDrive or SharePoint, or presenting it online.
6. **Export:** Export the workbook to a different file format such as PDF, CSV, or XML.
7. **Close:** Close the current workbook.
8. **Account:** View information about your Microsoft account and subscription, and sign out or switch accounts.
9. **Options:** Access Excel's settings and preferences, where you can customize various aspects of the application such as formulas, proofing, and add-ins.
10. **Feedback:** Provide feedback or suggestions to Microsoft about Excel.
11. **Exit:** Close Excel application.

The "File" tab in Excel serves as a consolidated area for managing workbooks, accessing settings, and executing file-related operations. It is frequently the beginning point for creating, opening, and saving Excel files, as well as accessing additional features and choices in the application. Excel, while best known as a spreadsheet application, can also be utilized as database software in specific instances. Let's look at how Excel can perform efficiently as a database.

1. Data Organization:

- Rows and Columns**:** Excel's grid structure allows for the organization of data into rows and columns, akin to traditional database tables. Each row represents a record (an entry or instance of data), while each column represents a field (an attribute or characteristic of the data).
- Tables: Excel's Table feature offers structured formatting and functionality, allowing users to easily manage and analyze data within defined ranges.

2. Data Entry and Editing:

- User-Friendly Interface: Excel's familiar interface makes it accessible for users to enter and edit data.
- Data Validation: Excel provides options for data validation, enabling users to enforce rules and restrictions on data entry to maintain data integrity.

3. Data Analysis and Manipulation:

- Formulas and Functions: Excel boasts a vast library of built-in functions and formulas for performing calculations, data manipulation, and analysis. These can be utilized to derive insights from the data stored in Excel.
- Filtering and Sorting: Users can filter and sort data based on specific criteria, facilitating analysis and visualization of subsets of data.
- PivotTables and Pivot Charts: PivotTables allow for dynamic summarization and analysis of large datasets, while Pivot Charts provide graphical representations of the summarized data, aiding in visual analysis.

4. Data Import and Export:

- External Data Connections: Excel supports connections to external data sources such as databases, web services, and other Excel files. This allows users to import data into Excel from various sources or establish links to live data for real-time updates.

- Exporting: Users can export Excel data to various formats such as CSV (Comma-Separated Values), TXT (Plain Text), or PDF (Portable Document Format), facilitating data sharing and collaboration.

5. Collaboration and Sharing:

- Sharing Workbooks: Excel workbooks can be shared among multiple users, allowing for collaborative data entry, editing, and analysis.
- Track Changes: Excel includes features for tracking changes made to the workbook by different users, enabling version control and auditing of edits.

6. Limitations:

- Scalability: While Excel is suitable for small to medium-sized datasets, it may encounter performance issues with extremely large datasets due to memory and processing constraints.
- Concurrent Access: Excel does not support concurrent access by multiple users as robustly as dedicated database management systems.
- Data Integrity: Maintaining data integrity (e.g., enforcing referential integrity, preventing data duplication) in Excel requires manual effort and is not as seamless as in dedicated database software.

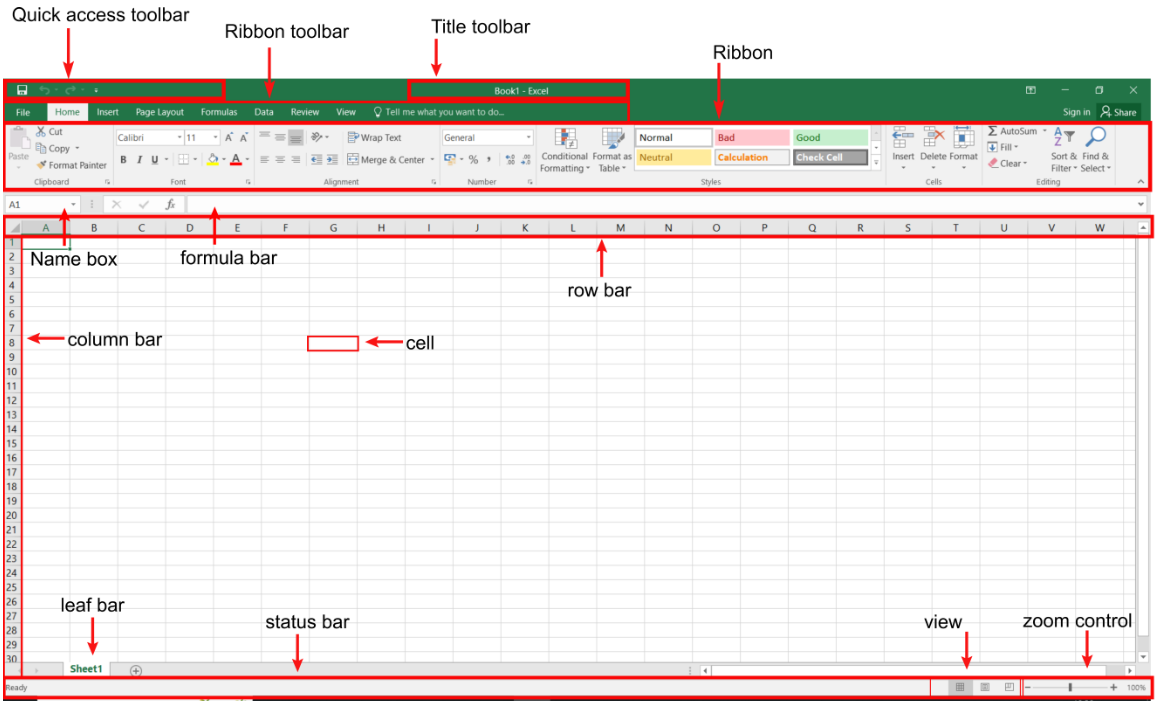
In summary, while Excel may not offer all the advanced features of dedicated database management systems, it serves as a versatile and user-friendly solution for managing, analyzing, and visualizing data in various contexts, particularly for smaller-scale databases and scenarios where ease of use and familiarity are paramount.

9.5.1. Creating worksheet in MS Excel:

The creating of work sheet in Microsoft Excel is a straightforward process. The involved in this process are:

Step 1: Open Microsoft Excel

- Launch Microsoft Excel on your computer. You can typically find it in the Start menu on Windows or in the Applications folder on macOS.



Step 2: Create a New Workbook

- Once Excel is open, you'll see a blank workbook with a grid of cells.
- If you want to start with a completely blank workbook, you're ready to begin entering your data. If you want to start with a template or a specific type of workbook, you can choose from the available templates or browse online templates.

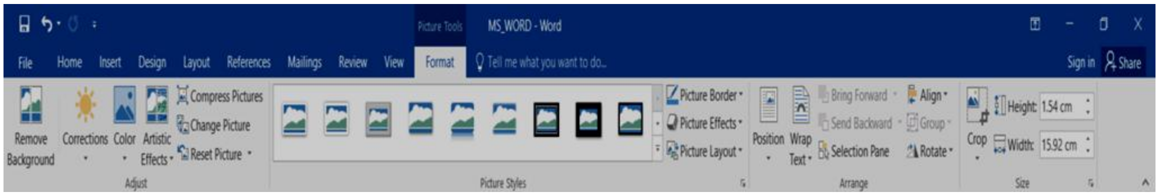
Step 3: Enter Data and Labels

- Click on a cell to select it, and start typing to enter your data.
- Use the Tab key to move to the cell to the right or Enter to move to the cell below.
- You can also copy and paste data from another source, such as a text document or another Excel workbook.

Step 4: Format Your Worksheet

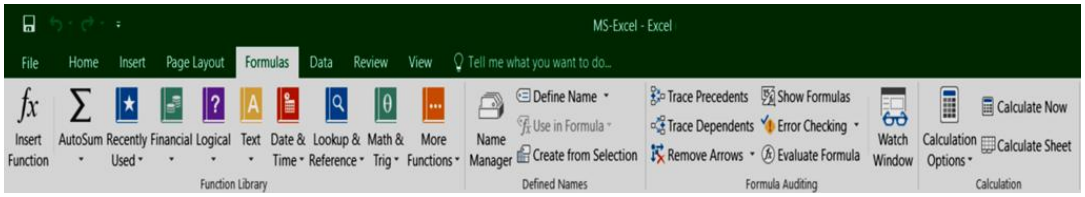
- Format your data as needed by selecting cells, rows, or columns and using the formatting options in the Home tab.

- You can change the font, font size, font color, cell alignment, borders, and more to make your worksheet easier to read and understand.

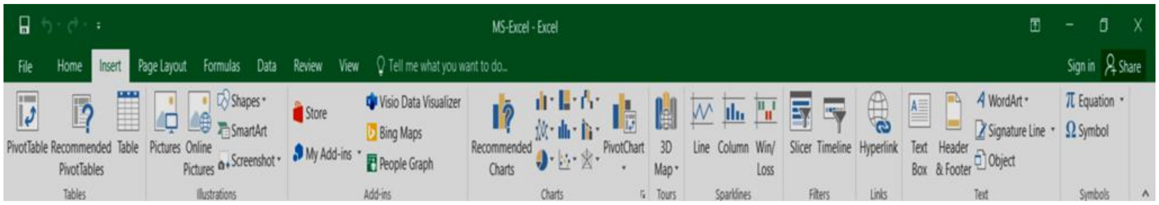


Step 5: Add Formulas and Functions

- Use formulas and functions to perform calculations or manipulate your data.
- Start a formula with an equal sign (=) and use cell references, mathematical operators (+, -, *, /), and functions to perform calculations.



Step 6: Insert Rows, Columns, and Worksheets (if needed)



- To insert a row or column, right-click on the row or column header where you want to insert the new row or column, and select "Insert" from the context menu.
- To add a new worksheet, click on the "+" icon at the bottom of the Excel window, or right-click on an existing worksheet tab and select "Insert" from the context menu.

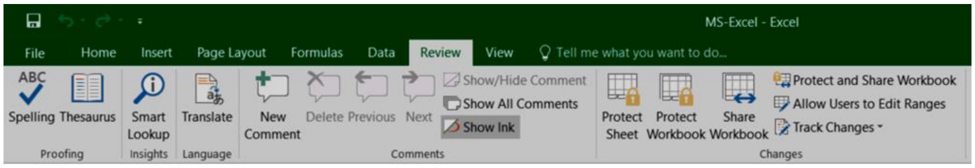
Step 7: Save Your Workbook

- Once you've created your worksheet and entered your data, it's important to save your work.
- Click on File > Save As, choose a location to save your workbook, enter a file name, and click Save.

- You can also use the keyboard shortcut Ctrl + S (Cmd + S on Mac) to save your workbook.

Step 8: Review and Proofread

- Before finalizing your worksheet, take some time to review and proofread your data, formulas, and formatting.
- Make any necessary adjustments to ensure that your worksheet is accurate and well-presented.



Step 9: Print or Share Your Worksheet

- Once you're satisfied with your worksheet, you can print it or share it with others.
- Click on File > Print to print your worksheet, or use the Share options to send it via email, share it on OneDrive, or collaborate with others in real-time.

By following these steps, you can create a new worksheet in Microsoft Excel and start entering and organizing your data.

9.5.2. Entering dates

Entering dates in Microsoft Excel is simple. You can manually type dates into cells, use keyboard shortcuts, or use Excel's built-in date functions. Here's how you can enter dates:

1. Manual Entry:

- Click on the cell where you want to enter the date.
- Type the date in one of the following formats:
- Month/Day/Year: For example, 6/1/2024 for June 1, 2024.
- Day/Month/Year: For example, 1/6/2024 for January 6, 2024. (This format depends on your regional settings.)
- Press Enter to confirm.

2. Keyboard Shortcut:

- You can use a keyboard shortcut to enter the current date into a cell:
- Ctrl +; (semicolon): This shortcut automatically inserts the current date into the selected cell.
- After pressing the shortcut, press Enter to confirm.

3. Using Excel's Built-in Date Functions:

- Excel offers several date functions that you can use to enter or manipulate dates:
- TODAY: This function returns the current date.
- DATE (year, month, day): This function constructs a date using the provided year, month, and day values.
- EOMONTH (start date, months): This function returns the last day of the month, a specified number of months before or after the start date.
- To use a date function:
- Click on the cell where you want to enter the date.
- Type the function with its arguments, for example: =TODAY()` or `=DATE(2024, 6, 1).
- Press Enter to confirm.

4. Date Picker (Excel for Windows):

- In Excel for Windows, you can use the Date Picker tool to select a date from a calendar:
- Click on the cell where you want to enter the date.
- Click on the drop-down arrow in the cell's formula bar.
- Select the desired date from the calendar that appears.

Note:

- ➢ Excel recognizes dates in various formats, but it's generally recommended to use a consistent date format to avoid confusion.

- ➢ Dates entered in Excel are stored as serial numbers, where each date is represented by a unique number. Excel then formats these serial numbers to display them as dates.
- ➢ These methods should cover most scenarios for entering dates in Microsoft Excel. Let me know if you need further assistance.

9.5.3. Alphanumeric Values

Alphanumeric values are those that contain both letters and numbers. In Microsoft Excel, you can enter alphanumeric values into cells just like any other type of data. Here's how:

Manual Entry:

1. Click on the cell where you want to enter the alphanumeric value.
2. Type the alphanumeric value directly into the cell.
3. Press Enter to confirm.

Examples of Alphanumeric Values:

- "ABC123"
- "X7Y9Z2"
- "Hello123"
- "Data2024"

Special Characters:

In addition to alphanumeric values, you can include special characters such as punctuation marks or symbols. For example:

- "ABC123!"
- "X7Y9Z2#"
- "Hello@123"
- "Data_2024"

9.5.4. Formatting

By default, Excel interprets alphanumeric values as text. However, if you wish to make calculations or use them in formulas, you may need to convert them to numbers using the 'VALUE' function or another suitable conversion function.

Notes:

- In Excel, alphanumeric values are often used to represent codes, identifiers, passwords, and other data kinds.
- When sorting alphanumeric values, Excel normally follows a sorting order where numbers come before letters, and uppercase letters come before lowercase letters.

Saving and quitting

Saving and quitting a worksheet in Microsoft Excel is a simple process. Here's how to do it:

Saving a Worksheet:

1. Save: To save your worksheet, click on the "File" tab in the top-left corner of the Excel window.

2. Save As: If you're saving the worksheet for the first time or want to save it with a new name or in a different location, choose "Save As" instead of "Save".

3. Choose Location and Name: Select the location where you want to save the worksheet, enter a name for the file, and click "Save".

4. Shortcut: You can also use the keyboard shortcut Ctrl + S (Cmd + S on Mac) to quickly save the worksheet.

9.5.5. Quitting Excel:

1. Close Excel: To quit Excel, you can simply close the application window by clicking the "X" button in the top-right corner of the window.

2. Prompt to Save: If you have unsaved changes in your worksheet, Excel will prompt you to save them before quitting. Choose "Save" to save the changes, "Don't Save" to discard the changes, or "Cancel" to return to the worksheet without quitting Excel.

AutoSave (Optional):

- Excel offers an AutoSave feature that automatically saves your changes periodically
- You can enable or disable this feature from the "AutoSave" toggle button in the top-right corner of the Excel window.

Working with formulas and cell referencing

Working with formulas and cell referencing is a fundamental aspect of Microsoft Excel. Formulas allow you to perform calculations and manipulate data dynamically. Cell referencing enables you to refer to the content of other cells within your formulas. Here's how you can work with formulas and cell referencing in Excel:

Basic Formula Syntax:

- Formulas in Excel always start with an equal sign (=).
- You can perform mathematical operations like addition (+), subtraction (-), multiplication (*), and division (/) directly in your formulas.
- Excel also supports a wide range of built-in functions for performing more complex calculations.

Cell Referencing:

- Relative Reference: When you reference a cell in a formula without any dollar signs ($), Excel uses relative referencing. For example, if you enter a formula like `=A1+B1` into cell C1, Excel will add the values in cells A1 and B1.
- Absolute Reference: If you want a reference to a cell to remain constant even if you copy the formula to other cells, you can use absolute referencing by adding dollar signs ($). For example, `=$A$1+$B$1` will always refer to cells A1 and B1, regardless of where the formula is copied.

- Mixed Reference: You can also have mixed references where only the row or column is absolute. For example, `$A1` will keep the column constant but allow the row to change, while `A$1` will keep the row constant but allow the column to change.

Using Functions:

- Excel provides a vast array of built-in functions for performing various calculations and data manipulations.
- To use a function, start typing the function name followed by an opening parenthesis, and then enters the arguments inside the parentheses. For example, `=SUM(A1:A5)` calculates the sum of values in cells A1 through A5.
- You can also use the function wizard to help you choose and enter functions. It's accessible from the "Formulas" tab in the ribbon.

Examples:

1. Simple Calculation:

- `=A1+B1`: Adds the values in cells A1 and B1.

2. Sum Function:

- `=SUM (A1:A10)`: Calculates the sum of values in cells A1 through A10.

3. Average Function:

- `=AVERAGE (A1:A10)`: Calculates the average of values in cells A1 through A10.

4. IF Function:

- `=IF (A1>10, "Yes", "No")`: Checks if the value in cell A1 is greater than 10. If true, returns "Yes"; otherwise, returns "No".

9.5.6. Shortcuts keys

The comprehensive list of keyboard shortcuts in Microsoft Excel for Windows is:

Navigation and Selection:

- Arrow Keys: Move one cell in the direction of the arrow.
- Ctrl + Arrow Key: Move to the edge of the current data region in the worksheet.
- Ctrl + Home: Move to the beginning of the worksheet.
- Ctrl + End: Move to the last cell with data on the worksheet.
- Ctrl + Shift + Arrow Key: Select a range of cells.
- Shift + Arrow Keys: Extend the selection of cells by one cell.
- Ctrl + Space: Select the entire column.
- Shift + Space: Select the entire row.
- Ctrl + A: Select all cells in the worksheet.

Editing:

- F2: Edit the active cell.
- Ctrl + X: Cut selected cells.
- Ctrl + C: Copy selected cells.
- Ctrl + V: Paste copied or cut cells.
- Ctrl + Z: Undo your last action.
- Ctrl + Y: Redo your last undone action.
- Alt + Enter: Start a new line in the same cell.
- Ctrl + D: Fill down from the cell above.
- Ctrl + R: Fill right from the cell to the left.

Formatting:

- Ctrl + B: Bold selected text.
- Ctrl + I: Italicize selected text.
- Ctrl + U: Underline selected text.
- Ctrl + 1: Format cells dialog box (for Number tab)
- Ctrl + Shift + $: Apply the currency format.
- Ctrl + Shift + %: Apply the percentage format.

- Ctrl + Shift + : Apply the number format with two decimal places, thousands separator, and minus sign for negative values.

Working with Formulas:

(Equals Sign): Start a formula.

- Tab: Complete a function or formula entry.
- Ctrl + A: Insert argument names into a formula after typing a function name in a formula.
- F4: Repeat the last action (such as formatting or inserting a row/column) or cycle through absolute/relative reference options when editing a formula.
- Ctrl + Shift + Enter: Enter a formula as an array formula.
- Alt + =: AutoSum selected cells.
- F9: Calculate all worksheets in all open workbooks.

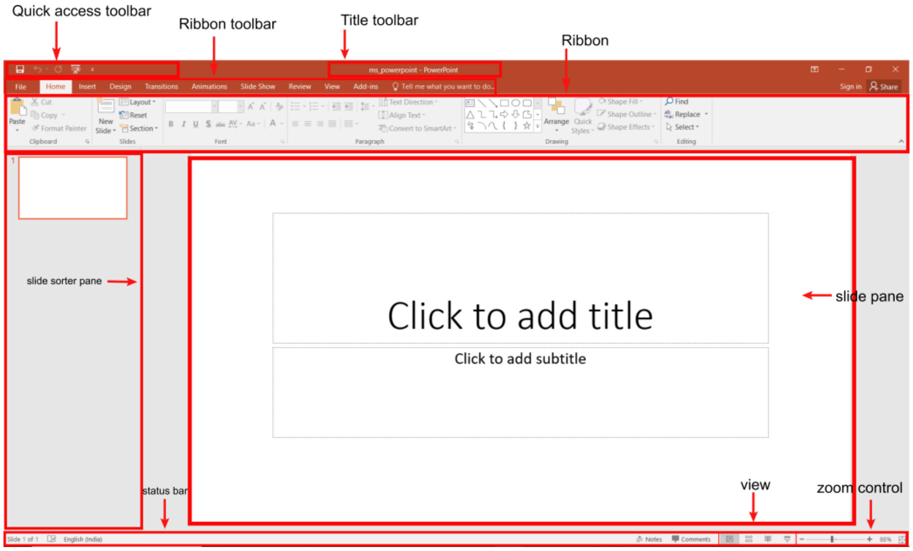
Other Useful Shortcuts:

- Ctrl + S: Save the workbook.
- Ctrl + P: Print the workbook.
- Ctrl + F: Open the Find dialog box.
- Ctrl + H: Open the Replace dialog box.
- Ctrl + Shift + L: Turn on/off filter (toggle AutoFilter).
- Ctrl + K: Insert a hyperlink.
- F1: Open Excel Help.

This list covers some of the most commonly used keyboard shortcuts in Excel for Windows. Feel free to explore and practice these shortcuts to improve your productivity!

9.6. Power Point

PowerPoint is popular presentation software made by Microsoft. It is typically used to create slideshows for business presentations, educational reasons, and other professional or personal projects. It includes a profusion of capabilities including as text formatting, slide transitions, animations, multimedia insertion (photos, videos, and music), and chart, graph, and diagram tools. Users can select from a number of pre-designed themes or construct unique layouts to meet their specific needs. The interface is commonly made up of a slide pane where individual slides are created and updated, a thumbnail view for navigating through the presentation, and several toolbars and menus for accessing different functions. PowerPoint integrates seamlessly with other Microsoft Office products like Word and Excel, allowing you to include data and content from these apps into your presentations.



It encourages collaboration by including tools such as comments, track changes, and real-time co-authoring, which allow several users to work on the same presentation at the same time. Furthermore, PowerPoint presentations can be exported in a variety of formats, including PDF, video, and HTML, for easy sharing and dissemination. Despite its extensive use and popularity, PowerPoint has been criticized for fostering a "slide-centric" culture in which complex ideas are oversimplified and communication becomes unduly reliant on visual aids. However, with correct design and delivery skills, PowerPoint remains a powerful tool for effectively communicating information and engaging people in a variety of circumstances. Tabs in PowerPoint are essentially top-level organizational units that group similar commands and functions. Each tab in PowerPoint has a specific purpose and gives you access to a variety of tools and features to help you develop, design, and deliver effective presentations. Here's an overview of the primary tabs commonly found in PowerPoint.

1. **Home Tab:** The Home Tab is the default tab with regularly used formatting commands for text, presentations, and objects. It provides options for modifying fonts, alignment, bullet points, and basic shapes.

2. **Insert Tab**: The Insert tab allows you to add many forms of information to your presentations, including photographs, shapes, charts, text boxes, videos, and audio recordings. It also has capabilities for putting headers, footers, and other content.

3. **Design Tab:** Add pre-designed themes and templates to your presentation for a consistent and professional look. You can also change the color scheme, fonts, and effects of your slides using this option.

4. **Transitions Tab:** This tab is for adding transition effects between presentations. Transitions determine how one slide flows into the next throughout a presentation, bringing visual interest and fluidity to your delivery.

5. **Animations Tab:** The Animations option allows you to animate text, images, and objects on your slides. It provides a wide range of animation effects as well as features for controlling animation timing and sequence

6. **Slide Show Tab:** The Slide Show menu allows you to start, rehearse, and customize your slideshow presentation. This page allows you to modify parameters including slide timing, narration, and navigation choices.

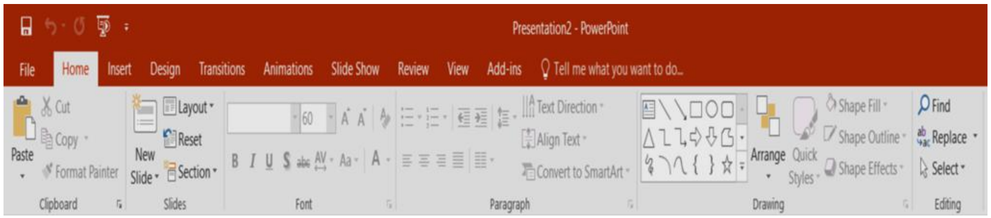
7. **Review Tab:** This tab is used for collaboration and presentation review. It provides facilities for spell-checking, commenting, tracking changes, comparing presentations, and password-protected content.

8. **View Tab:** Customize your presentation with various views and layouts. You can toggle between Normal, Slide Sorter, Reading, and Outline views, as well as modify the magnification level and use presentation tools such as gridlines and rulers..

9. **Developer Tab:** The Developer tab is not visible by default and must be enabled manually through PowerPoint settings. It offers advanced features for generating and maintaining macros, form controls, and add-ins, making it more useful to developers and power users.

9.6.1. The Home tab

The Home tab in PowerPoint serves as a center for a variety of regularly used commands and formatting choices. It contains tools for manipulating text, shapes, photos, and the actual slides. Below is a full overview of the commands accessible on the Home tab:



1. **Clipboard:** This section includes fundamental clipboard functions including cut, copy, and paste. It also provides choices for styling copied information with the Format Painter tool, which allows you to apply the same formatting to numerous objects or text.

2. **Slides:** Add, duplicate, or delete slides in your presentation. You may also change the order of the slides using the Arrange dropdown option.

3. **Font:** This area provides text formatting settings such as font type, size, color, and highlighting. It also has buttons for bold, italic, and underline formatting, as well as text alignment choices (left, center, right, and justify).

4. **Paragraph**: Adjust the spacing and alignment of text paragraphs. You can change the line spacing, paragraph spacing, and indentation settings, as well as the alignment of text within a text box or placeholder..

5. **Drawing**: This area offers tools to create and format forms, lines, and other drawing objects. You can select from a variety of shapes, add lines or arrows, and change the fill and outline colors, among other formatting options.

6. **Editing:** Access commands for altering objects and text in your presentation. This includes the ability to pick, find, and replace text, as well as access sophisticated editing tools such as grouping and ungrouping objects.

7. **Find:** Use this command to look for certain text or objects in your presentation. You can search for text or shapes on your slides using keywords or phrases.

8. **Select**: This dropdown menu allows you to choose individual items or features from your presentation. You can select all objects on a slide, objects by type (such as shapes or photos), or use the selection window to perform more advanced selections.

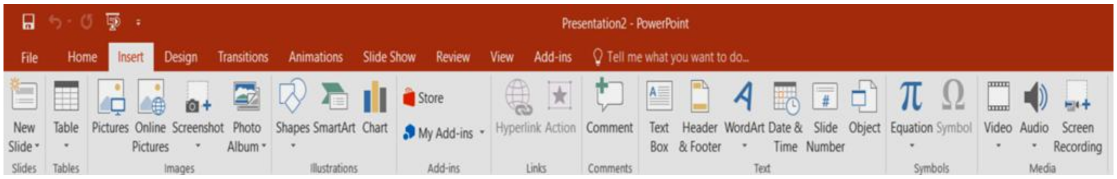
9. **Editing Group**: The Editing Group offers additional editing tools, such as undo and redo commands, as well as cutting, copying, and pasting possibilities.

Overall, PowerPoint's Home tab acts as a comprehensive toolset for basic slide creation, text formatting, and object manipulation, giving users the tools they need to develop and update presentations efficiently.

9.6.2.The Insert tab

The Insert tab in PowerPoint contains all of the tools and options for adding different types of information to your presentation slides. Below is a full explanation of the commands accessible on the Insert tab:

1. **Slides:** This group allows you to add new slides to your presentation. You can choose between numerous layouts or integrate a specific slide layout into your presentation.



2. **Tables:** This group has options for inserting tables into your slides. You can select from pre-designed table layouts or create your own table by choosing the number of rows and columns.

3. **Illustrations:** You can add photographs, online images, forms, smart art, and icons. This group lets you add graphic components to your slides to improve your presentation.

4. **Charts**: Use charts and graphs to visually show data in your slides. You can select from a variety of chart kinds, including column, bar, line, and pie, and then personalize the chart with your data.

5. **Add-ins**: This group offers choices for including add-ins in your PowerPoint presentation. Add-ins is features or tools that can extend PowerPoint's capability, such as third-party programs or service integrations.

6. **Links:** Add interactive aspects to your presentation by including hyperlinks, action buttons, or settings. Hyperlinks allow you to direct people to external websites or other slides in your presentation, whilst action buttons allow them to browse between slides or perform specified tasks.

7. **Text:** This group allows you to put text boxes, headers, footers, and word art into your slides. You can place text directly on your slides or use text boxes for more exact positioning and formatting.

8. **Media**: Add multimedia elements like audio and video files to your presentation. You can easily embed audio or video files into your slides, as well as link to external files on your computer or network.

9. **Screenshots:** This group allows you to include screenshots straight into your presentation slides. You can take screenshots of specific windows or regions on your screen and use them as images in your slides.

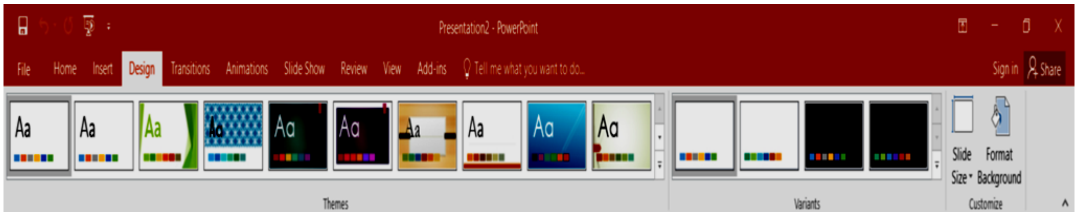
10. **Comments:** This command allows you to add comments to your presentation slides for collaboration and feedback purposes. You can leave comments on individual sections of your slides and respond to comments from other participants.

Overall, PowerPoint's Insert tab provides a variety of tools and options for adding material, drawings, media, and interactive features to your presentation slides, enabling you to build dynamic and interesting presentations.

9.6.3. The Design tab

The Design tab in PowerPoint contains tools and options for modifying the overall appearance and layout of your presentation. Below is a full explanation of the commands accessible on the Design tab:

1. **Themes:** This collection includes pre-designed themes for your presentation. Themes are a collection of colors, fonts, and effects that give your presentation a uniform and professional appearance. You can use the built-in themes or browse and apply custom themes.



2. **Variant:** The Variant dropdown menu in the Themes group allows you to customize the selected theme's color schemes and font combinations. This allows for additional customization and customization of your presentation's design.

3. **Customize:** The Customize group offers choices to customize the existing theme. You can change the theme colors, fonts, and effects to better reflect your preferences or your company's branding guidelines. You can also design and save your own custom themes.

4. **Slide Size:** This group lets you adjust the size and direction of your slides. You can pick between conventional slide sizes such as 4:3 (normal) and 16:9 (widescreen), as well as a custom slide size. You can also choose between landscape and portrait orientations.

5. **Slide Orientation**: Adapt your slides' backgrounds. You can choose a solid color backdrop, a gradient fill, a texture fill, or a photo or pattern for the slide background. You can also reset the background to its default settings.

6. **Slide Background:** This function suggests layout possibilities for your slides based on their content. PowerPoint analyzes your slide content and suggests design changes to improve the visual appeal and readability of your presentation.

7. **Design Ideas:** This function suggests layout possibilities based on your slides' content. PowerPoint analyzes your slide content and suggests design ideas to improve the visual appeal and readability of your presentation.

8. **Format Background**: Open the Format backdrop box to modify your slide backdrop. You can change the backdrop's transparency, blur, and other effects, as well as apply background styles and fills.

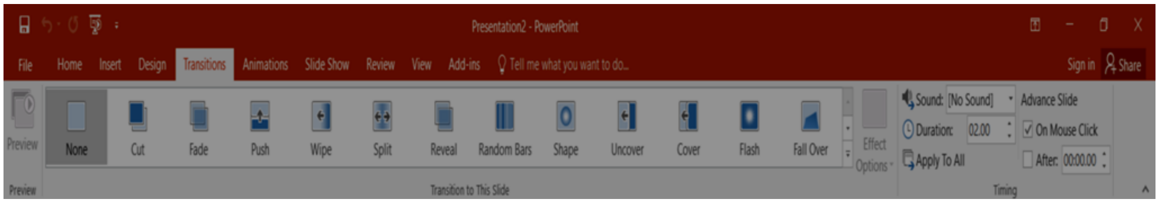
9. **Hide Background Graphics:** Use this command to toggle between showing and hiding background visuals in your presentation. Background graphics may comprise photos, textures, or patterns that are applied to the slide background.

Overall, the Design tab in PowerPoint provides a wide range of tools and options for applying themes, customizing slide layouts, and improving the visual aspect of your presentation slides, allowing you to produce polished and professional-looking presentations.

9.6.4. The Transition tab in

The Transition tab in PowerPoint allows you to build transition effects between slides in your presentation. These effects regulate how one slide transitions to the next during a slideshow, enhancing the visual appeal and smoothness of your presentation delivery. Below is a thorough explanation of the commands accessible on the Transition tab:

1. **Transition to This Slide**: Use this section to preview and apply transition effects to the currently chosen slide. By clicking on the desired transition effect, you can select one of several options, including fade, dissolve, zoom, or slide.



Sound: [No Sound]
Duration: 02.00
☑ On Mouse Click
□ After: 00:00.00

2. **Effect Options:** After picking a transition effect, this dropdown menu offers more customization possibilities. Depending on the transition, you may be able to change factors like direction, speed, or timing.

3. **Duration:** Select the desired transition effect duration from the dropdown menu. You can define how long it takes for the transition to complete, from a fraction of a second to many seconds.

4. **Apply To All:** This command applies the chosen transition effect to all slides in your presentation. It's an easy technique to ensure consistency across your slideshows and keep a consistent visual style throughout the presentation.

5. **Preview:** The evaluate group includes buttons to evaluate transition effects on your presentations. You can sample the transition for the current slide or see the entire slideshow to see how the transitions flow between slides.

6. **Advance Slide**: This part allows you to regulate how slides advance throughout a presentation. You can advance slides manually (by clicking the mouse or typing a key) or automatically (after a set time).

7. **On Mouse Click**: Set this option to progress slides when the mouse is clicked. It's excellent for presentations where you want the presenter to manage the slideshow's pace.

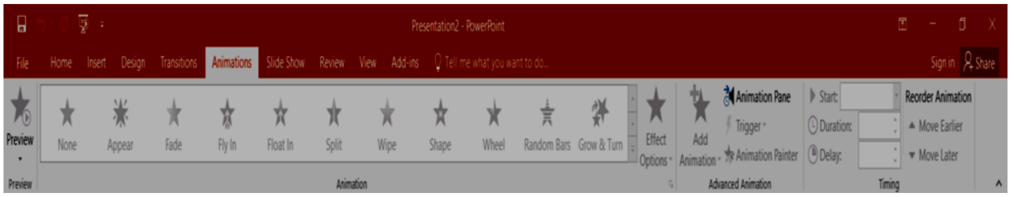
8. **After:** Set this option to automatically advance slides after a specified time. You can set the time (in seconds) after which the slide shall move to the next slide automatically.

9. **Transition Sound:** Use this dropdown menu to add sound effects to your transition. You can use the built-in sound effects or select "No Sound" if you prefer silent transitions.

Overall, the Transition tab in PowerPoint offers a variety of tools and options for adding transition effects between slides, adjusting the timing of slide transitions, and modifying your slideshow's behavior, all of which improve your audience's overall presentation experience.

9.6.5. Animation tab

The Animation tab in PowerPoint includes tools and settings for animating text, graphics, and other objects on your slides. Animations enhance your presentation by adding movement and visual appeal to your material.



1. **Add Animation**: In the Add Animation section, you may select animation effects for your slide's objects. Animate text, images, forms, and other objects using effects like entry, emphasis, exit, and motion pathways.

2. **Animation Gallery:** The "Add Animation" option opens a dropdown menu with animation effects organized by kind (Entrance, Emphasis, Exit, and Motion Paths). You can test many animation effects and choose the one that best fits your material.

3. **Animation Pane**: Open the Animation Pane to view animations on the current slide in detail. The Animation Pane lets you customize the sequence, timing, and attributes of animations.

4. **Timing:** Set the timing and length of animations. You can define when animations begin (on click, with previous, or after previous) and how long they last (in seconds).

5. **Delay:** Set a delay before the animation starts. You can define the amount of time (in seconds) before the animation starts (for example, after the previous animation or after a mouse click).

6. **Duration:** Set the duration of an animation effect. You can define the duration of the animation (in seconds), as well as the pace and timing of the animation movement.

7. **Delay between Animations:** Set a delay between numerous animations on the same item. You can define how long (in seconds) to wait before launching the next animation effect.

8. **Effect Options:** After applying an animation effect, this dropdown menu allows for more customization. Depending on the animation effect, you may be able to change characteristics like direction, speed, or style.

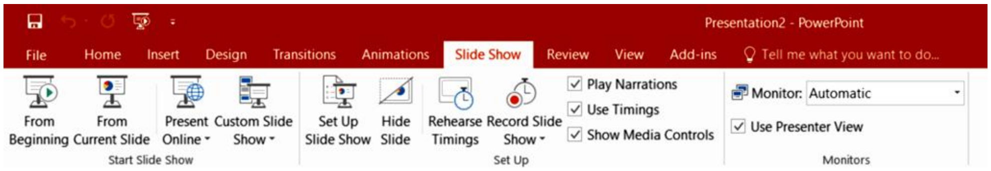
9. Advanced Animation**: This section provides advanced options for customizing animation** effects, including options for adding sound effects, setting repeat intervals, and configuring triggers for animations.

Overall, PowerPoint's Animation tab provides a variety of tools and options for adding animation effects to your presentation slides, allowing you to create dynamic and interesting presentations that hold your audience's attention.

9.6.6. Slide Show Tab

In PowerPoint, the "Slide Show" tab contains all of the tools and settings for presenting your slides.

1. **Start Slide Show**: You can start your presentation from the current slide or from the beginning.



☑ Play Narrations
☑ Use Timings
☑ Show Media Controls
Monitor: Automatic
☑ Use Presenter View

2. **Set Up Slide Show**: This section allows you to start the presentation from the current slide or from the beginning, loop the slides, and display timings or narrations if applicable.

3. **Monitors:** Select the display to use for your presentation, especially if you have many displays.

4. **Show/Hide Slide Show Controls:** You can toggle the visibility of on-screen navigation controls during the presentation.

5. **Use Presenter View**: Presenter View is useful for managing numerous displays. It displays your speaker notes and forthcoming slides, while your audience sees

6. **Rehearse Timings**: Use this tool to time how long you spend on each slide during your presentation practice.

7. **Record Slide Show:** Record your presentation, including voice narration and pointer motions

8. **Presentation Views:** Switch between Normal, Slide Sorter, Reading View, and Slide Master.

9. **Zoom In/Out**: Control the zoom level of your slides during the presentation.

10. **End Slide Show**: To conclude the presentation, press the End Slide Show button.

The "Slide Show" category contains all you need to know about delivering and rehearsing PowerPoint presentations.

9.6.7. The "Review" tab

The "Review" tab in PowerPoint includes tools and capabilities for proofreading, collaborating, and managing comments and modifications to your presentation. The "Review" tab often includes the following features:

1. **Spelling:** This function checks your presentation for misspelled terms and provides corrections.

2. **Research**: Search for material online without leaving PowerPoint. You can use tools like online dictionaries, thesauruses, and translation services.

3. **Thesaurus**: Identify synonyms and antonyms for specific terms.

4. **Language**: Select a language for presentation proofreading and translation.

5. **Accessibility:** Provides tools for checking the accessibility of your presentation, ensuring it can be easily understood and used by people with disabilities.

6. **Comments:** Allows you to add, edit, delete, and manage comments in your presentation. Comments are useful for collaboration and feedback.

7. **Compare:** Enables you to compare different versions of a presentation to see changes made between them.

8. **Protect:** Allows you to protect your presentation by adding passwords or restricting editing permissions.

9. **Inspect:** Helps you remove hidden metadata and personal information from your presentation before sharing it.

10. **Translate:** Lets you translate selected text or the entire presentation into another language.

The "Review" tab is essential for ensuring the quality, accuracy, and accessibility of your PowerPoint presentations, as well as facilitating collaboration and feedback among team members.

9.6.8. The Power Point

In PowerPoint, the "View" tab provides various options for customizing how you view and work with your presentation. Here are some common features you'll find in the "View" tab:

1. **Presentation Views:** This section allows you to switch between different views of your presentation, such as Normal, Slide Sorter, Reading View, and Slide Master. Each view offers a different perspective on your slides and allows you to perform different tasks.

2. **Show/Hide:** These options let you show or hide specific elements in the PowerPoint interface, such as rulers, gridlines, guides, and the navigation pane.

3. **Zoom:** You can use this feature to adjust the zoom level of your presentation, allowing you to focus on specific details or view the entire slide.

4. **Color/Grayscale:** This option allows you to switch between color and grayscale modes, useful for checking the appearance of your slides in different contexts.

5. **Arrange All**: If you have multiple presentations or windows open, this feature arranges them all on the screen for easier navigation and comparison.

6. **Window:** Here, you can manage multiple presentation windows, including options to switch between open presentations, arrange them, or switch to different windows.

7. **Macros:** If you've created or imported macros into your PowerPoint presentation, you can access them and run them from this menu.

8. **Presentation Views Group:** This section provides shortcuts to commonly used presentation views, such as Normal, Slide Sorter, Reading View, and Slide Master.

9. **Zoom Group:** You can quickly adjust the zoom level of your presentation or use the Zoom Slider to zoom in or out.

10. **Color/Grayscale Group:** This section provides options for viewing your presentation in color or grayscale mode.

The "View" tab in PowerPoint offers a range of tools and options to customize your working environment and optimize your workflow while creating and editing presentations.

9.6.9.The Developer" tab

In PowerPoint, the "Developer" tab is not displayed by default and needs to be enabled manually. This tab provides advanced tools and features for creating and working with macros, add-ins, form controls, and ActiveX controls. Here are some common features you'll find in the "Developer" tab:

1. **Code:** This section provides access to the Visual Basic for Applications (VBA) editor, where you can write, edit, and manage macros to automate tasks or enhance functionality in your presentations.

2. **Add-Ins**: Allows you to manage add-ins that extends the functionality of PowerPoint. You can load, unload, or create add-ins from here.

3. **Controls:** This section provides various form controls and ActiveX controls that you can insert into your presentation to create interactive elements such as buttons, checkboxes, combo boxes, and more.

4. **XML Mapping**: Allows you to map XML data to specific elements in your presentation, enabling dynamic updating of content from external data sources.

5. **Macros:** Provides access to a list of macros that you've created or recorded, allowing you to run them or assign them to buttons or other form controls.

6. **Record Macro:** Allows you to record a series of actions as a macro, which can then be replayed to automate repetitive tasks.

7. **Visual Basic:** Opens the Visual Basic Editor, where you can write, edit, and debug VBA code for macros.

8. **Insert:** This section provides options for inserting various types of form controls, ActiveX controls, and other objects into your presentation.

The "Developer" tab is mostly utilized by advanced users, such as developers and power users, who want to customize and automate PowerPoint presentations with macros, add-ins, and other advanced capabilities. If you don't see the "Developer" tab in your PowerPoint application, you can enable it in the PowerPoint settings or preferences menu.

9.7. Summary

Microsoft Office is a suite of productivity software produced by Microsoft, designed to help users create, edit, and manage numerous sorts of documents, presentations, spreadsheets, and more. A word processing tool used for generating documents such as letters, reports, essays, and resumes. It offers features like spell check, grammar check, formatting settings, and collaboration capabilities. A spreadsheet program for organizing, analyzing, and displaying data. Excel has sophisticated features for constructing formulas, charts, graphs, and pivot tables, making it excellent for budgeting, financial analysis, and information management. Presentation software used to create slideshows for meetings, lectures, and other events. PowerPoint provides a variety of slide layouts, animations, and transitions to assist users build compelling presentations. An email client and personal information manager used for organizing emails, calendars, contacts, and tasks. Outlook interfaces with other Office products and includes capabilities including email filtering, scheduling, and collaboration tools. A database management system used for developing and managing databases. Access allows users to develop bespoke database applications for tracking information, generating reports, and automating business operations.

9.8. Terminal questions

Q. 1. What is Microsoft office? Discuss the features of MS office.

Answer:---

Q. 2. What are Microsoft words? Discuss the features of Microsoft words.

Answer:---

Q. 3. How Microsoft Excel is very useful in data management and calculation?

Answer:---

Q. 4. Discuss the Creating worksheet in MS Excel.

Answer:---

Q. 5. What are the short cut keys? Discuss the shortcut key of MS Excel.

Answer:---

Q. 6. What is the power point? Discuss about power point presentation software.

Answer:---

9.9. Further suggested readings

1. Rajaraman, V. and Radhakrishnan T.; Digital Logic and Computer Organisatin, 1st ed., Prentic-Hall of India, New Delhi, (2006).

2. Hennessy, J.L. and Patterson, D. A.; Computer Oranization and Design: The Hardware/Software Interfaces, Morgan Kauffman Publishers, San Mateo, CA, (1994).

3. Chauhan Sunil, Saxena Akash and Gupta Kratika; Fundamentals of Computer, Laxmi Publication, (2006).

4. Andrie de Vries and Joris Meys, R Programming for Dummies, 2edition, Wiley publication.

5. John Paul Mueller and Luca Massaron, Machine Learning (in Python and R) for Dummies, Wiley publication

Block-4 PGEVS-106N

Rajarshi Tandon Open

Universitv. Pravaarai

Numerical
and
Statistical Computing

Block- 4

Computational analyses

Unit-10

Algorithm and Flow Charts

Unit-11

Computation with MS Excel

Unit-12

Statistical Software

Rajarshi Tandon Open

University, Prayagraj

Numerical and Statistical Computing

Course Design Committee

Prof. Ashutosh Gupta **Chairman**
School of Science, UPRTOU, Prayagraj

Dr. Uma Rani Agarwal **Member**
Rtd. Professor, Department of Botany
CMP Degree College, Prayagraj

Dr. Ayodhaya Prasad Verma **Member**
Red. Professor, Department of Botany
B.S.N.V. P.G. College, Lucknow

Dr. Sudhir Kumar Singh **Member**
Assistant Professor
K. Banerjee Centre for Atmospheric and Ocean Studies
University of Allahabad, Prayagraj

Dr. Ravindra Pratap Singh **Member**
Assistant Professor (Biochemistry)
School of Science, UPRTOU, Prayagraj

Dr. Dharmveer Singh **Course Coordinator**
Assistant Professor (Biochemistry)
School of Science, UPRTOU, Prayagraj

Course Preparation Committee

Dr. Anuj Kumar Singh **Author** **Block-1** (Unit: 1)
Assistant Prof. (Statistics)
School of Sciences, UPRTOU, Prayagraj

Dr. Upasana Singh **Author** **Block-1&2** (Unit: 2-6)
Assistant Professor-Zoology
Prof. Rajendra Singh Rajju Bhaiya
University, Prayagraj

Dr. Jaspal Singh **Author** **Block-1&3,4** (Unit: 1,7,8,9,10,11)
Assistant Professor
Department of Environmental Science,
Bareilly College, Bareilly

Dr. Nishtha Seth **Author** **(All blocks and units)**
Associate Professor
Department of Environmental Science,
Bareilly College, Bareilly

Dr. Dharmveer Singh
(Course and SLM Coordinator)
School of Sciences, UPRTOU, Prayagraj

Introduction

This fourth block of numerical and statistical computing, this consists of following three units:

Unit-10: The introduction, goals, and an example of an algorithm with its properties are covered in this unit. This unit covers the many studies of algorithms, flow charts, and their properties.

Unit-11: This unit explains how to compute using Microsoft Excel. Functions in particular are covered, including logical, loop, statistical, numeric, and mathematical functions; statistical analysis utilizing Excel's descriptive statistics; curve fitting; correlation and regression analysis; and graphing.

Unit-12: The basic overview of statistical software (SPSS, R, etc.) is covered in this section. We go over the fundamentals of R, R studio, data file creation, and statistical analysis using R.

Unit-10: Algorithm and Flow Charts

10.1. Introduction

A computer is a device that performs predefined actions in response to a given set of instructions. Sadly, computers follow our instructions, which is not necessarily what we would like them to do. The instructions we provide to software must be clear and devoid of any room for misunderstanding. The computers will always make some effort to guarantee that the outcomes meet our requirements.

The most fascinating tasks appear to be complex from a programming perspective. Some problems must be difficult by definition. In many circumstances, however, it can be due to another aspect over which we have influence, such as an insufficient or imprecise problem specification. Complexity does not have to be a problem while developing computer programs provided it is handled and controlled properly. Computer programming can be challenging. It is challenging in large part because it is a complex task that involves multiple mental processes at once. We can do a lot to make things easier. For instance, the task of programming can be made much more manageable by systematically breaking it up into a number of less complex subtasks.

This may be referred to as the divide and conquer approach. The approach of subdividing a task has made considerable success in practice. Given a task, we separate it into two important phases, namely the problem-solving phase and the implementation phase. In the problem-solving phase, we concentrate on subdividing a task into a number of comparatively simpler subtasks. This is equivalent to say that we concentrate on designing an algorithm to solve the stated problem (see figure 1.1 below). Only after we are satisfied that we have developed a suitable algorithm do we proceed to the intricacies of its implementation in some programming language.

An algorithm is a step-by-step method or collection of rules for solving a problem or completing a task. It provides as a blueprint for how a computer program should carry out a certain activity, offering a methodical approach to problem resolution. Algorithms can be simple or complex, depending on the problem they are designed to address. Flowcharts are graphical representations of algorithms or processes that use symbols and arrows to show the flow of control within a program or system. They depict the steps of an algorithm, making it easier to understand and examine the program's logic. Flowcharts usually incorporate start and finish points, decision points, input/output activities, and various control structures including loops and conditionals.

Objectives

After the study of this unit, you will be able to:

- Define a algorithm
- Set various steps in a program using algorithm
- Define a flowchart
- Draw flow chart of various problem

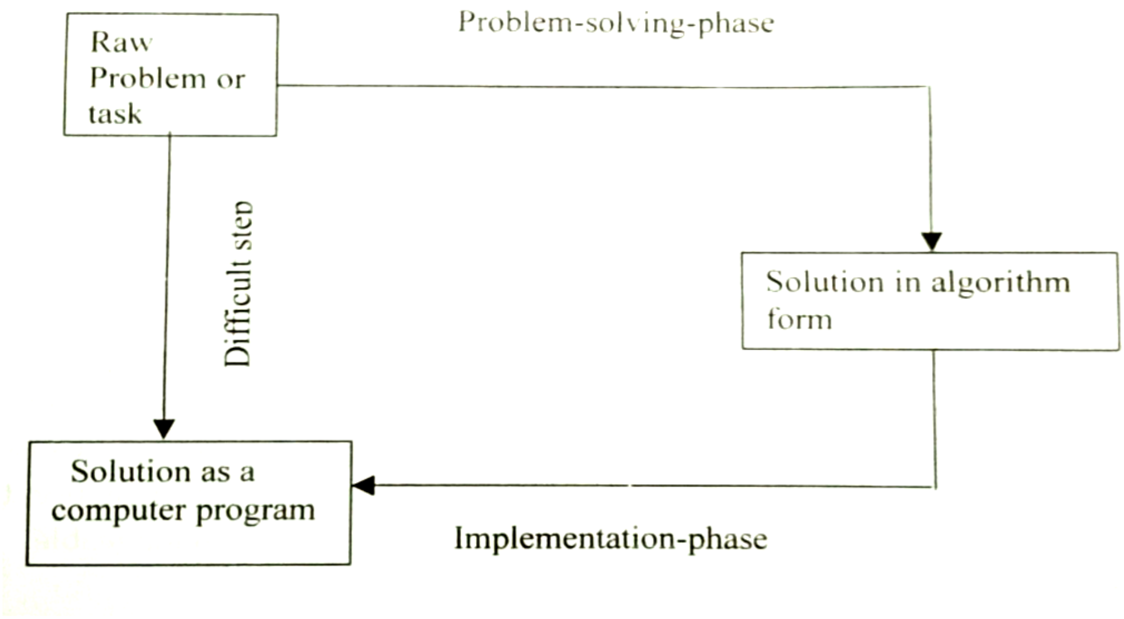


Fig. 10.1: Problem solving and Implementation phases

10.2. Algorithm

Algorithm refers to "A set of finite rules or instructions to be followed in calculations or other problem-solving operations" or "A procedure for solving a mathematical problem in a finite number of steps that frequently involves recursive operations" .

An algorithm is a clear and sequential set of actions that solve a specific problem. Any algorithm will have a start and end point. Algorithms can be expressed in any language, including natural languages such as English or French and programming languages. The term "algorithm" may be new, but the notion should be familiar. Directions to a specific street represent an algorithm for locating the street. A recipe is a relatively common type of algorithm. A blueprint serves the same purpose in a construction project. At Christmas, many parents spend hours following algorithms to assemble their children's new toys. We will require that our algorithm have a number of crucial qualities. First, the steps in an algorithm must be basic and straightforward, and they must be executed in a specific order. Second, we will always insist on our algorithms being effective, which means they must always solve the problem in a finite number of steps. We couldn't afford to pay the computing costs if this wasn't the case.

102.1. Characteristics of algorithms

An algorithm is a well-defined series of steps or commands intended to solve a specific problem or complete a certain activity. It is a scientific problem-solving strategy that defines the key movements or processes to be performed in a clear and unambiguous manner. Algorithms can be expressed in different ways, including natural language, pseudo-code, flowcharts, and programming languages

Algorithms are essential in computer science for solving computational problems. They serve as the foundation for software development, allowing for the efficient execution of tasks and the manipulation of data. Algorithms can be advanced to achieve a variety of goals, including searching for data, sorting facts, calculating mathematical capabilities, and addressing complex optimization problems. A proper algorithm must have multiple acceptable features, including correctness, performance, clarity, and generality. Correctness ensures that the method yields the expected result for all potential inputs. Efficiency refers to the set of rules' ability to solve problems in a timely manner, preferably by making the best use of available resources. Clarity refers to how clear and understandable the set of rules is, making it easier for builders to implement and maintain. Generality implies that a set of principles can be applied to specific examples of a problem or modified to various contexts.

1. **Well-described steps:** Algorithms are made up of a defined and unambiguous set of instructions or steps that may be followed to complete a given task or solve a problem. Each phase should be clearly described, with no possibility for ambiguity or uncertainty.
2. **Input and output:** Algorithms process inputs, such as preliminary records or facts, and generate results or solutions based on a set of rules. The algorithm's good judgment determines the relationship between inputs and outputs.
3. **Finiteness:** Algorithms should have a clear termination condition. This strategy implies that they eventually reach an endpoint or change after a finite number of steps. If a collection of rules continues to run indefinitely without being terminated, it is considered incorrect or incomplete.
4. **Determinism:** Algorithms are deterministic, meaning they consistently return the same results with the same inputs and conditions. The behavior of a set of rules

should be predictable and consistent. Algorithms aim to be efficient in terms of time and resources. They aim to clear up issues or accomplish tasks in an inexpensive amount of time and with the best use of computational resources like memory, processing power, or garage.

5. **Correctness:** Algorithms must generate accurate outputs for all valid inputs within their area. They must accurately solve the problem for which they are built, and their outputs must correspond to the expected outcomes.
6. **Modularity and reusability:** Algorithms can be broken down into smaller subproblems or features that can be utilized in different parts of the algorithm or other algorithms. This encourages code agency, maintainability, and reuse. Clear and easy-to-understand algorithms are essential for successful implementation. Well-documented and legible code can help an algorithm become more understandable.

10.2.2. Examples of algorithm

Examples of algorithm are omnipresent even in our everyday life. An interesting example is given below:

Example 1. Sorting mail:

A detailed algorithm for sorting mail is as follows.

```
Step 1: Get all mail from mailbox
Step 2: Put mails on table
        While more mails to sort
Step 3: Get piece of mail from the table
Step 4: If piece is personal
        Read it
Step 5: If piece is magazine
        Put in the magazine rack
Step 6: Else if piece is bill
        Pay it
```

Step 7: Stop

When an algorithm is used to process information, data is usually read from an input source or device, written to an output sink or device, and/or stored for later processing. Stored data is considered part of the internal state of the entity conducting the algorithm. In practice, the state is maintained in a data structure, but an algorithm only needs the internal data for certain operation sets. Any computational process requires a carefully defined and stated method that applies to all potential scenarios. That is, all conditional actions must be handled systematically on a case-by-case basis, with clear and computable criteria for each.

Because an algorithm is a precise set of steps, the order of computation is usually always important to its proper operation. Instructions are typically supposed to be openly listed and characterized as beginning 'from the top' and progressing 'down to the bottom', a concept more technically represented by flow of control.

Algorithms can be expressed in a variety of notations, including natural language, pseudo-code, flow charts, and programming languages. Algorithms' natural language statements are verbose and unclear, and therefore are rarely employed to solve difficult or technical problems. Pseudocode and flowcharts are structured approaches to express algorithms that avoid many of the ambiguities found in natural language assertions, while staying independent of the implementation language. Programming languages are primarily designed to express algorithms in a computer-executable format, but they are also frequently used to define or document algorithms. Supplementing small flow charts with common language and/or arithmetic expressions put inside block diagrams to summarize what the flow charts are doing might be useful in the presentation of an algorithm. We will present a few standard examples below to help clarify the construction of an algorithm. The offered algorithms are simply for example purposes; alternate designs are always possible.

Example 2: Algorithm for finding out the frequency of a definite integer in a sequence of integers:

Step 1: Let the total number of integers =0

Step 2: Let the frequency of desired integers =0

Step 3: Repeat steps 4, 5, 6 and 7 until the end of the sequence is reached.

Step 4: Read one integer from the sequence

Step 5: Add 1 to the total number of integers

Step 6: If the integer read is the desired integer, add 1 to the frequency of desired integer.

Step 7: Move to the next integer in the sequence. If no more integer is left in the sequence to go step 8 otherwise go back to step 4

Step 8: Write the frequency of desired integers.

Step 9: Stop.

Example 3: Algorithm to pick the largest of three numbers:

Step 1: Input the numbers X, Y and Z

Step 2: If X >Y, go to step 3

Otherwise go to step 5

Step 3: If X > Z, Write X as the largest number

Otherwise Write Z as the largest number

Step 4: Stop

Step 5: If Y > Z Write Y as the largest number

Otherwise Write Z as the largest number

Step 6: Stop

Example 4: Algorithm to find the largest number in an unsorted list of numbers:

The solution necessarily requires looking at every number in the list, but only once at each. From this follows a simple algorithm, which can be stated in a high-level description, say English prose, as:

Step 1: Assume that the first item is largest.

Step 2: Look at each of the remaining items in the list and if it is largest than the largest item so far, make a note of it.

Step 3: The last noted items is the largest in the list when the process is complete.

Step 4: Stop the process.

The following is a more formal coding of the algorithm in pseudocode, written in prose but very similar to the high-level language of a computer program.

```
Step 1: Input the non-empty list of number L.
Step 2: Largest @ $L_0$
Step 3: for each item in the list L $\geq 1$, do
        If the item > largest, then
        Largest @ the item
Step 4: Write largest
Step 5: Stop.
```

Example 5: Algorithm to count the number of non-zero observations in a list of n observations where n is any positive integer.

The strategy is to read a specific observation from the list. Check if it is not zero, then increment a counter. The same method is followed until all n observations are considered. The whole algorithm is provided below.

```
Step 1: initialize none zero observation counter ‘nzo’ to ‘zero’
Step 2: Repeat for the values of I from 1 to n.
Step 3: Input an observation, say O.
Step 4:  if $O_i$ = zero go to step 5
        Otherwise nzo= nzo+1
Step 5: Go to Step 2 for next i unless $i \leq n$.
Step 6: Write the counter nzo.
```

Step 7: Stop.

Example 6: Algorithm to find the roots of a quadratic equation $ax^2+bx+c=0$ when discriminate is non negative. The roots are to be stored in R_1 and R_2:

Step 1: Input a, b, c

Step 2: Evaluate the discriminate D = b^2-4ac

Step 3: Check; if D < zero, go to step 4

Otherwise evaluate $R_1 = \{(-b-\sqrt{D})/2a\}, R_2 = \{(-b-\sqrt{D})/2a\}$ and go to step 6

Step 4: Writ4e a message “discriminate is negative”

Step 5: Go to Step 7

Step 6: Write R_1 and R_2

Step 7: Stop.

Different algorithms may execute the same task using a different set of instructions, using less or more time, space, or effort than others. For example, given two distinct recipes for creating potato salad, one may have peeled the potato before boiling it, while the other presents the stages in the opposite sequence, but both call for these steps to be repeated for all potatoes and to be completed when the potato is ready to be eaten. The analysis and study of algorithms is a branch of computer science that is frequently conducted abstractly, without the use of a specific programming language or implementation. In this respect, algorithm analysis is similar to other mathematical disciplines in that it focuses on the algorithm's basic features rather than the intricacies of any given implementation. Pseudocode is commonly used for analysis since it is the most simple and universal form.

10.2.3. Various analyses of algorithms

Algorithm analysis in computer science is evaluating algorithm performance and behavior in terms of time complexity, space complexity, and other considerations. Here are the many sorts of analysis frequently performed on algorithms.

- **Time Complexity Analysis:** Time complexity analysis determines how long an algorithm takes to run as a function of input size. It estimates how the running time grows with additional input. Time complexity is commonly expressed using notations such as Big O, Big Theta, and Big Omega.
- **Space Complexity Analysis:** An algorithm's memory requirements are determined by space complexity analysis, which takes into account the size of the input. It assesses memory consumption and aids in determining if an algorithm can handle big inputs or is practical for contexts with limited memory. Big O notation is used to indicate space complexity, much like it is for time complexity.
- **Worst-Case Analysis:** Worst-case analysis establishes the maximum time or memory that an algorithm needs for each input size. It takes into account the situation in which the algorithm will run the longest or use the most memory. It gives an upper bound on the algorithm's performance.
- **Average-Case Analysis**: When calculating an algorithm's expected performance over all possible inputs of a given size, average-case analysis accounts for the likelihood that distinct inputs will occur. It offers a more accurate evaluation of an algorithm's performance by taking the input distribution into account.
- **Asymptotic Analysis:** The behavior of an algorithm as the input size gets closer to infinity is the main subject of asymptotic analysis. It offers a high-level comprehension of the algorithm's scalability with big inputs. Big O notation is most frequently used in asymptotic analysis to indicate the upper bound of an algorithm's growth rate.
- **Empirical Analysis:** Empirical analysis is the process of applying an algorithm to real inputs and evaluating its output. It offers practical measures of memory use and algorithm execution time. Validating theoretical analysis and evaluating an algorithm's performance in real-world scenarios can both benefit from empirical investigation.
- **Algorithmic Paradigms:** Algorithm analysis also entails investigating several algorithmic paradigms, including divide and conquer, dynamic programming, greedy algorithms, and others. Understanding the strengths and limits of each paradigm aids in the analysis and design of effective algorithms for certain problem domains.

Researchers and practitioners can assess algorithm efficiency, scalability, and viability using these numerous analyses. This analysis aids in determining the best method for a given problem and maximizing its performance across various input sizes and conditions.

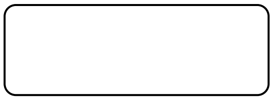
10.3. Flow Chart

A flowchart is a visual depiction of a process or algorithm that uses symbols and arrows to show the flow of control inside a system. It graphically depicts the steps required to complete a task or solve an issue. Flowcharts typically comprise symbols for start and finish points, processing steps, decision points, input/output activities, and connectors that depict the flow of control between processes. Flowcharts assist clarify the logic and structure of algorithms by mapping out the sequence of activities and decision points in a process. This makes them easier to comprehend, evaluate, and convey. Before writing a program of great complexity, it is vital to clearly specify it. In the early days of writing programs, it was unclear how they should be specified. Because there was no concept of software engineering as a formal discipline at the time, it was assumed that defining the execution sequence of a program was sufficient, and flow charts were developed. A flow chart is a graphical depiction of an algorithm that is primarily used to formulate and understand the technical aspects of the program. A flow chart is drawn using a standard convention that includes a variety of shapes. Each shape represents a certain lesson. The step-by-step process is depicted with lines, arrows, and boxes of various forms to demonstrate the flow of the process.

A flow chart is a graphical depiction of an algorithm that is primarily used to formulate and understand the technical aspects of the program. A flow chart is drawn using a standard convention that includes a variety of shapes. Each shape represents a certain lesson. The step-by-step process is depicted with lines, arrows, and boxes of various forms to demonstrate the flow of the process.

Flow charts are extremely valuable in program development, providing excellent documentation with a strong visual effect. The key advantage of designing a flow chart is that one is not bothered with the complexities of programming language and instead focuses on the logic of the activity to be completed. Furthermore, a flow chart's graphical format aids in the detection of logical sequence flaws. The commonly used symbols in a flow chart are given below.

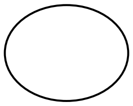
1. **Terminator:** An extended oval flow chart (rectangle with rounded ends) that indicates the start or end of the process. It is the first and last symbol used in program logic, and it typically contains the words "Start" and "End".



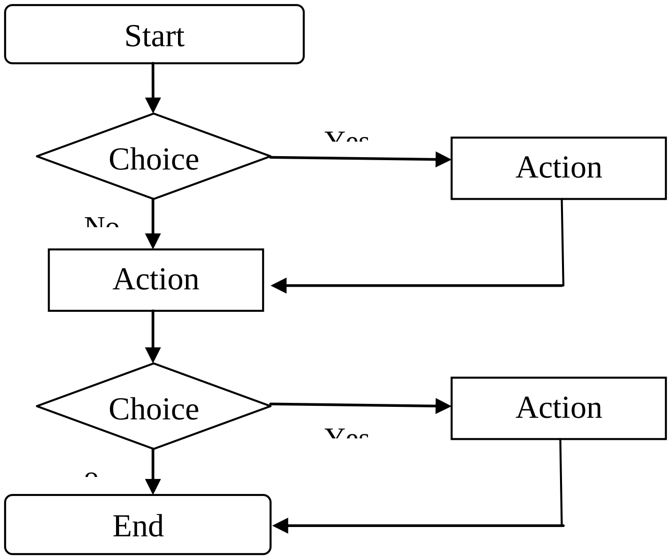
2. **Input/ Output:** A box shaped like an aparallelogram represents either an input (such as Read) or an output (such as Write).

3. **Processing:** The processing symbol is a rectangle, which is used in flowcharts to indicate arithmetic and data transfer instructions.

4. **Decision:** The symbol used for this purpose is a rhombus (a diamond-shaped box), indicating the point at which a decision must be made and allows for branching. The criteria for making a decision should be clearly stated in the discussion box. A diamond typically includes one arrow leading in and two or more leading out, indicating many ways the control can proceed from that point. A diamond is used in decision statements such as "If A is less than 10, proceed to add B to C; otherwise, multiply C and D".
5. **Flow lines:** Flow lines, represented by arrows, are used to show the flow of operations. Thus, the purpose of flow lines is to depict the exact sequence in which the instructions will be performed.
6. **Connector:** When the number and direction of flow lines become tangled, it is useful to use the connector sign as a substitute. A connector is symbolized by a circle, which may contain a letter or digit to signify the connectivity.



A simple flowchart defines the starting and finishing points of a process, the sequence of events in the process, and the decision or branching points along the way. This is what a basic flow chart looks like.



10.3.1. Characteristics of flowcharts

- Flowcharts are visual representations of processes or algorithms. Different shapes, such as rectangles, diamonds, circles, and arrows, are used to represent different aspects and their interactions, making it easier to grasp the flow of control and data.
- Flowcharts show the steps of a process or algorithm. The flowchart depicts the order in which the instructions are carried out, beginning at the top and progressing through the steps indicated by arrows or lines.
- Flowcharts incorporate decision points that direct control flow depending on certain conditions. These decision points are generally depicted by diamond-shaped symbols, indicating that the algorithm will take several courses based on the outcome of a condition.

- Flowcharts show the input and output parts of a process or algorithm. Input is typically represented by parallelogram-shaped symbols, whereas output is represented by rectangles or other appropriate shapes.
- Flowcharts can represent modular structures, with subroutines or modules shown as distinct flowcharts or sub-sections within the main flowchart.
- Loop symbols, such circles or arrows, can be used in flowcharts to illustrate iterative structures.
- Flowcharts are easy to understand for both technical and non-technical users.
- Standardized symbols and conventions ensure uniform representation across flowcharts.
- Flowcharts serve as documentation and analytical tools. They can be used to describe current algorithms or processes, comprehend and debug complex code, and assess the efficiency and efficacy of an algorithm or program.
- Flowcharts provide versatility in design and representation.

10.3.2. Benefits of flowchart

The benefits of flowchart are as follows:

1. Visual Clarity: One of the most significant advantages of a flowchart is its ability to show several steps and their sequence within a single page. Stakeholders throughout an organization can readily grasp the workflow and determine which steps are superfluous and which should be enhanced.
2. Instant Communication: Teams can utilize flowcharts instead of meetings. Simply explaining progresses provides a simple, visual technique for team members to quickly understand what they need accomplish step by step.
3. Effective Coordination: A flowchart can help project managers and resource schedulers order events and limit the possibility of overburdening team members. Eliminating unnecessary steps saves time and resources.
4. Efficiency Increase: Flowcharts offer substantial benefits in terms of increased efficiency. The flowchart depicts each step required to complete a procedure. A flowchart can assist a designer in removing unnecessary stages and errors from a

workflow. The flowchart should only include steps required to complete the procedure.

5. Effective Analysis: A flowchart can assist assess the situation more effectively. It specifies what type of action each step in a process demands. In general, a rectangle with rounded corners denotes the start or conclusion of the process, a diamond shape indicates the moment at which a decision is required, and a square block represents an action made within the process. A flowchart may also incorporate symbols indicating the type of media on which data is stored, for as a rectangle with a curved bottom for a paper document or a cylinder for a computer hard drive.
6. Problem-Solving: Flowcharts divide an issue into easily identifiable components. The flowchart explains how to solve a complex problem. A flowchart lowers the likelihood that a crucial step for solving an issue will be overlooked because it appears obvious. This saves money and time.
7. Proper Documentation: Digital flowcharts are an excellent form of paperless documentation that is required for a variety of applications, hence increasing efficiency.

Some examples on flow- chart are given below

Example 6: Flow chart to pick the largest of three numbers:

Algorithm to pick the largest of three numbers was attempted in example 3. The following figure shows the flow chart of the same example.

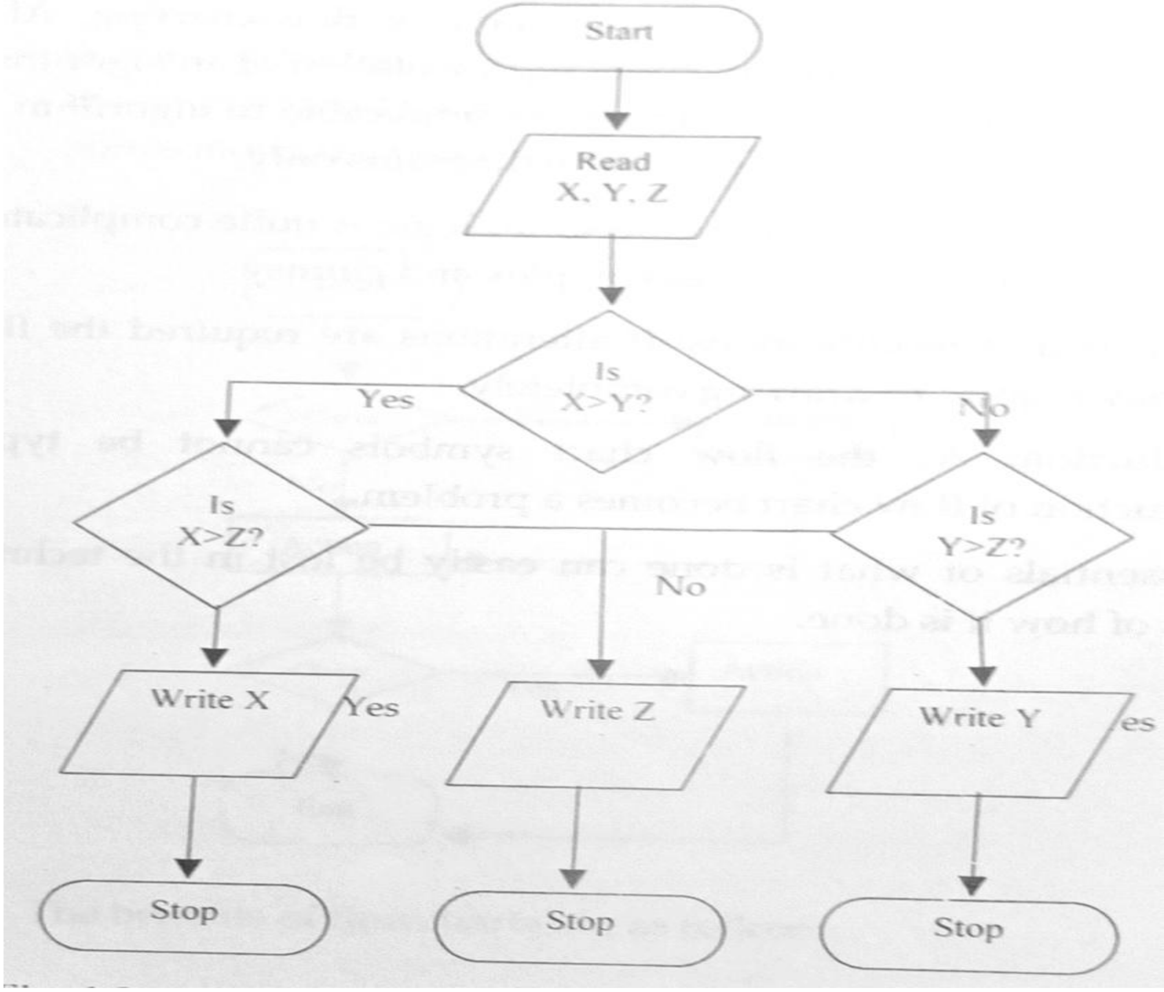


Fig. 10.2: A flow chart to pick the largest of three numbers X, Y and Z.

Example 7: Flow chart to count numbers of non zero observation in a list of n observations:

The solution to this example in the form of algorithm was provided in example 5. We shall draw below a flow chart based on the various steps of the algorithm. In the illustration, we have used a new pictorial representation, a hexagon, in which the number of repetitions and the last of repetition are shown.

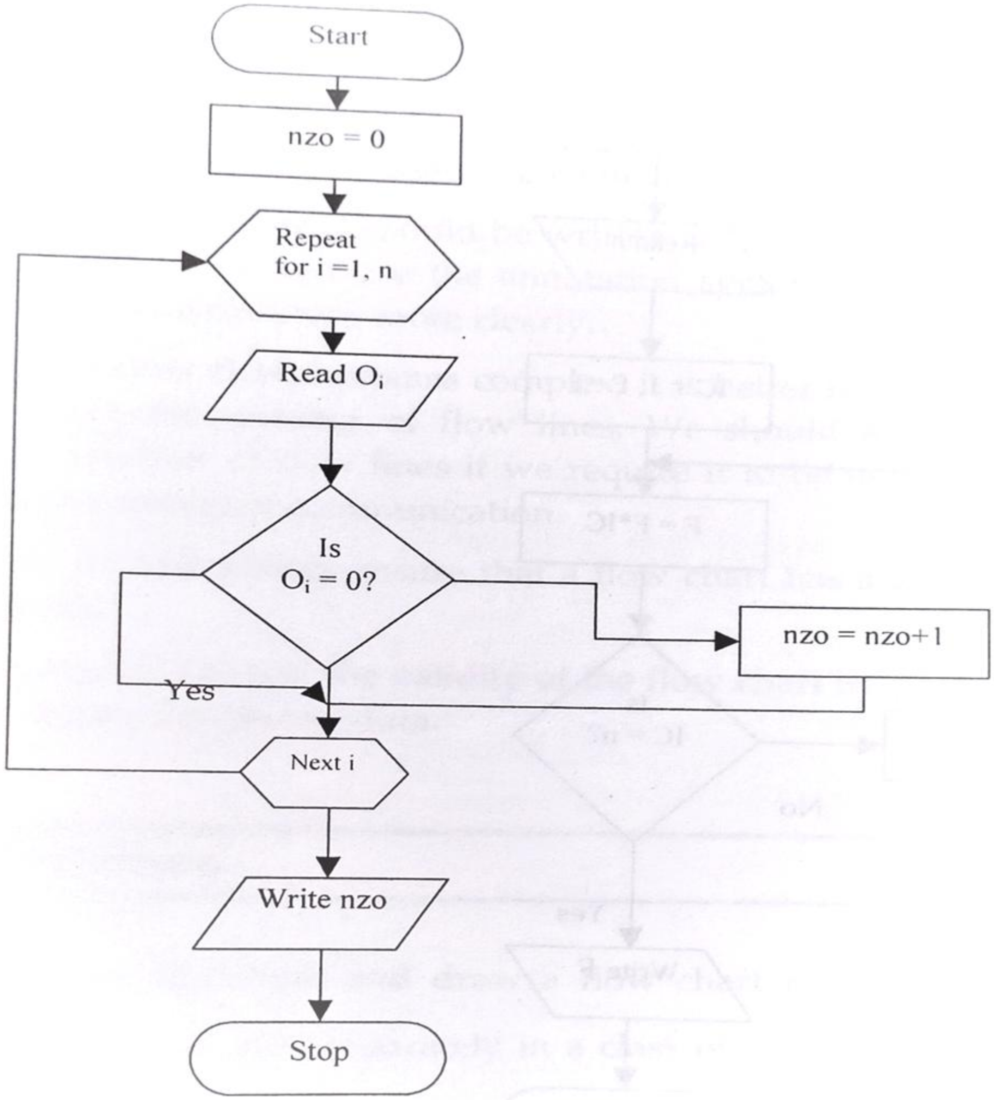


Fig.10.3: A flow chart to count number of non-zero observation in a list of n observations.

Example 8: Flow Chart to obtain factorial of positive integer n:

The factorial of n is the product of the first n natural integers. To build a flow chart, first enter the number of observations, then initialize and set two variables to unity. The first of these two variables will simply serve as a counter, while the second will be updated at each step and eventually yield the appropriate factorial. The graphic below depicts the flow chart for calculating the factorial of a given number n. The algorithm for the problem has not been provided, although it is assumed that it can be done routinely if the flow chart is known.

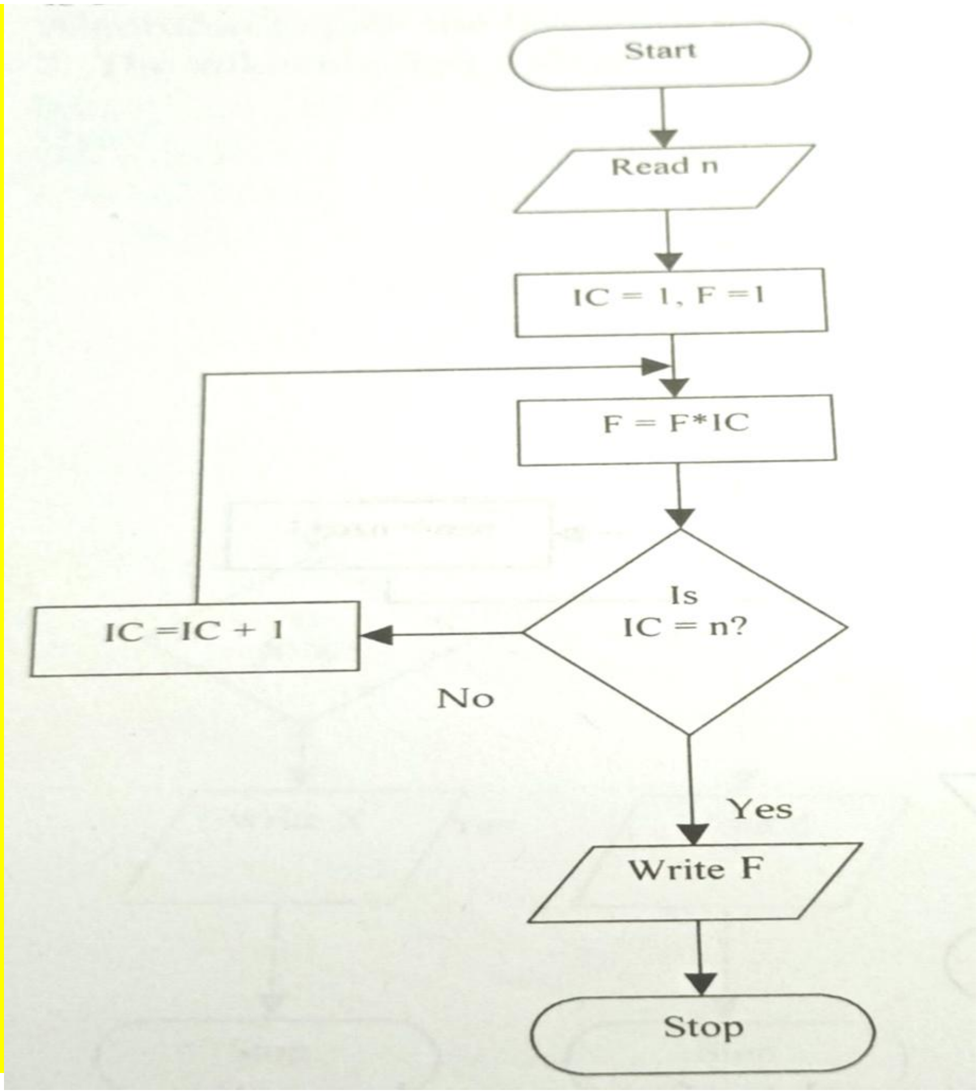


Fig. 10.4: A flow chart to obtain factorial of a positive integer n.

103.3. Some Guidelines on Flow charting

a. While drawing a flow chart all necessary requirements should be listed out in logical order.

b. The flow chart should be clear, neat and easy to follow. There should not be any room for ambiguity in understanding the flow chart.

c. The usual direction of the flow of a procedure must be from left to right or from top to bottom.

d. Only one flow line should come out from a process symbol.

e. Only one flow line should enter a decision symbol but two or three flow lines, one for each possible answer, may leave the decision symbol.

f. Only one flow line is used in conjunction with a terminal symbol.

g. A brief description should be written within a standard symbol. If necessary, we can use the annotation symbol to describe data or computational steps more clearly.

h. If the flow chart becomes complex, it is better to use connectors to reduce the number of flow lines. We should always avoid the intersection of flow lines if we require it to be more effective and better means of communication.

i. We should always ensure that a flow chart has a logical start and finish.

j. It is useful to test the validity of the chart by passing through it with a simple test data.

10.4. Summary

A flowchart template is a graphic that depicts a linear sequence of steps to define a broader process. Distinct steps in the process will be represented by distinct forms, which will be linked by directing arrows to send the user in the appropriate way. Flowcharts have traditionally been used to help individuals make rapid decisions or automate repetitive procedures, and they can also be used to visually describe workflows. An algorithm is a series of procedures for addressing a certain problem. To be considered an algorithm, a set of rules must be unambiguous and have a defined end point. Algorithms can be written in any language, including natural languages like English or French and computer languages like FORTRAN or C. We employ algorithms on a daily basis; for example, a cake recipe is an algorithm. Most programs, with the exception of some artificial intelligence applications, are made up of algorithms. One of the most difficult aspects of programming is developing elegant algorithms, which are straightforward and involve as few steps as feasible. A flow chart is a visual explanation of an activity or process using pictorial representations. A flow chart is a visual explanation of an activity or process using pictorial representations. Each action is represented by a shape that leads to the next action or activities, with each shape connected to the next by a line to indicate the flow of the activity. Flowcharts have traditionally been used to help individuals make rapid decisions or automate repetitive procedures, and they can also be used to visually describe workflows. A flowchart's various symbols and forms assist the user understand direction, action, and outcomes.

10. 5.Further suggested readings

Q.1. Write an algorithm and draw a flow chart to pick the largest of four real numbers.

Answer:---

Q.2. Write an algorithm and draw a flow chart to convert Centigrade temperature to Fahrenheit.

Answer:---

Q.3. Write an algorithm and draw a flow chart to evaluate $S_1 = \sum_{i=1}^{n} X_i$ and $S_2 = \sum_{i=1}^{n} X_1^2 - (s_1/n)^2$.

Answer:---

Q.4. A Fibonacci sequence is defined as 0,1,1,2,3,5,8,13,21,34,55, 89……… The first and second terms in the sequence are 0 and 1, respectively and the third and subsequent terms are found by adding the preceding two terms. Draw a flow chart to obtain all the numbers in Fibonacci sequence that are less than 200.

Answer:---

Q.5. Draw a flowchart to arrange a given set of data in an ascending order.

Answer:--

10.6. Further Readings

1. Chauhan Sunil, Saxena, Akash and Gupta Kratika. *Fundamentals of computer*, Laxmi Publications, (2006).
2. Cormen Thomas H.; Leiserson, Charles E.; Rivest, Ronald L.; *Introduction to Algorithm*, First edition, MIT Press and McGraw- Hill, (1990).
3. Forsythe,Alexandra I. *Computer Science: A Primer*,Wiley, (1969).
4. Lipshutz, Seymour, *Schaum's Outline of Essential Computer Mathematics,* McGraw-Hill, (1987).
5. Wilde, Daniel Underwood *An Introduction to Computing: Problem Solving, Algorithms and Data Structures*, Prentice-Hall, (1973).

Unit-11: Computation with MS Excel

Contents

11.1. Introduction

MS Excel is used to build electronic computation worksheets. These worksheets are used for organizing, computing, analyzing, and evaluating data, as well as doing financial calculations. Electronic workbooks are simple to create, manipulate, and update. Electronic spreadsheets are superior to manual spreadsheets. When there are too many adjustments, you must write a new manual worksheet, making electronic worksheet editing more convenient. A variety of calculations can be programmed and performed automatically in MS Excel. There are specific routines for performing financial computations and statistical analyses. Different books and writers use the phrases worksheet, spreadsheet, and workbook interchangeably to refer to a calculation worksheet created using Microsoft Excel. The MS Excel workbook is made up of cell grids organized in a tabular format with rows and columns crossing. Each cell serves as an autonomous record-keeping unit. Data stored in cells can be numeric, textual, date, or time. This unit will cover all topics, from MS Excel basics to advance. There are specific routines for performing financial computations and statistical analyses. Different books and writers use the phrases worksheet, spreadsheet, and workbook interchangeably to refer to a calculation worksheet created using Microsoft Excel. Here are some of the most common uses of Microsoft Excel. MS Excel is a key component of Microsoft Office. It is used to create math worksheets, analyze data, and present results in the form of charts.

Objectives:

After reading this unit, you should be able to:

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

11.2. Formula

There are numerous Excel formulae and functions available depending on the type of action you wish to conduct on the dataset. We will look at formulae and functions for mathematical operations, character-text functions, data and time, sum if-count if, and a few lookup functions. Let's look at the top 25 Excel formulas you should know. In this part, we've organized 25 Excel formulae by their operations. Let's start with the first Excel formula on the list. In Microsoft Excel, a formula is an expression that operates on values in a set of cells. These formulas produce a result, even if it is a mistake. Excel formulae allow you to do mathematical operations such as addition, subtraction, multiplication, and division. In addition, we may use Excel to calculate averages and percentages for a range of cells alter date and time variables, and much more. Formulas are used to make calculations automatically. When we apply a formula to a cell, its value is mathematically dependent on the values of the other fields. For example, the formula c3 = b3 + a3 indicates that the value of c3 is the total of the values entered in the b3 and a3 fields. As you modify the value of b3 or a3 cells, the value of c3 changes automatically. Every formula starts with an equals symbol (=) or a plus sign. These are recognized as reserved words for formula entry in a worksheet. Click on the cell where the formula is to be entered. Enter the = or + symbol before typing a formula into the formula bar, and then type the formula. For example, to calculate c3 = b3 + a3, follow the steps below. First, enter values into the b3 and a3 fields. Clicking over the c3 cell moves the focus to it. Enter the following string "= b3 + a3" into the formula text box and click enter; the sum of b3 and a3 is now available in c3. Instead of entering a formula, just put = or + symbols into the formula bar; then move your mouse to the b3 cell and click; b3 will appear on the formula bar. Enter the + sign from the keyboard to see it appear on the formula bar next to b3. Then, click on the a3 cell to complete your formula (type "="; click on b3; type +; click on a3). Thus, you can either type the URL directly or choose it by clicking on a specific cell to put it into the calculation. If you have assigned names to cells as described in subsection 8.10, you can use their names rather than their addresses. If a formula is not preceded by the + or = sign, it will be interpreted as a string. Sometimes formulas employ standard functions rather than straightforward mathematical calculations. MS Excel defines the following process for performing the computation.

- Select a cell.
- Enter an equal sign by clicking the cell and typing =.
- Select a cell by entering its address or from the list.
- You have to input an operator.
- Type the address of the next cell in the selected cell.
- Press Enter.

Another term that comes up frequently in Excel formulas is "function". The words "formulas" and "functions" are occasionally used interchangeably. They are similar but not identical. A formula begins with the equal sign. Meanwhile, functions are employed to execute sophisticated calculations that cannot be done by hand. Excel functions are named according to their intended application. There are numerous Excel formulae and functions available depending on the type of action you wish to conduct on the dataset. We will look at formulae and functions for mathematical operations, character-text functions, data and time, sum if-count if, and a few lookup functions. Let's look at the top 25 Excel formulas you should know. In this article, we've organized 25 Excel formulae by their operations. Let's start with the first Excel formula on the list.

11.3. Function Specifically

In Microsoft Excel, a function is referred to as a pre-defined or pre-structured formula that may compute particular outputs based on the variables that have been chosen. A function is built upon a foundational syntax. It consists of the function name (such as SUM, PRODUCT, IF), an equal sign ("="), and the argument (s). Arguments are enclosed in parenthesis and contain the data that has to be computed.

Inserting Functions

To insert a function, see the below steps:-

1) Select the Formulas tab on the menu ribbon.

2) Select a function from the horizontal list, as shown below

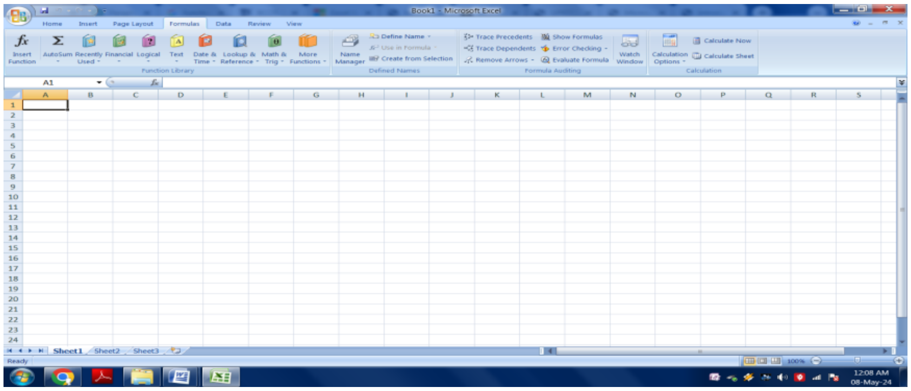
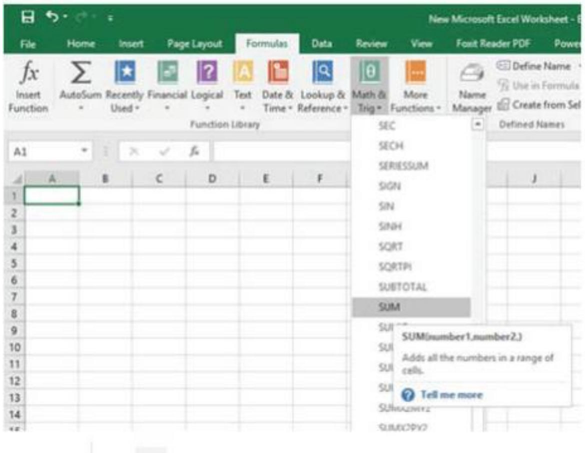


Fig. 11.1: Formulas Ribbon

3) Let us take an example of SUM.

4) This function is available under Math & Trig tab

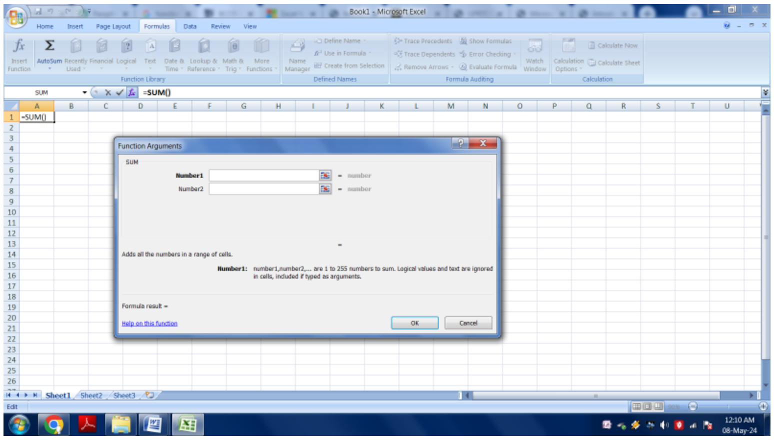


SUM(number1,number2,...)

Adds all the numbers in a range of cells.

Fig.11.2: Math & Trig Tab Section

5) When we click on this option, syntax will appear on the selected cell, and a dialog box would appear.



=SUM()

Function Arguments

SUM

Number1 = number

Number2 = number

Adds all the numbers in a range of cells.

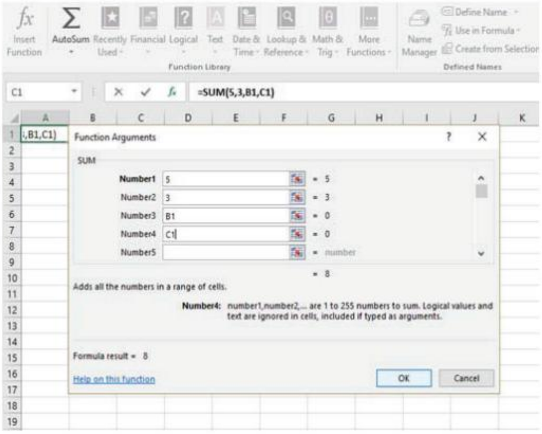
Number1: number1,number2,... are 1 to 255 numbers to sum. Logical values and text are ignored in cells, included if typed as arguments.

Formula result =

Help on this function

Fig.11.3: Sum Formula

6) We need to put in the required information for the function to calculate, in this case the cells that need to be added.



C1 =SUM(5,3,B1,C1)

Function Arguments

SUM

Number1 5 = 5

Number2 3 = 3

Number3 B1 = 0

Number4 C1 = 0

Number5 = number

= 8

Adds all the numbers in a range of cells.

Number4: number1,number2,... are 1 to 255 numbers to sum. Logical values and text are ignored in cells, included if typed as arguments.

Formula result = 8

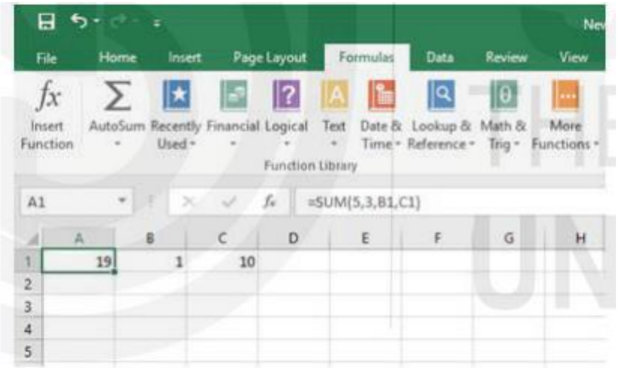
Help on this function

OK Cancel

Fig.11.4: Sum Formula Data

7) After putting in the information, press OK.

8) This will result in provision of selected cells sum, as in the figure, needful is to contain numerical values in the cells under process.



A1 =SUM(5,3,B1,C1)

	A	B	C	D	E	F	G	H
1	19	1	10					
2								
3								
4								
5								

Fig.11.5: Sum Formula Result

Built-in Function

The graphic below illustrates the six main categories of built-in functions found in Microsoft Excel:

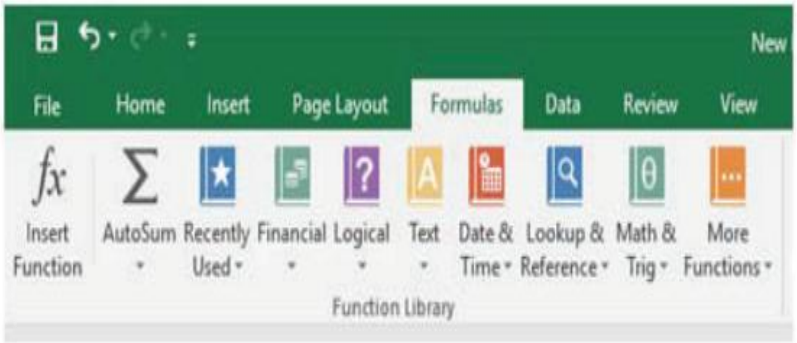


Fig.11.6: Formulas Ribbon

1. **Financial Functions**
   a. **PMT:** This function is used for calculating periodical installments for a loan at a constant interest rate.
   b. **PV:** This function is used for calculating the Present Value of a given Loan as an Investment Amount at fixed rate of interest.
   c. **FV:** This function is used for calculating the Future Value of the current investment amount; calculation is dependent on the fixed rate of interest.
2. **Logical Functions**

a. IF: It presents a reasonable contrast between the actual value and the predicted value.

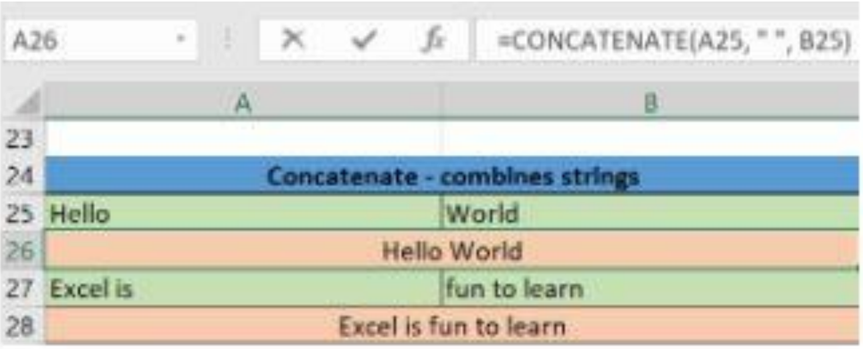
b. AND: Essentially, this serves as a helpful utility to help with more reasonable comparisons. When one or more inputs are FALSE, the output becomes FALSE. Otherwise, it stays TRUE, retaining all inputs TRUE.

c. OR: This function is also supporting logical in nature, but it operates on the basis of the following reasoning: if one or more of the inputs supply TRUE, the resultant outcome becomes TRUE; if all of the inputs provide FALSE, the final outcome becomes FALSE.

3. **Text Functions**

a. **Concatenate:** This function combines or connects many text strings into a single string. Given below are the different ways to perform this function. For example, we used the following syntax:

CONCATENATE =CONCATENATE(A25, " ", B25)



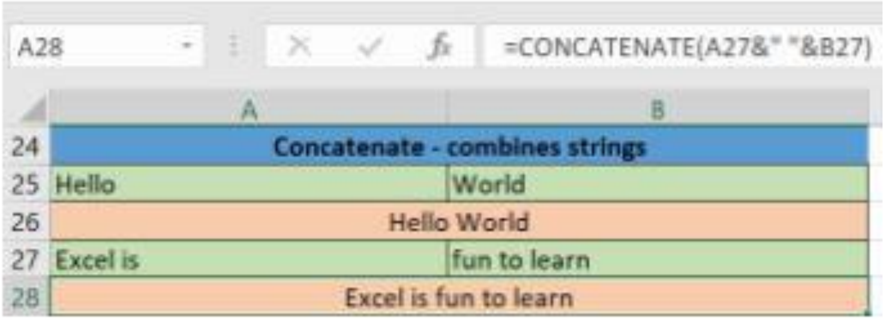
A26 =CONCATENATE(A25, " ", B25)

	A	B
23		
24	Concatenate - combines strings	
25	Hello	World
26	Hello World	
27	Excel is	fun to learn
28	Excel is fun to learn	

Fig.11.7: Concatenate function in Excel

In example, we have operated with the syntax:

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



A28 =CONCATENATE(A27&" "&B27)

	A	B
24	Concatenate - combines strings	
25	Hello	World
26	Hello World	
27	Excel is	fun to learn
28	Excel is fun to learn	

Fig.11.8: Concatenate function in Excel

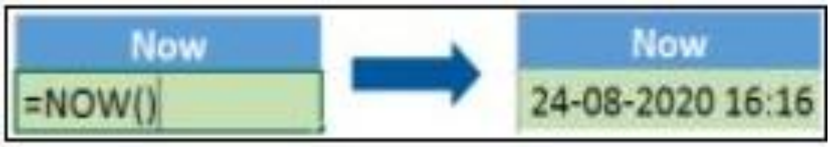
Those were the two ways to implement the concatenation operation in Excel.

b. **Dollar:** It converts a number to text with a dollar sign as its prefix.

4. Date & Time Functions

a. NOW()

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



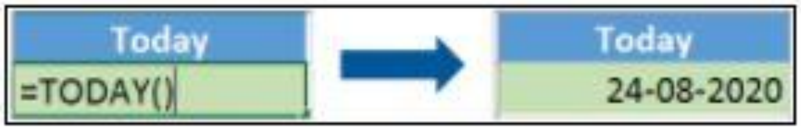
Now		Now
=NOW()	→	24-08-2020 16:16

Fig.11.9: Now function in Excel

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

b. TODAY()

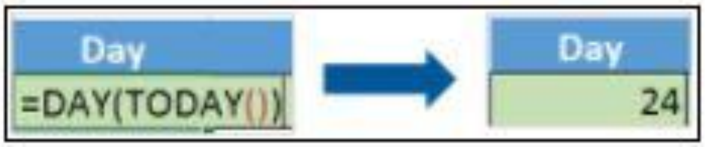
The TODAY() function in Excel provides the current system date.



Today		Today
=TODAY()	→	24-08-2020

Fig.11.20: Today function in Excel

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

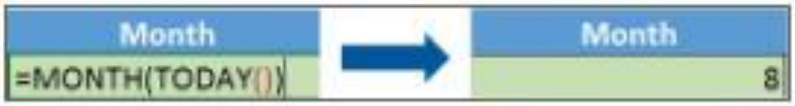


Day		Day
=DAY(TODAY())	→	24

Fig.11.21: Day function in Excel

c. **The MONTH ()**:

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

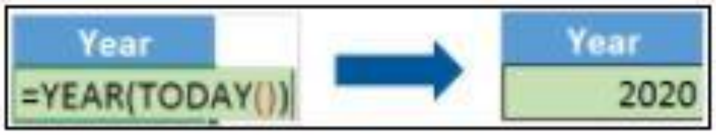


Month		Month
=MONTH(TODAY())	→	8

Fig.11.22: Month function in Excel

d. **The YEAR():**

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

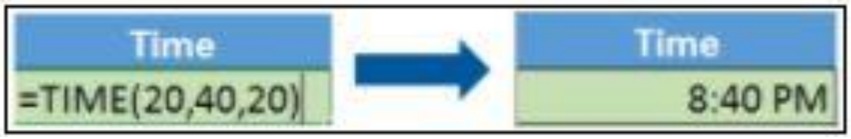


Year		Year
=YEAR(TODAY())	→	2020

Fig.11.23: Year function in Excel

e. TIME():

The TIME() function turns hours, minutes, and seconds entered as numbers into an Excel serial number formatted with a time.



Time

=TIME(20,40,20)

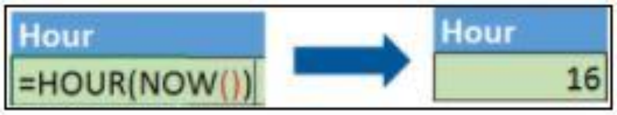
Time

8:40 PM

Fig.11.24: Time function in Excel

f. HOUR, MINUTE, SECOND

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



Hour

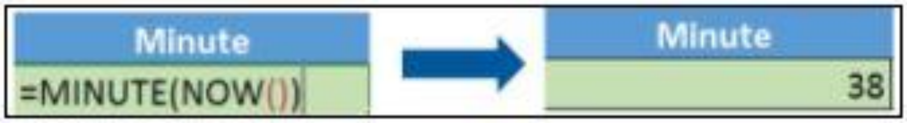
=HOUR(NOW())

Hour

16

Fig.11.25: Hour function in Excel

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



Minute

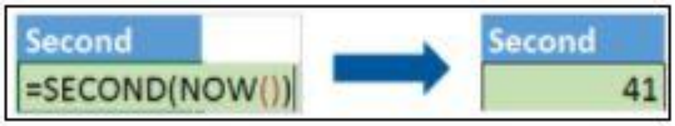
=MINUTE(NOW())

Minute

38

Fig.11.26: Minute function in Excel

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



Second

=SECOND(NOW())

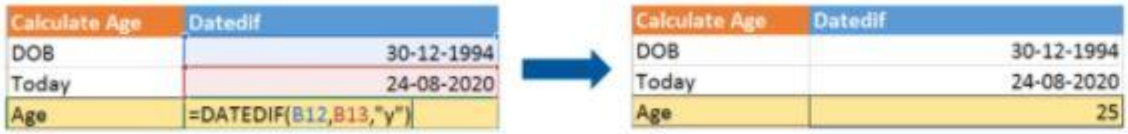
Second

41

Fig.11.27: Second function in Excel

g. DATEDIF

The difference between two dates expressed in years, months, or days is provided by the DATEDIF() function. The DATEDIF function is demonstrated in the example below, which determines an individual's age based on two dates: their birthdate and the current date.



Calculate Age	Datedif
DOB	30-12-1994
Today	24-08-2020
Age	=DATEDIF(B12,B13,"y")

Calculate Age	Datedif
DOB	30-12-1994
Today	24-08-2020
Age	25

Fig.11.28: Dated if function in Excel

Let's now go over a few essential advanced Excel functions that are frequently used for data analysis and report creation.

5. Lookup & Reference Functions

a. VLOOKUP

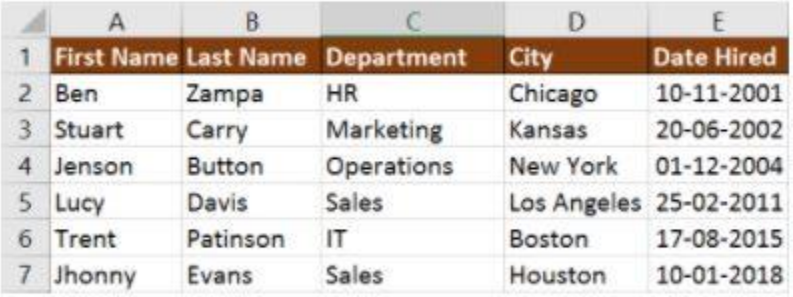
The VLOOKUP() function will be discussed next in this article. This is an acronym for the vertical lookup function, which searches the leftmost column of a table for a specific value. After that, it gives back a value from the specified column in the same row. The arguments for the VLOOKUP function are listed below: Lookup value: This is the value you need to search for in a table's first column.

Table: This denotes the table that the value was obtained from. Col index: The table column from which the value needs to be obtained.

Lookup by range: [optional] TRUE = roughly correspond (by default). False: a perfect fit. The table below will help us understand how the VLOOKUP function functions.

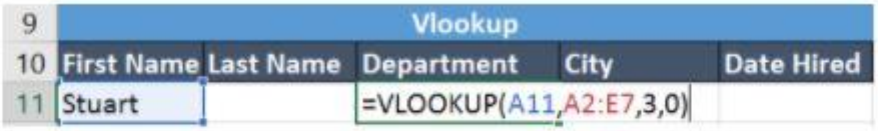
Stuart's department might be located by using the VLOOKUP function, as demonstrated below:

Table 11.1: Vlookup function in Excel



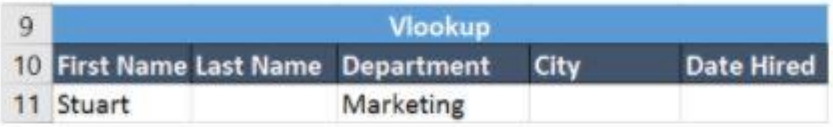
	A	B	C	D	E
1	First Name	Last Name	Department	City	Date Hired
2	Ben	Zampa	HR	Chicago	10-11-2001
3	Stuart	Carry	Marketing	Kansas	20-06-2002
4	Jenson	Button	Operations	New York	01-12-2004
5	Lucy	Davis	Sales	Los Angeles	25-02-2011
6	Trent	Patinson	IT	Boston	17-08-2015
7	Jhonny	Evans	Sales	Houston	10-01-2018

In this case, the lookup value is in cell A11; the table array is in cell A2: E7; the column index number containing departmental information is 3; and the range lookup is 0.



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

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



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

5.2.21. HLOOKUP

We also have a function called HLOOKUP(), or horizontal lookup, which is similar to VLOOKUP. The HLOOKUP function searches a table or array of benefits' top row for a value. It displays the value from a specified row in the same column.

The HLOOKUP function's arguments are listed below:

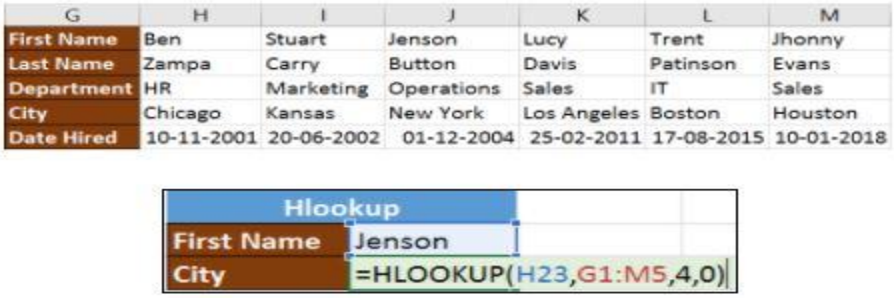
Lookup value - This indicates the value to lookup.

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

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

Range lookup - [optional] this is a Boolean to indicate an exact match or approximate match. The default value is TRUE, meaning an approximate match.

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

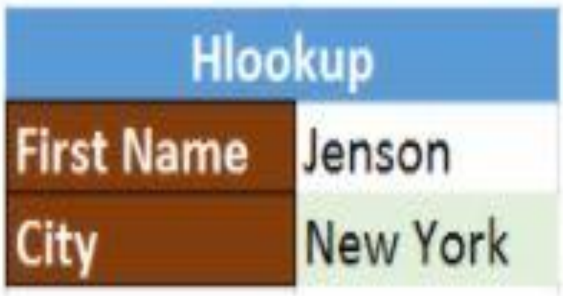


G	H	I	J	K	L	M
First Name	Ben	Stuart	Jenson	Lucy	Trent	Jhonny
Last Name	Zampa	Carry	Button	Davis	Patinson	Evans
Department	HR	Marketing	Operations	Sales	IT	Sales
City	Chicago	Kansas	New York	Los Angeles	Boston	Houston
Date Hired	10-11-2001	20-06-2002	01-12-2004	25-02-2011	17-08-2015	10-01-2018

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

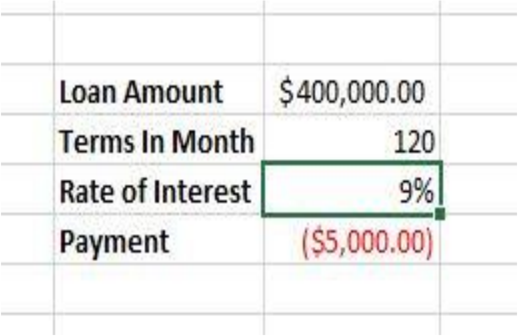
Fig11.29: Hlookup function in Excel

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



Hlookup	
First Name	Jenson
City	New York

You may become an excellent data analyst by learning analytics tools and techniques from our Master's program in data analysis! This training is ideal for you to launch your career. Enroll right away!



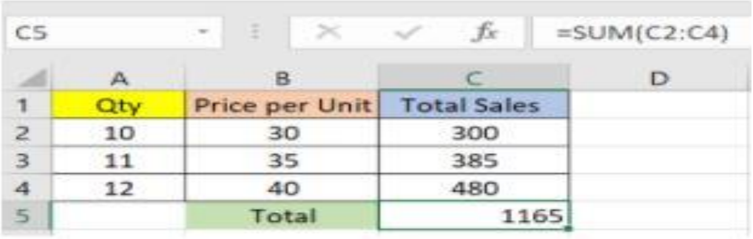
Loan Amount	$400,000.00
Terms In Month	120
Rate of Interest	9%
Payment	($5,000.00)

6. Math & Trig Functions

a. SUM

The SUM () function, as the name implies, returns the sum of the selected range of cell values. It performs the mathematical action called addition. Here is an example below:

Sum = "(C2:C4)"



C5 =SUM(C2:C4)

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

Fig.11.30: Sum function in Excel

As shown above, to calculate the total amount of sales for each unit, we simply used the function "= SUM (C2:C4)". This automatically sums c2, c3, and c4. The results are saved in C5.

11.4. Numeric/Mathematical Functions

Any decision-making process must include mathematics, and Microsoft Excel offers a comprehensive range of features to address mathematical issues. A distinct class of Math & Trig functions is dedicated to supplying help for trigonometric and mathematical problems.

Excel makes adding and subtracting simple; all you need to do is build a formula. Recall that the formula bar allows you to create and modify formulas, and that all formulas in Excel start with the equal symbol (=).

In one cell, add two or more integers. To begin a formula, click any blank cell and type the equal sign (=). Type a few integers followed by a plus sign (+) after the equal sign. As an instance, 40+21+14+3.Hit RETURN; 78 is the outcome

To begin a formula, subtract two or more numbers from a cell, click any blank cell, and then input the equal symbol (=). Type a few numbers separated by a minus sign (-) after the equal sign.

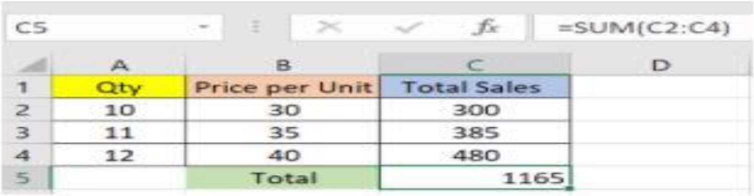
Let's say 70-20-15-3. If you use the sample numbers, presses RETURN; the output will be 32.

To begin a formula, subtract two or more numbers from a cell, click any blank cell, and then input the equal symbol (=). Type a few numbers separated by a minus sign (-) after the equal sign.For example, 70-20-15-3.

If you use the sample numbers, press RETURN; the output will be 32. The sum() mathematical function is explained below:

a. Sum() function:

Using their individual cell addresses, two or more numbers can be added using this function. The cell addresses are given to the function either directly or through the use of operators. In general, the outcome is the sum of the given integers; however, when there are gaps in the cell ranges, the outcome varies. The SUM () function, as the name implies, returns the sum of the selected range of cell values. It performs the mathematical action called addition. The examples are below:
Sum = "(C2:C4)"



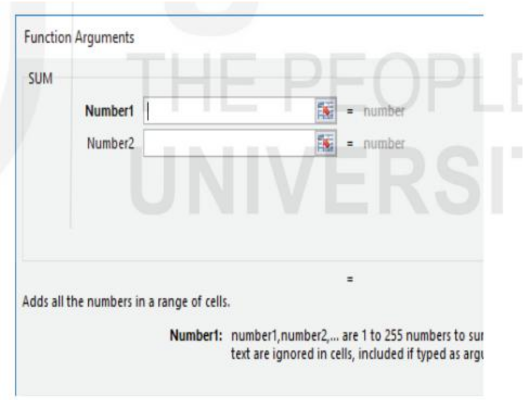
C5 =SUM(C2:C4)

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

Fig11.31: Sum function in Excel

As shown above, to calculate the total amount of sales for each unit, we simply used the function "= SUM (C2:C4)". This automatically sums c2, c3, and c4. The results are saved in C5.

Cell addresses of numbers must be supplied as parameters to the sum() function. Cells do not contain Text or Logical Values; they are only included if they are typed as arguments.



Function Arguments

SUM

Number1 = number

Number2 = number

=

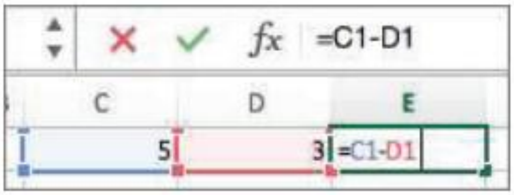
Adds all the numbers in a range of cells.

Number1: number1,number2,... are 1 to 255 numbers to sur
text are ignored in cells, included if typed as argu

Fig.11.32: Sum Formula Dialog Box

Subtract numbers using cell references A cell reference combines the row number and column letter, like A1 or F345. When you use cell references in a formula instead of the cell value, you can change the value in that cell without having to change the formula.

1) Type a number in cells C1 and D1.For example, 5 and 3.
2) In cell E1, type an equal sign (=) to start the formula
3) After the equal sign, type C1-D1



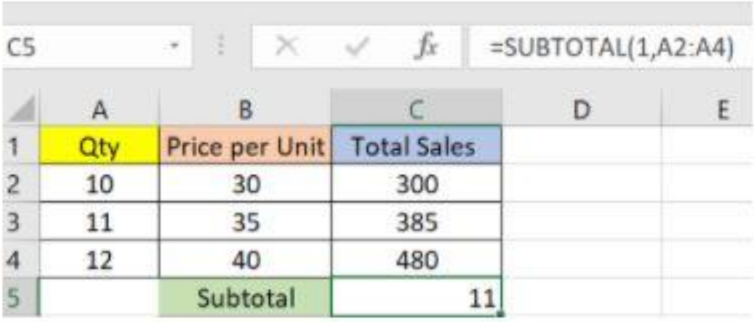
=C1-D1

	C	D	E
	5	3	=C1-D1

4) Press RETURN, If you used the example numbers the result is 2

b. Subtotal

Moving on, let us examine how the subtotal function works. The SUBTOTAL () function retrieves the subtotal in a database. You can choose between average, count, sum, min, max, min, and other options. Let's look at two such situations.

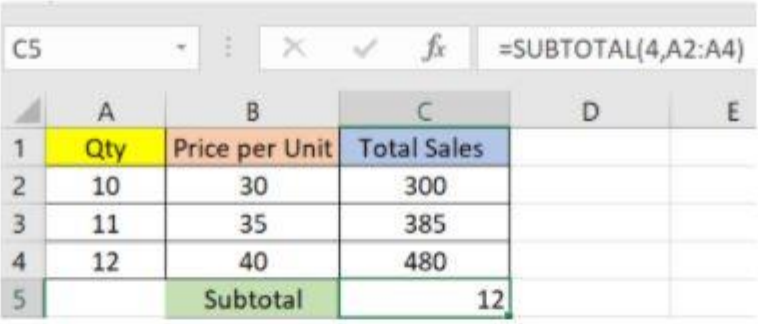


C5 =SUBTOTAL(1,A2:A4)

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

Fig.11.33: Subtotal function in Excel

In the above example, we calculated the subtotal for cells A2 through A4. The function used is: SUBTOTAL = SUBTOTAL(1, A2, A4). In the subtotal list, "1" represents average. As a result, the above code will return the average of A2 and A4, which is 11, as stored in C5. Similarly, "=SUBTOTAL(4, A2:A4)". This selects the cell with the greatest value between A2 and A4, which is 12. Incorporating "4" inside the function produces the best results.

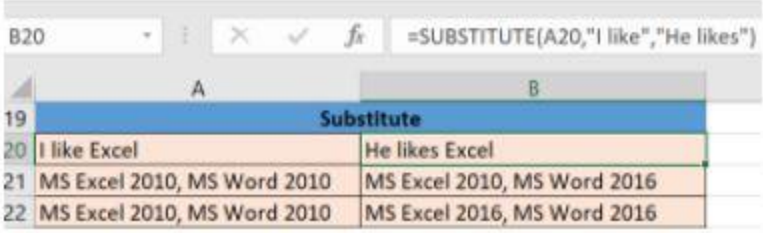


=SUBTOTAL(4,A2:A4)

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

c. Substitute

The SUBSTITUTE() function substitutes current text with new text in a text string. The syntax is: "=SUBSTITUTE(text, old text, new text, [instance num])". Here, [instance num] refers to the index position of the current texts more than once. The following are some examples of this function: Typing "=SUBSTITUTE(A20, "I like","He likes")" replaces "I like" with "He likes".

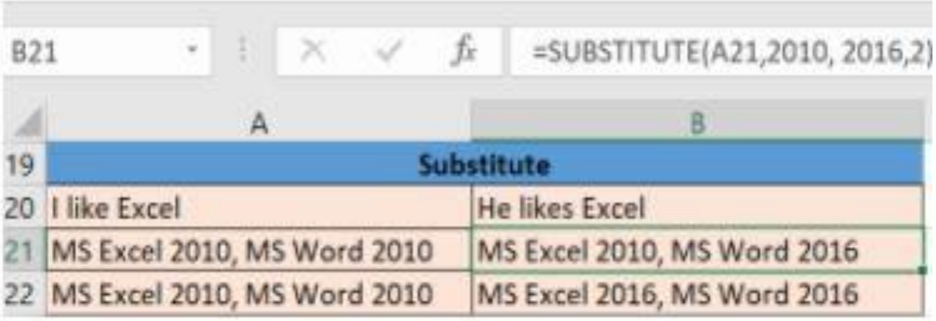


=SUBSTITUTE(A20,"I like","He likes")

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

Fig.11.34: Substitute function in Excel

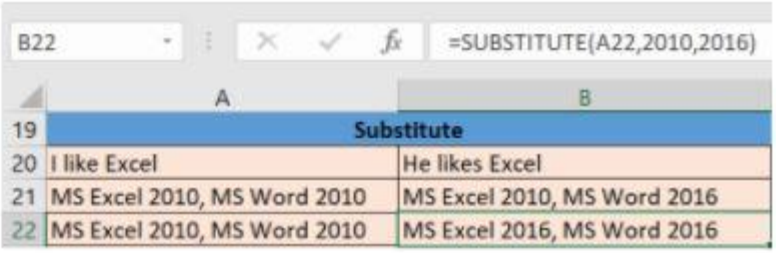
Next, we replace the second 2010 in cell A21 with 2016 by typing "=SUBSTITUTE(A21,2010, 2016,2)". Substitute function (i)



=SUBSTITUTE(A21,2010, 2016,2)

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

We are now replacing both 2010s in the original text with 2016 by putting "=SUBSTITUTE(A22,2010,2016)".



=SUBSTITUTE(A22,2010,2016)

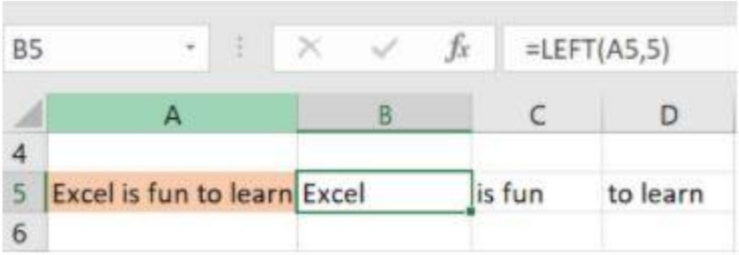
	A	B
19	Substitute	
20	I like Excel	He likes Excel
21	MS Excel 2010, MS Word 2010	MS Excel 2010, MS Word 2016
22	MS Excel 2010, MS Word 2010	MS Excel 2016, MS Word 2016

Fig.11.35: Substitute function in Excel

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

d. Left, Right, Mid

The LEFT() function returns the number of characters from the beginning of a text string. Meanwhile, the MID() function retrieves the characters in the center of a text string based on a beginning location and length. Finally, the right () function returns the number of characters remaining at the end of a text string. Let us understand these functions by a few examples. For example, we can use the function left to get the leftmost word in the statement in cell A5.

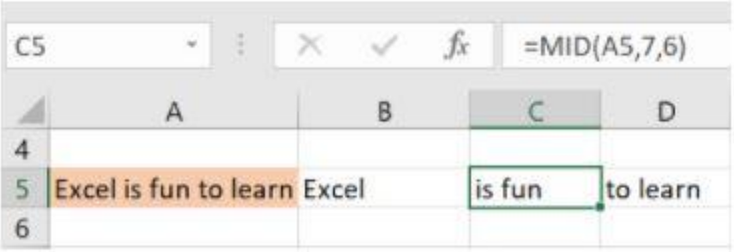


=LEFT(A5,5)

	A	B	C	D
4				
5	Excel is fun to learn	Excel	is fun	to learn
6				

Fig.11.36: Left function in Excel

Shown below is an example using the mid function.

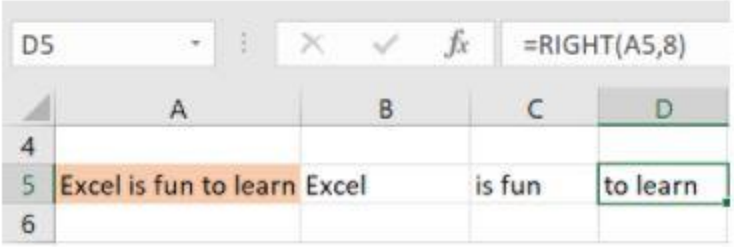


=MID(A5,7,6)

	A	B	C	D
4				
5	Excel is fun to learn	Excel	is fun	to learn
6				

Fig.11.37: Mid function in Excel

Here, we have an example of the right function.



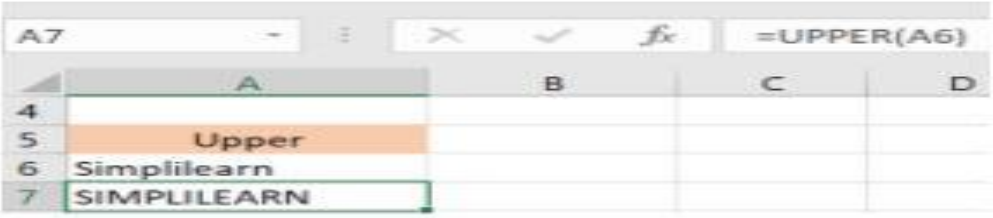
=RIGHT(A5,8)

	A	B	C	D
4				
5	Excel is fun to learn	Excel	is fun	to learn
6				

Fig.11.38: Right function in Excel

e. Upper, Lower, Proper

Any text string can be made uppercase by using the UPPER() function. The LOWER() function, on the other hand, lowercases any text string. Any text string can be converted to proper case using the PROPER() method, which means that all of the letters in a word will be lowercase except for the first one. To gain a better understanding, let us consider the following examples: Here, the content in A6 has been changed to a complete uppercase text in A7.

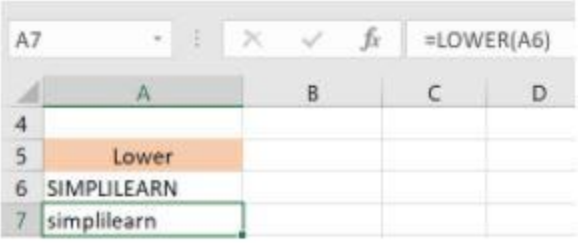


=UPPER(A6)

	A	B	C	D
4				
5	Upper			
6	Simplilearn			
7	SIMPLILEARN			

Fig.11.39: Upper function in Excel

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

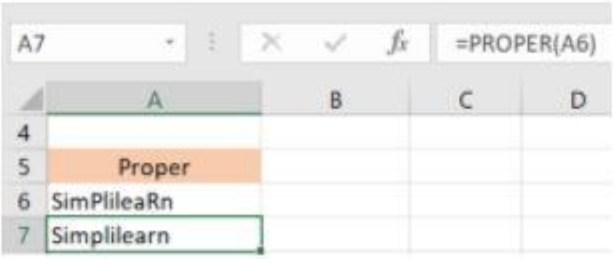


A7 =LOWER(A6)

	A	B	C	D
4				
5	Lower			
6	SIMPLILEARN			
7	simplilearn			

Fig.11.40: Lower function in Excel

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



A7 =PROPER(A6)

	A	B	C	D
4				
5	Proper			
6	SimPlileaRn			
7	Simplilearn			

Fig.11.41: Proper function in Excel

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

11.5. Statistical Functions

a. Average

When numbers are submitted as arguments to the average() function, the arithmetic mean of those numbers is produced. For example, if A2:A10 contains numerical values, the formula =AVERAGE(A2:A10) yields the average of those values. For instance, the grades that five students out of twenty received are listed below. The average of those marks is now determined using the formula [=average(B2:B6)]; this yields a result of 13.8.

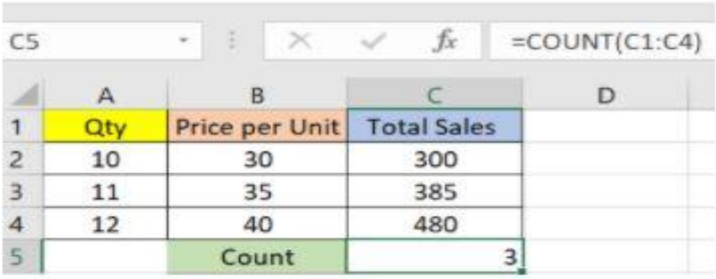
Table: 11.2: Calculation of Average

Student Name	**Marks [20]**	**Functions**
Arun	**12**	
Varun	**13**	
Ram	**15**	
Kriti	**14**	Average(B2:B6)
Kiran	**15**	**13.8**

b. COUNT

The COUNT() function computes the total number of cells in a range containing a number. It excludes the blank cell and those that contain data in any format other than numeric.

COUNT = C



C5 =COUNT(C1:C4)

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

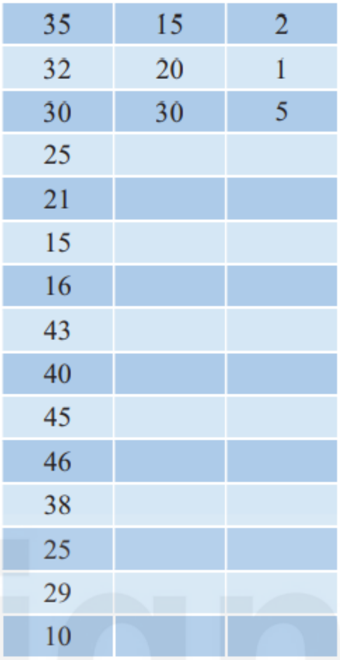
Fig.11.42: Microsoft Excel Function - Count

As shown above, we are counting from C1 to C4, which is ideally four cells. However, because the COUNT function only considers cells with numerical values, the answer is 3, as the "Total Sales" cell is excluded. If you need to count all cells containing numerical values, text, or any other data format, use the 'COUNT A ()' function. However, COUNT A () does not include any blank cells. COUNT BLANK () is used to determine how many blank cells are there in a range of cells.

c. Frequency

Frequency() is an array-based function that does exactly what its name suggests: it counts the number of arguments that are supplied as range to the method along with the criteria. For instance, the count of items in the data_array given as bins_array is provided by the number of items mentioned in the range for frequency(). = frequency between bins and data arrays The function's output reflects the count of items up to the amount specified in bins_array. In the data_array list, for instance, there are two numbers ranging from 0 to 15, one number from 15 to 20, and five numbers ranging from 20 to 30.

Table.11.3: Dataset for frequency

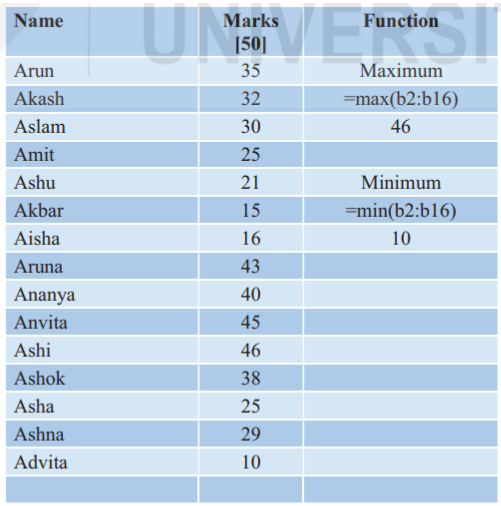


35	15	2
32	20	1
30	30	5
25		
21		
15		
16		
43		
40		
45		
46		
38		
25		
29		
10		

d. MAX, MIN

These two functions are providing the maximum or minimum out of the range of input values. Maximum number can be derived as outcome of =max (range) function. Minimum number can be derived as outcome of =min (range) function.

Table11.4: Dataset for MAX, MIN

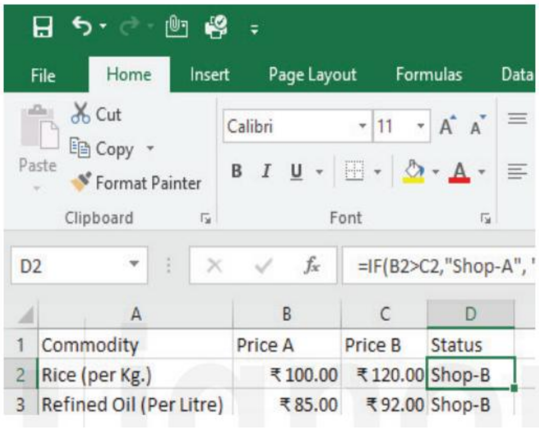


Name	Marks [50]	Function
Arun	35	Maximum
Akash	32	=max(b2:b16)
Aslam	30	46
Amit	25	
Ashu	21	Minimum
Akbar	15	=min(b2:b16)
Aisha	16	10
Aruna	43	
Ananya	40	
Anvita	45	
Ashi	46	
Ashok	38	
Asha	25	
Ashna	29	
Advita	10	

11.6. Logical Function

The cell ranges mentioned in the formula have logical linkages that are implemented by means of logical functions. IF(): During the decision-making process, this function branches. This function is supplied three arguments.

- logical_test (required): for a binary outcome in the form of true or false.
- value_if_true (required): branching operation when logical_test return true.
- value_if_false (optional): branching operation when logical_test return false.



D2 =IF(B2>C2,"Shop-A",'

	A	B	C	D
1	Commodity	Price A	Price B	Status
2	Rice (per Kg.)	₹ 100.00	₹ 120.00	Shop-B
3	Refined Oil (Per Litre)	₹ 85.00	₹ 92.00	Shop-B

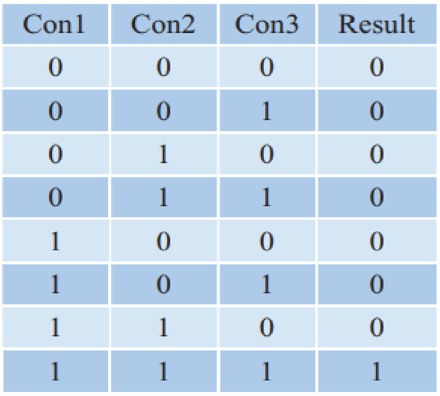
Fig.11.43: IF Function

In the example above, the consumer compares the pricing of goods from two separate stores using reasoning. The status indicates the store that was selected to purchase the goods.

AND

In order to work, this logical test needs at least two input parameters. It verifies if the conditions supplied as arguments in the function TRUE[1] or FALSE[0] result in a binary output. This function yields TRUE [1] as output only after all inputs reflect TRUE [1] as the result.

Table.11.5: Dataset for AND



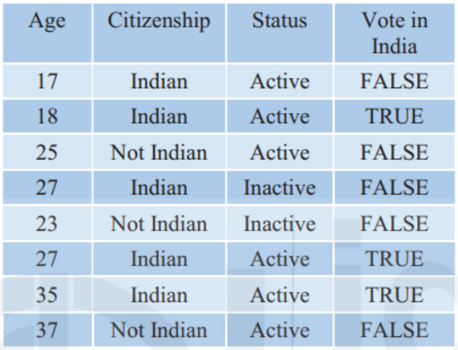
Con1	Con2	Con3	Result
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	0
1	1	0	0
1	1	1	1

In above example 3 different conditions with logical inputs are mentioned as Con1, Con2, Con3 and Result is based on logical inputs to the function as =AND(Cond1, Cond2, Cond3). Once it gets all three conditions TRUE, the result becomes TRUE.

For instance, we can quickly indicate whether or not an entity is eligible to vote in India by using the formula in D11 as the relative cell address and copying it into every other cell in the same column.

=AND(A11>=18,B11="Indian",C11="Active")

Table.11.6: Dataset 2 for AND



Age	Citizenship	Status	Vote in India
17	Indian	Active	FALSE
18	Indian	Active	TRUE
25	Not Indian	Active	FALSE
27	Indian	Inactive	FALSE
23	Not Indian	Inactive	FALSE
27	Indian	Active	TRUE
35	Indian	Active	TRUE
37	Not Indian	Active	FALSE

TRUE & FALSE

These two functions, TRUE() and FALSE(), are mostly used to verify compatibility with other spreadsheets in the workbook. They return the logical values TRUE and FALSE, respectively.

11.7. Loop functions

Loop functions in Excel enable users to automate repetitive activities, repeat over data sets, and execute calculations more effectively. Although Excel lacks built-in loop capabilities like traditional programming languages, users can obtain comparable results by utilizing other techniques and functions accessible inside Excel, such as VBA (Visual Basic for Applications), array formulae, and other built-in features.

Visual Basic for Applications (VBA):

It is the programming language used in Excel for creating macros and automating tasks. VBA allows users to write custom functions and procedures to perform complex operations, including loops. One common loop structure in VBA is the For...Next loop, which repeats a block of code for a specified number of times. For example:

```
Sub Example For Loop()

 Dim i As Integer

For i = 1 To 10

 Cells(i, 1).Value = i * 2  'Multiply each cell value by 2
```

```
Next i

End Sub
```

In this example, the for...Next loop iterates from 1 to 10, updating the value of each cell in column A with the result of multiplying the loop counter i by 2.

Another useful construct in VBA is the Do...Loop loop, which repeats a block of code until a certain condition is met. For instance:

```
Sub Example Do Loop()

Dim i as Integer

i = 1

 Do While i <= 10

Cells (i, 1).Value = i * 2 ‘Multiply each cell value by 2

 i = i + 1

Loop

End Sub
```

This code achieves the same result as the previous example using a Do While loop.

Array Formulas:

Array Formulas allow users to perform calculations on arrays or ranges of data without the need for explicit loops. They process entire ranges of data at once, making them efficient for handling large datasets. By pressing Ctrl + Shift + Enter after entering an array formula, Excel treats it as an array operation. For instance:

```
{=SUM (A1:A10 * B1:B10)}  'Multiply corresponding values in two ranges and sum the products (array formula)
```

Here, the formula multiplies each value in range A1:A10 with the corresponding value in range B1:B10 and then sums the products.

ROW () and COLUMN ():

It functions return the row and column numbers of a cell, respectively. They are often used within other functions to create dynamic references or perform calculations. For example:

=ROW () 'Returns the row number of the cell

=COLUMN () 'Returns the column number of the cell

These functions are useful in constructing formulas that adapt to changes in the spreadsheet structure or size.

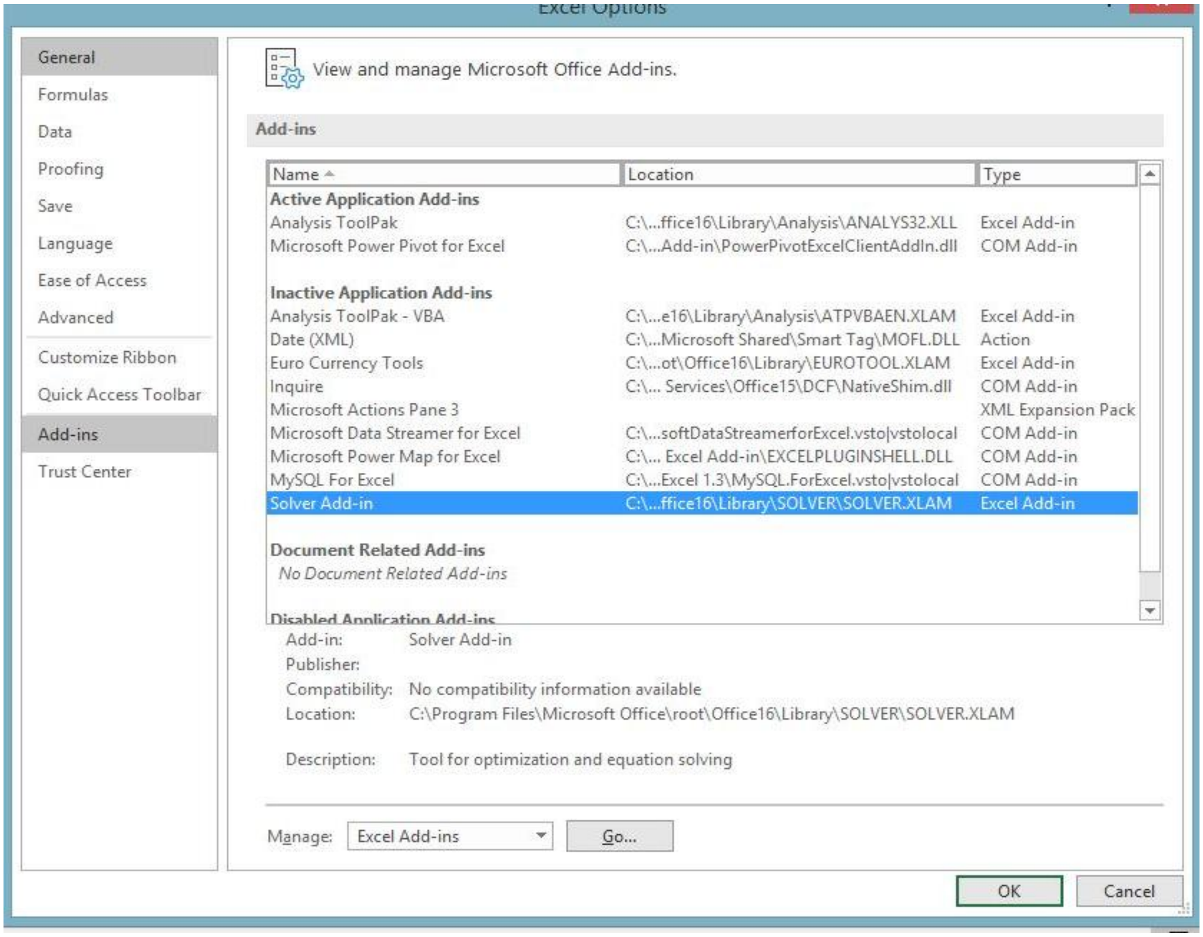
The loop functions in Excel enable users to automate repetitive tasks, iterate through data sets, and perform calculations efficiently. While Excel lacks traditional loop constructs, techniques such as VBA, array formulas, and the use of built-in functions like ROW () and COLUMN () provide powerful alternatives for achieving similar results. By leveraging these tools effectively, users can streamline their workflows and increase productivity in Excel.

11.8. Statistical Analysis with solver by MS Excel

The technique known as "What-If Analysis" involves altering the values of Advanced Excel formulas to experiment with various outcomes. One or more of these Advanced Excel formulae can be applied with numerous distinct sets of values to examine the various outcomes. The best tool for what-if analysis is a solver. It is a Microsoft Excel add-in application that offers numerous benefits. In the cell referred to as the target cell, the characteristic can be utilized to determine the ideal value for a formula. On the other hand, some worksheet formula cell values are subject to certain limitations or restrictions. Decision variables are a set of cells that the solver uses to calculate the formulas in the objective and constraint cells. The solution modifies the value. To work on the constraints on constraint cells, the solver modifies the value of decision variable cells. This procedure helps identify the intended outcome for the target cell. Turning on the Solver Add-in

- On the File tab, click Options.

- Go to Add-ins, select Solver Add-in, and click on the Go button



Name	Location	Type
Active Application Add-ins		
Analysis ToolPak	C:\...ffice16\Library\Analysis\ANALYS32.XLL	Excel Add-in
Microsoft Power Pivot for Excel	C:\...Add-in\PowerPivotExcelClientAddIn.dll	COM Add-in
Inactive Application Add-ins		
Analysis ToolPak - VBA	C:\...e16\Library\Analysis\ATPVBAEN.XLAM	Excel Add-in
Date (XML)	C:\...Microsoft Shared\Smart Tag\MOFL.DLL	Action
Euro Currency Tools	C:\...ot\Office16\Library\EUROTOOL.XLAM	Excel Add-in
Inquire	C:\... Services\Office15\DCF\NativeShim.dll	COM Add-in
Microsoft Actions Pane 3		XML Expansion Pack
Microsoft Data Streamer for Excel	C:\...softDataStreamerforExcel.vsto|vstolocal	COM Add-in
Microsoft Power Map for Excel	C:\... Excel Add-in\EXCELPLUGINSHELL.DLL	COM Add-in
MySQL For Excel	C:\...Excel 1.3\MySQL.ForExcel.vsto|vstolocal	COM Add-in
Solver Add-in	C:\...ffice16\Library\SOLVER\SOLVER.XLAM	Excel Add-in
Document Related Add-ins		
No Document Related Add-ins		
Disabled Application Add-ins		

Add-in: Solver Add-in
Publisher:
Compatibility: No compatibility information available
Location: C:\Program Files\Microsoft Office\root\Office16\Library\SOLVER\SOLVER.XLAM

Description: Tool for optimization and equation solving

Manage: Excel Add-ins ▾ Go...

Fig. 11.44: Excel option for Add-ins

- Check Solver Add-in and click OK

Numerous individuals plan to employ statistical methods in their investigations. Data-driven substantiation of claims is undoubtedly a smart practice. Numerous functions of statistics fall into the following general categories:

a. Summaries and Data Descriptions: To read data quickly, one summarizes and describes the data. Cross-tabulations and graphs are made for nominal or ordinal data, while z-scores are computed for interval or ratio data.

b. Variance and distribution of the data: For nominal and ordinal data, one creates tables, charts, and graphs; for interval and ratio data, one creates box plots with the interquartile range or histograms with a normal curve.

c. Compare groups: When comparing two or more populations, cross-tabulations are used for nominal and ordinal data, and hypothesis testing is used for continuous and numeric data that has been separated into groups.

d. Determine relationships: For nominal or ordinal data, cross-tabulations are used; for interval or ratio data, the correlation coefficient and scatter plot are computed; and for data containing one dependent variable and two or more predictor variables, linear regression or ANOVA is used.

e. Find groups of related cases: Using hierarchical cluster analysis, the k-means cluster analysis problem of finding groups of related instances is resolved. Discriminate analysis is used to determine the traits of established groupings.

f. Find groups of related variables: To find groups of related variables, factor analysis is used.

11.9. Statistical Analysis Using Excel-Descriptive Statistics

Descriptive statistics are used to summarize and describe the important characteristics of a dataset. These characteristics can include measures of central tendency (mean, median, mode), measures of dispersion (range, variance, standard deviation), and measures of distribution (skewness, kurtosis). Performing descriptive statistics in Excel is a common task and can be easily accomplished using built-in functions. The step-by-step guide on how to perform descriptive statistics in Excel is as:

Prepare your data: Enter your data into an Excel spreadsheet. Make sure each column contains data for a single variable, and each row contains a single observation.

Select the data: Click and drag to select the range of cells containing your data.

Access the Data Analysis Toolpak: If you haven't already, you'll need to enable the Data Analysis Toolpak add-in. Go to the "Data" tab on the Excel ribbon, and then click on "Data Analysis" in the Analysis group. If you don't see "Data Analysis," you may need to install the add-in first.

Choose the Descriptive Statistics tool: In the Data Analysis dialog box, scroll down and select "Descriptive Statistics," then click "OK."

Configure the Descriptive Statistics dialog box: In the Descriptive Statistics dialog box, select the input range (your data), and check the appropriate options based on the statistics you want to calculate. You can choose to output the results to a new worksheet or a new location on the current worksheet.

Click "OK": Once you've configured the options, click "OK" to perform the descriptive statistics analysis.

View the results: If you chose to output the results to a new worksheet, Excel will create a new worksheet containing the summary statistics. If you chose to output to a new location on the current worksheet, the results will be displayed there.

Interpret the results: Review the summary statistics to understand the central tendency, dispersion, and shape of your data. Common statistics include mean, median, mode, standard deviation, variance, minimum, maximum, range, skewness, and kurtosis.

Optional: Create charts or graphs: You can also create visualizations of your data, such as histograms or box plots, to further explore its distribution and characteristics.

11.10. Curve Fitting

Curve fitting, also known as regression analysis, is a statistical approach for determining the best-fitting curve or line that characterizes a collection of data points. Excel allows you to do curve fitting using a variety of approaches, including linear regression, polynomial regression, exponential regression, and others. Curve fitting entails determining the mathematical function that best describes a collection of data points. The goal is to reduce the discrepancy between the observed data points and the values predicted by the fitted curve. This enables analysts to spot trends, make predictions, and derive conclusions from the data. the curve fitting in Excel are:

Linear Regression:

Linear regression is used when the relationship between the variables can be approximated by a straight line equation (y = mx + b). Excel provides built-in functions to perform linear regression analysis, such as the LINEST function.

=LINEST(y_data, x_data, TRUE, TRUE) Where:

y_data is the range of dependent variable data.

x_data is the range of independent variable data.

The LINEST function returns an array of regression statistics, including the slope and intercept of the regression line.

Polynomial Regression:

Polynomial regression is used when the relationship between the variables can be approximated by a polynomial equation. Excel does not have a built-in function specifically for polynomial regression, but you can use the LINEST function along with array formulas to perform polynomial regression.

=LINEST(y_data, x_data^{1,2,3,...}, TRUE, TRUE)

Where x_data^{1,2,3,...} represents the independent variable raised to different powers, depending on the degree of the polynomial you want to fit.

Exponential Regression:

Exponential regression is used when the relationship between the variables can be approximated by an exponential equation (y = a * e^(bx)). You can perform exponential regression in Excel using the LINEST function with logarithmic transformation.

```
=LINEST (LN (y_data), x_data, TRUE, TRUE)
```

This formula fits an exponential curve to the data by taking the natural logarithm of the dependent variable.

Other Regression Methods:

Excel also provides additional regression methods such as logarithmic regression, power regression, and moving average regression. These methods can be implemented using a combination of built-in functions and formulas tailored to the specific regression model.

Charting and Visualization:

After performing curve fitting, you can visualize the results by creating a scatter plot of the data points along with the fitted curve. Excel's charting tools allow you to add trend lines to your scatter plot, which can display the regression equation and R-squared value to assess the goodness of fit.

By utilizing these methods and functions, you can perform curve fitting in Excel to analyze and model relationships within your data effectively. Whether you're fitting a straight line, a polynomial curve, or an exponential function, Excel provides the tools necessary to perform regression analysis and gain insights from your data.

Curve fitting has numerous applications in economics, engineering, biology, and social sciences. Curve fitting is used by financial analysts to model stock price fluctuations and predict future trends. In engineering, curve fitting is used to assess experimental data and optimize operations. In biology, it aids researchers in comprehending the link between biological variables.

While curve fitting is an effective tool for data analysis, it is critical to recognize its limitations. Over fitting, multicollinearity, and outliers can reduce the fitted curve's accuracy and reliability. Analysts should carefully analyze the regression model's appropriateness and proceed with caution when interpreting the results.

11.11. Correlation and Regression Analysis and Graphs

Correlation and regression analysis are two important statistical approaches for investigating the relationship between variables and making predictions based on it. In Excel, you can perform these analyses using built-in functions and tools, and create corresponding graphs to visualize the results. This is a thorough guide:

Correlation analysis

Correlation analysis evaluates the strength and direction of a link between two variables. The correlation coefficient, indicated as r, can range from -1 to 1. A high correlation (r >0r>0) suggests a direct association, while a negative correlation (r <0r<0) indicates an inverse relationship. A correlation coefficient close to zero shows no linear relationship. Use the CORREL function in Excel to calculate the correlation coefficient. For example, if your data is in columns A and B, you might use this formula:

```
=CORREL (A:A, B:B)
```

This function returns the correlation coefficient between the two variables.

Regression analysis

Regression analysis models the relationship between a dependent variable and one or more independent variables. The most frequent type is linear regression, which uses a straight line to fit data. The equation for a simple linear regression model is y = mx + b, where m is the slope and b is the intercept. The LINEST function in Excel allows you to perform linear regression analysis. For instance, if your dependent variable is in column A and your independent variable is in column B, you would use the formula:

```
=LINEST (A: A, B: B, TRUE, TRUE)
```

Creating Graphs:

Understanding and interpreting correlation and regression analysis results requires a visual representation of the relationship between variables. Excel includes a variety of charting tools to help you construct graphs that effectively demonstrate these relationships.

1. Scatter Plot:

A scatter plot is a common way to visualize the relationship between two variables. Each data point represents a pair of values from the two variables. To create a scatter plot in Excel, select the data and choose "Scatter" from the Insert menu. You can customize the appearance of the plot, including adding axis labels and a title.

2. Trendline:

Adding a trendline to a scatter plot helps visualize the relationship more clearly. Excel allows you to add various types of trendlines, including linear, exponential, logarithmic, and polynomial. Right-click on the data series in the scatter plot, select "Add Trendline," and choose the desired type of trendline. You can also display the equation of the trendline on the chart.

3. Regression Line:

In regression analysis, you can plot the regression line along with the data points to visually represent the fitted model. After calculating the regression coefficients using the LINEST function, you can plot the regression line by adding a new series to the scatter plot and using the regression equation. This allows you to see how well the regression line fits the data.

4. Correlation Matrix:

For multiple variables, you can create a correlation matrix to visualize the correlation coefficients between each pair of variables. Excel's conditional formatting feature allows you to highlight cells based on the correlation strength, making it easy to identify strong and weak correlations.

11.12. Summary

Microsoft Excel allows you to do a variety of computations, from fundamental arithmetic to extensive statistical analysis. Users can enter data into cells and use built-in functions to do computations like SUM, AVERAGE, and MAX. Excel can do mathematical operations, statistical analyses, financial calculations, and logical judgments. Users can build formulas that automate calculations across numerous cells, allowing for more efficient data processing. Excel also has charting features that help you view data trends and patterns, which improves data interpretation. Overall, Excel's varied computing capabilities make it an effective tool for data analysis, financial modeling, and decision-making across multiple sectors. Numeric/Mathematical Functions in Excel execute mathematical operations on numerical quantities, such as addition, subtraction, multiplication, division, exponentiation, and rounding. Examples are SUM for adding values, SQRT for calculating square roots, and ROUND for rounding numbers. Statistical Functions evaluate data sets and provide insights into their properties. They include functions such as AVERAGE for finding the mean, STDEV for standard deviation, and COUNT for determining the amount of variables in a dataset. Logical Functions analyze logical conditions and return either TRUE or FALSE. Examples include IF for conditional statements, AND for testing several conditions, and OR for testing any combination of various conditions. Loop functions in Excel enable users to automate repetitive tasks, iterate through data sets, and perform calculations efficiently. While Excel lacks traditional loop constructs, techniques such as VBA, array formulas, and the use of built-in functions like ROW() and COLUMN() provide powerful alternatives for achieving similar results. By leveraging these tools effectively, users can streamline their workflows and increase productivity in Excel. Curve fitting is a versatile statistical technique used to model relationships between variables in data. In Excel, analysts can perform curve fitting using various methods, including linear regression, polynomial regression, and exponential regression. By visualizing and interpreting the fitted curves, analysts can gain insights into the underlying patterns and make data-driven decisions.

Curve fitting in Excel enables analysts to extract valuable information from data, identify trends, and make predictions, thereby facilitating informed decision-making across various domains. Correlation and regression analysis are strong tools for determining the relationships between variables and formulating data-driven predictions. Excel allows you to do these studies with built-in functions and tools, as well as construct graphs to efficiently illustrate the results. By understanding correlation coefficients, regression coefficients, and graphical representations, you can obtain insights into the data's underlying patterns and make more informed decisions.

11.13. Further suggested reading

Q. 1. Give some examples of financial functions.

Answer: ---

Q. 2. What do you understand by Array Formulas?

Answer: ---

Q. 3. Discuss the Correlation and Regression Analysis and Graphs.

Answer: ---

Q. 4. Write the Function Specifically.

Answer: ---

Q. 5. Discuss about Mathematical Functions.

Answer: ---

Q. 6. Write about Statistical Functions.

Answer: ---

Q. 7. Discuss the Logical Function.

Answer: ---

Q. 8. Discuss the Statistical Analysis Using Excel-Descriptive Statistics.

Answer: ---

11.14. Further suggested readings

1. Andrie de Vries and Joris Meys, R Programming for Dummies, 2edition, Wiley publication
2. John Paul Mueller and Luca Massaron, Machine Learning (in Python and R) for Dummies, Wiley publication
3. **www.support.office.com**
4. **www.tutorialspoint.com**
5. **www.chandoo.org**
6. **www.ignou.ac.in**
7. **www.contextures.com**

Unit -12: Statistical Software

Structure

12.1. Introduction.

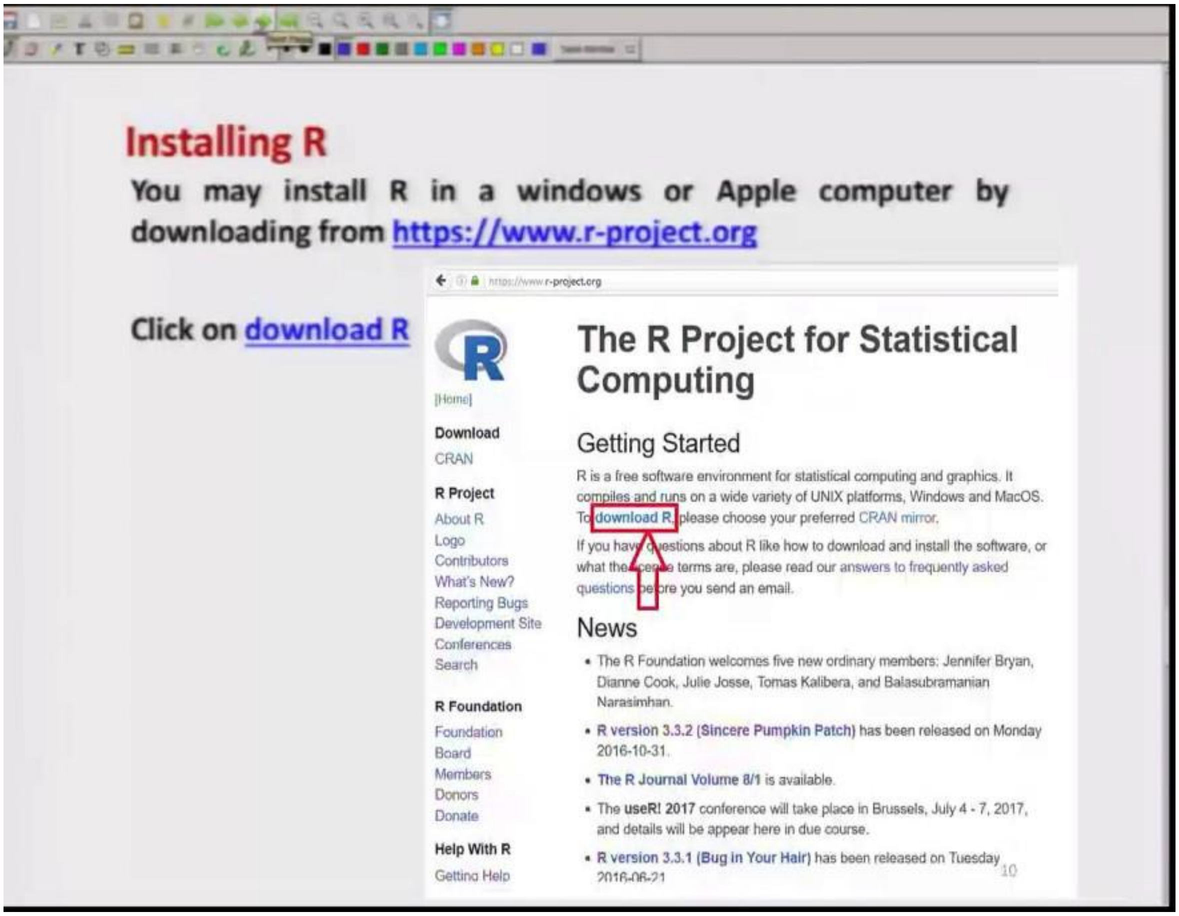
How we analyze data has changed dramatically in recent years. With the advent of personal computers and the internet, the sheer volume of data we have available has grown enormously. The science of data analysis (statistics, psychometrics, econometrics, and machine learning) has kept pace with this explosion of data. Before personal computers and the internet, new statistical methods were developed by academic researchers who published their results as theoretical papers in professional journals. Today new methodologies appear daily. The advent of personal computers had another effect on the way we analyze data. When data analysis was carried out on mainframe computers, computer time was precious and difficult to come by. Analysts would carefully set up a computer run with all the parameters and options thought to be needed. Today's data analysts need to access data from a wide range of sources (database management systems, text files, statistical packages, and spreadsheets), merge the pieces of data together, clean and annotate them, analyze them with the latest methods, present the findings in meaningful and graphically appealing ways, and incorporate the results into attractive reports that can be distributed to stakeholders and the public. As we will see in the following pages, R is a comprehensive software package that's ideally suited to accomplish these goals. R is essentially an environment for the data Manipulation, statistical computing as well as graphical display and data analysis, right. R is just like any other software. There are different types of software which helps us in mathematical calculation and statistical data analysis. R is software, and R has an advantage that R can do data manipulation; R can do statistical computing as well as simulations, R is capable of graphical display and R can also help us in doing different type of data analysis. R has an effective way of data handling. So, it can handle the data easily and it can store the input and output variables. This can store the outcome in the form of a scalar as well as form of a vectors or a matrix, and in R software, simple calculations are possible as well as complicated calculations are also possible and it is not difficult, the mathematical calculation like addition, subtraction and this vectors and matrices, everything is possible, just like any other software.

Objectives

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

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

12.2 Basics of R software.



Installing R

You may install R in a windows or Apple computer by downloading from https://www.r-project.org

Click on download R

https://www.r-project.org

[Home]

Download

CRAN

R Project

About R
Logo
Contributors
What's New?
Reporting Bugs
Development Site
Conferences
Search

R Foundation

Foundation
Board
Members
Donors
Donate

Help With R

Getting Help

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred CRAN mirror.

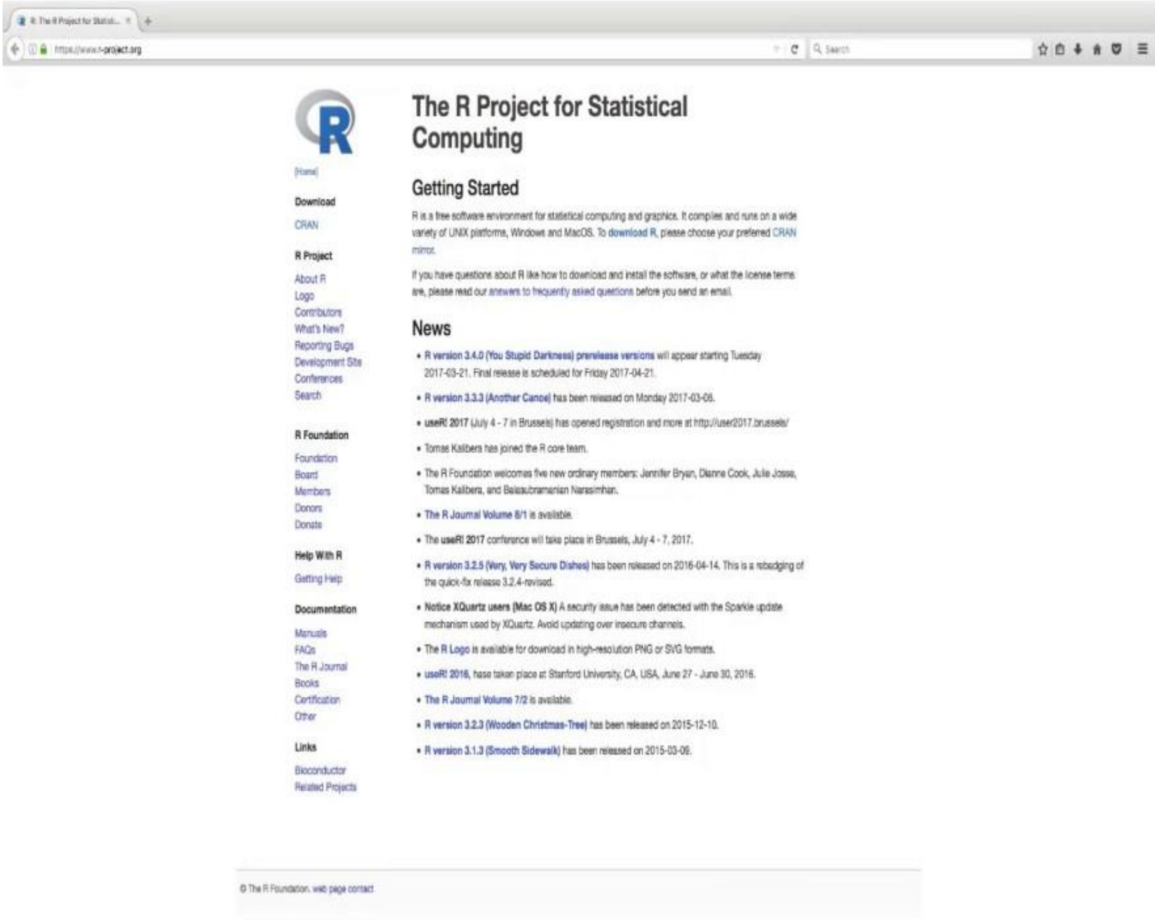
If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News

- The R Foundation welcomes five new ordinary members: Jennifer Bryan, Dianne Cook, Julie Josse, Tomas Kalibera, and Balasubramanian Narasimhan.
- R version 3.3.2 (Sincere Pumpkin Patch) has been released on Monday 2016-10-31.
- The R Journal Volume 8/1 is available.
- The useR! 2017 conference will take place in Brussels, July 4 - 7, 2017, and details will be appear here in due course.
- R version 3.3.1 (Bug in Your Hair) has been released on Tuesday 2016-06-21

10

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



The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

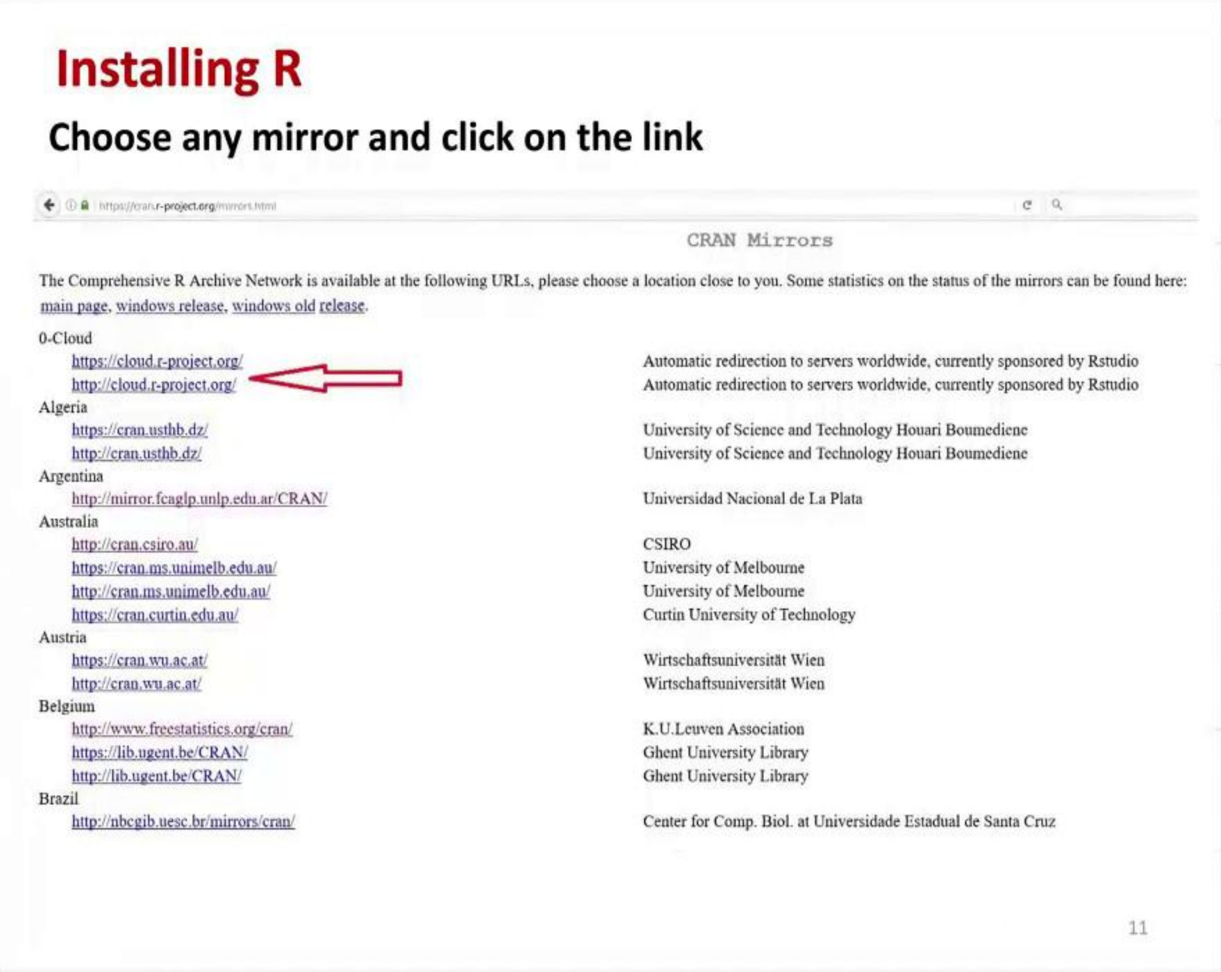
News

- R version 3.4.0 (You Stupid Darkness) prerelease versions will appear starting Tuesday 2017-03-21. Final release is scheduled for Friday 2017-04-21.
- R version 3.3.3 (Another Canoe) has been released on Monday 2017-03-06.
- useR! 2017 (July 4 - 7 in Brussels) has opened registration and more at http://user2017.brussels/
- Tomas Kalibera has joined the R core team.
- The R Foundation welcomes five new ordinary members: Jennifer Bryan, Dianne Cook, Julie Josse, Tomas Kalibera, and Balasubramanian Narasimhan.
- The R Journal Volume 8/1 is available.
- The useR! 2017 conference will take place in Brussels, July 4 - 7, 2017.
- R version 3.2.5 (Very, Very Secure Dishes) has been released on 2016-04-14. This is a rebadging of the quick-fix release 3.2.4-revised.
- Notice XQuartz users (Mac OS X) A security issue has been detected with the Sparkle update mechanism used by XQuartz. Avoid updating over insecure channels.
- The R Logo is available for download in high-resolution PNG or SVG formats.
- useR! 2016, hase taken place at Stanford University, CA, USA, June 27 - June 30, 2016.
- The R Journal Volume 7/2 is available.
- R version 3.2.3 (Wooden Christmas-Tree) has been released on 2015-12-10.
- R version 3.1.3 (Smooth Sidewalk) has been released on 2015-03-09.

© The R Foundation. web page contact

So, you can see here. Now what we have done just for the sake of convenience and in order to illustrate it, we have taken a screenshot of this webpage and we have copied it here. So, we showed that this will not create any confusion for you, but it will help us in understanding the things. So, once you come to this homepage, then what you have to do here that you need to go here, there is a command here download R.

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



Installing R

Choose any mirror and click on the link

CRAN Mirrors

The Comprehensive R Archive Network is available at the following URLs, please choose a location close to you. Some statistics on the status of the mirrors can be found here: main page, windows release, windows old release.

0-Cloud	
https://cloud.r-project.org/	Automatic redirection to servers worldwide, currently sponsored by Rstudio
http://cloud.r-project.org/	Automatic redirection to servers worldwide, currently sponsored by Rstudio
Algeria	
https://cran.usthb.dz/	University of Science and Technology Houari Boumediene
http://cran.usthb.dz/	University of Science and Technology Houari Boumediene
Argentina	
http://mirror.fcaglp.unlp.edu.ar/CRAN/	Universidad Nacional de La Plata
Australia	
http://cran.csiro.au/	CSIRO
https://cran.ms.unimelb.edu.au/	University of Melbourne
http://cran.ms.unimelb.edu.au/	University of Melbourne
https://cran.curtin.edu.au/	Curtin University of Technology
Austria	
https://cran.wu.ac.at/	Wirtschaftsuniversität Wien
http://cran.wu.ac.at/	Wirtschaftsuniversität Wien
Belgium	
http://www.freestatistics.org/cran/	K.U.Leuven Association
https://lib.ugent.be/CRAN/	Ghent University Library
http://lib.ugent.be/CRAN/	Ghent University Library
Brazil	
http://nbcgib.uesc.br/mirrors/cran/	Center for Comp. Biol. at Universidade Estadual de Santa Cruz

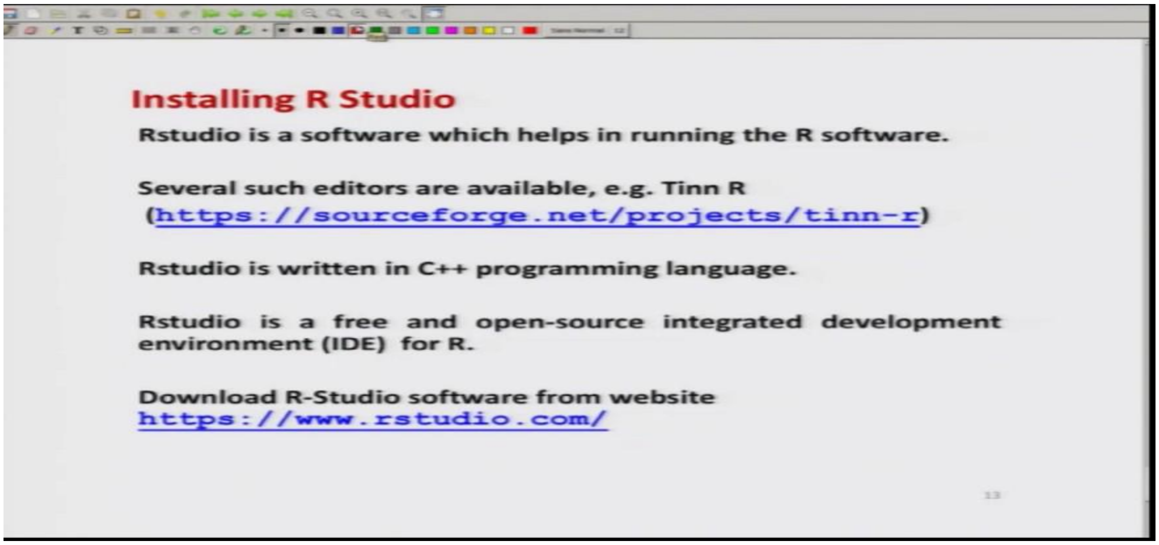
11

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

12.3. R studio and R-Commander.

R Studio software as we had discussed earlier, that is free software that can be downloaded from the website and this is actually a sort of interface between R software and us. Whatever information is contained in R, whatever execution is being done in R, they can be seen through R Studio also. And usually it is more helpful to work in R Studio rather than working directly into the R software. And particularly if you are beginning to learn the R software, this is more helpful because you can see each and every thing just before your eyes in a single shot. And whenever you are trying to write down the program, you are trying to write down the code of a program, and then it is easier actually, at every step. For example, you can highlight you can run and you can check whether your commands are working fine or not. In case if you find any mistake, at the same step you can correct it.

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



Installing R Studio

Rstudio is a software which helps in running the R software.

Several such editors are available, e.g. Tinn R (https://sourceforge.net/projects/tinn-r)

Rstudio is written in C++ programming language.

Rstudio is a free and open-source integrated development environment (IDE) for R.

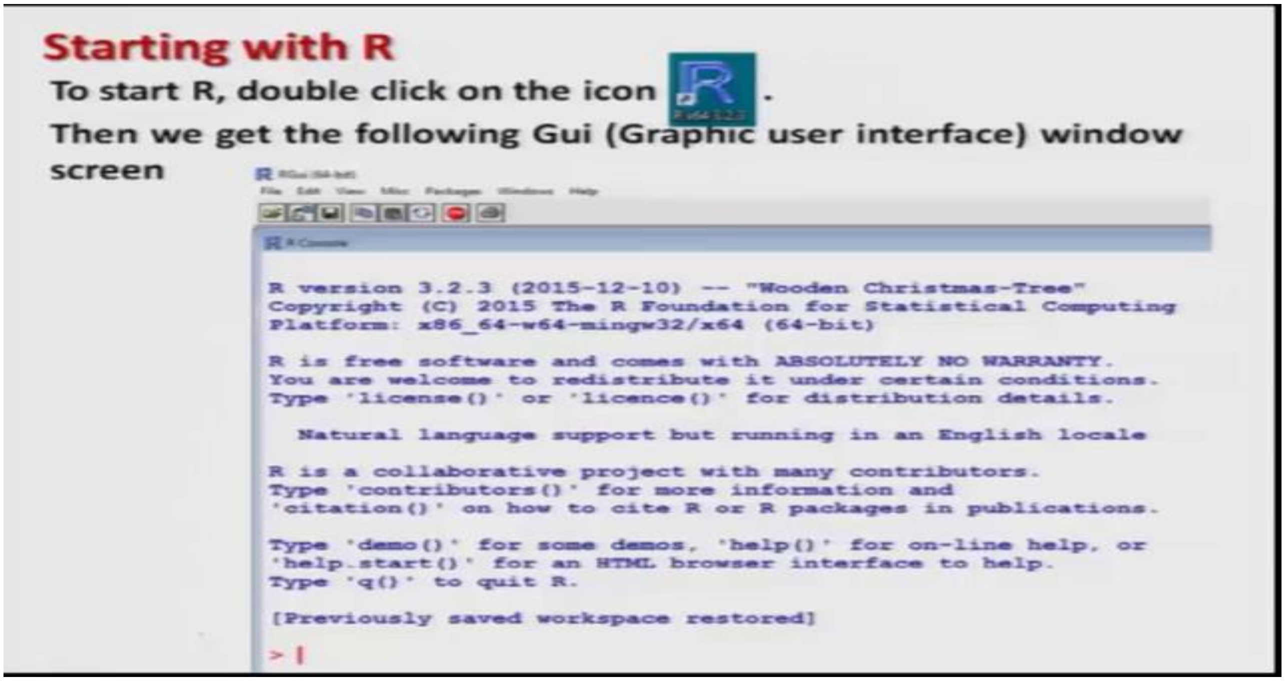
Download R-Studio software from website https://www.rstudio.com/

13

Now, after this, we come to another aspect. Whenever we are working with R software, we have two options- either we can work directly with the R software, we can execute the command inside the R software and second option is that I can take help of supplementary software which works inside the R from outside. So, so there are different types of software which are which are available and in this course, we are going to use software, what is called as R Studio. So, R Studio is essentially software which helps in the execution of R software and beside R Studio, there are other types of software which are available. For example, one of another software is the Tinn R and this can be downloaded from this site and, but I can use any one actually. we are not trying to say at all that either Tinn R is better than R Studio or vice versa. We have chosen R Studio to work. So, this R Studio is actually written in the C++ Programming language.

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

12.4. Creation of Data files.



Starting with R

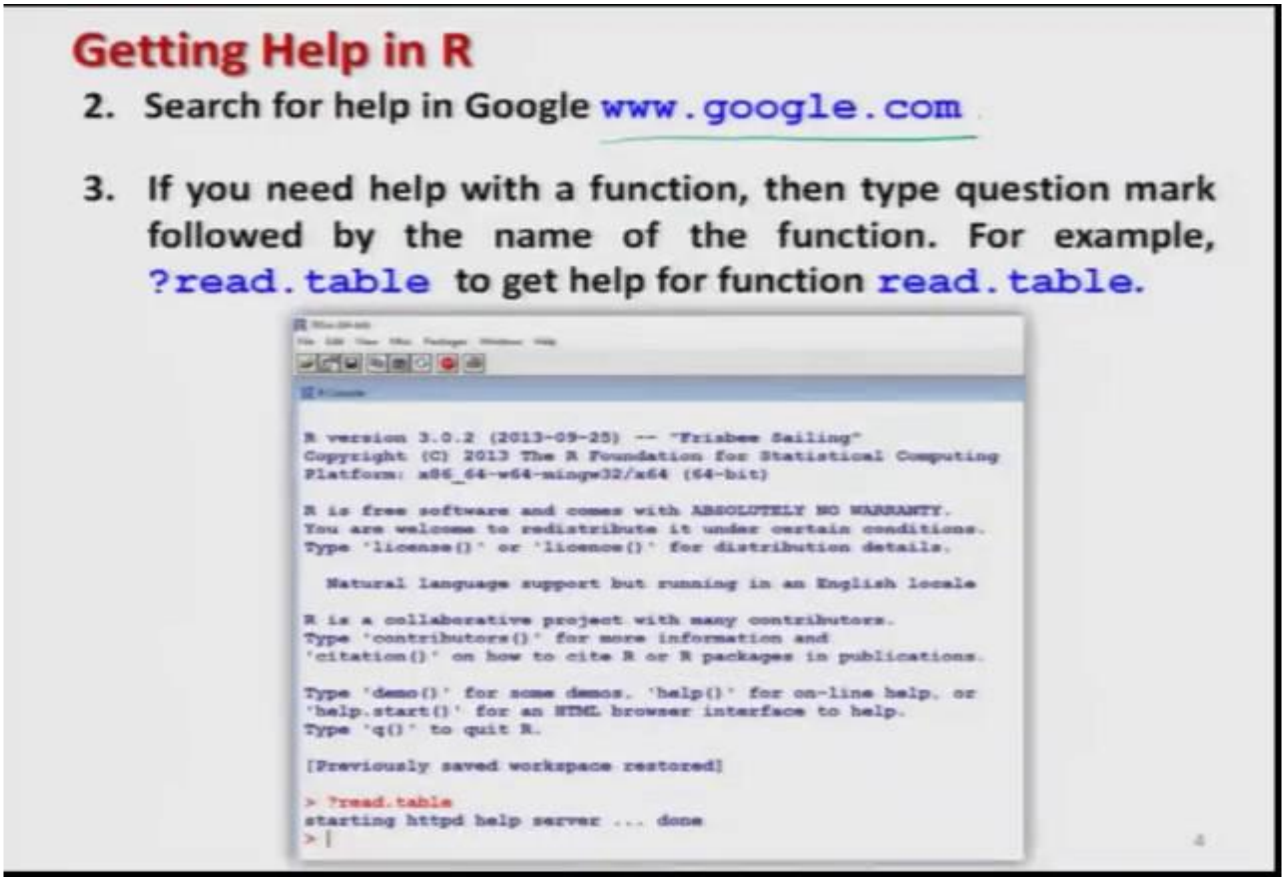
To start R, double click on the icon .

Then we get the following Gui (Graphic user interface) window screen

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

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

  Natural language support but running in an English locale

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

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

[Previously saved workspace restored]

> |
```

So now, you can see here on the desktop of your computer that there is going to be an icon of here R like this. So you need to double click on this icon, right. So, let us try to do it here. So now, we have taken a screenshot of the R GUI and this is pasted over here and now let us tries to see what is there in this window. If you try to read it here for the first 2 paragraph, it is trying to give you the details about the R software and in the last 2 paragraphs; it is trying to give you different options. For example, how to get a citation here and how to obtain the Demonstration of a particular function, how to get here a help and how to quit the R program and so on.

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



Getting Help in R

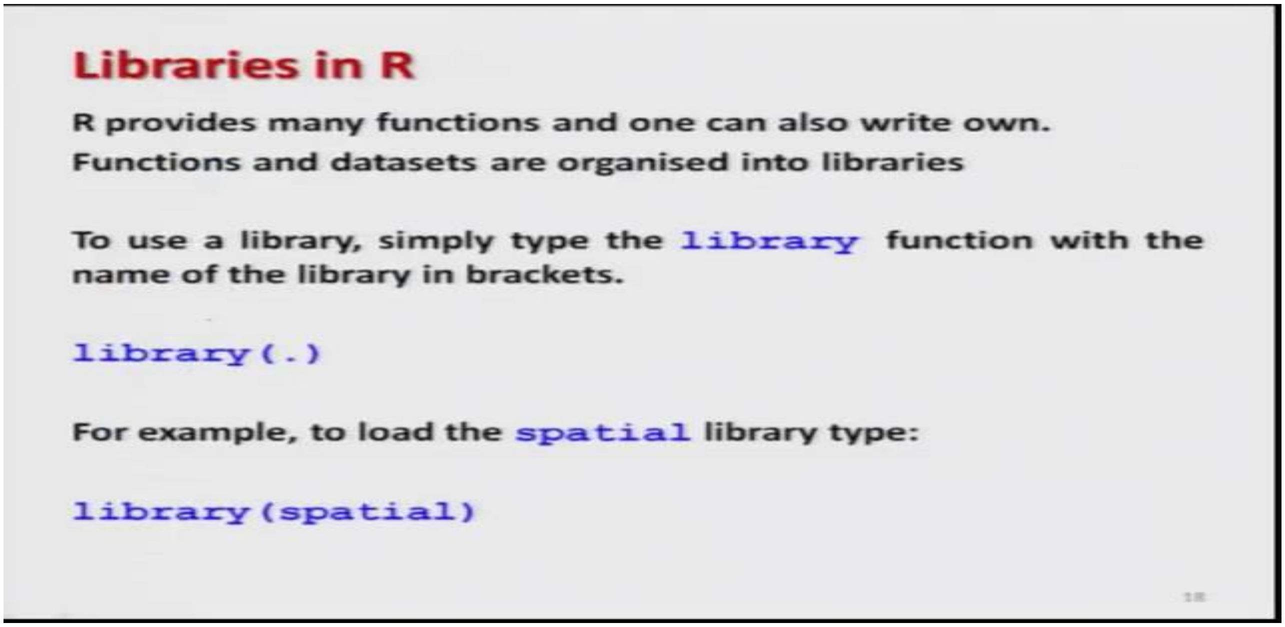
2. Search for help in Google www.google.com

3. If you need help with a function, then type question mark followed by the name of the function. For example, ?read.table to get help for function read.table.

```
R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

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

  Natural language support but running in an English locale

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

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

[Previously saved workspace restored]

> ?read.table
starting httpd help server ... done
> |
```

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

So, suppose we want to obtain some help on the read.table function, what we have to do here, we simply have to write say question mark and after this we have to type here the function name. For example, here we have done here question mark followed by read.table and we try to type this read.table on the command line. For example, we will simply try to copy and paste over here so that we save some time and we tried to paste it over here.



Libraries in R

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

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

```
library(.)
```

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

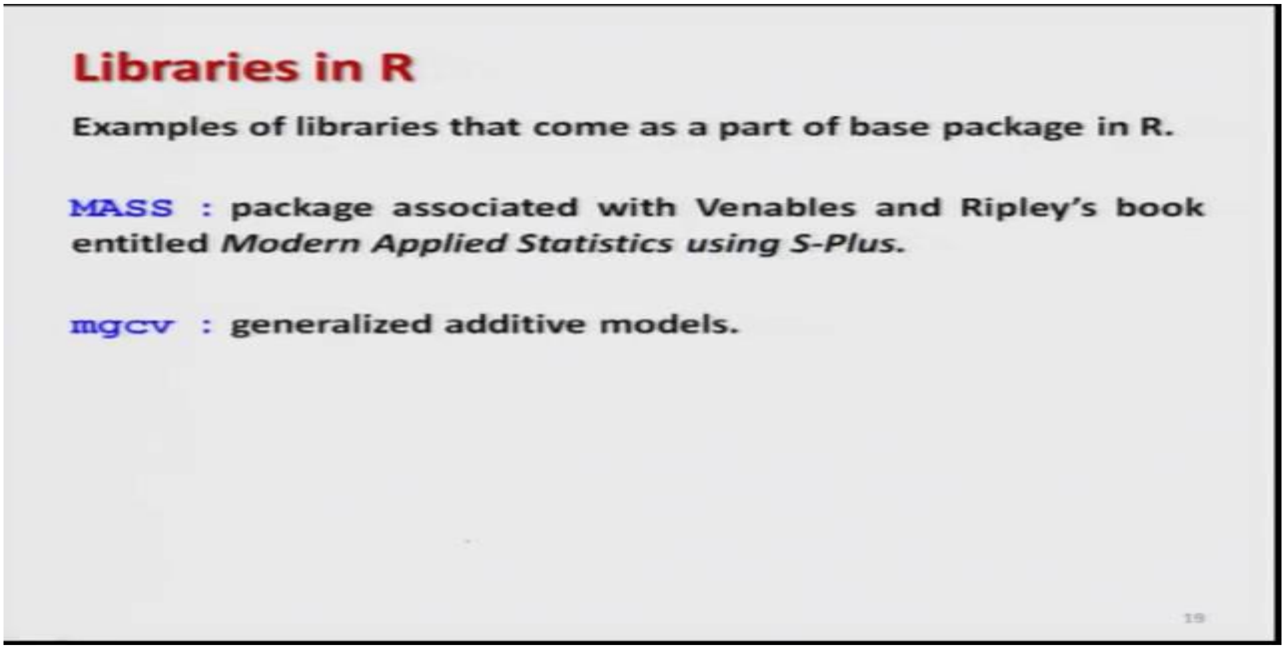
```
library(spatial)
```

18

The next aspect on which we would like to concentrate is the libraries in R. The first question comes, what is a library? So, if you try to see, in simple words, the literal meaning of the library is that a place where there is a collection of many books. So similarly, these libraries are also the collection of several types of commands to execute different types of tasks and R provides many many functions and out of which there are two types of functions one function which are built in inside the R software and say another type of software which you can create.

Beside those things, when you try to install the R software, you will see that there are several options given. One option is when you are starting the R for the first time, and then you are actually downloading the base software. So in the base software they have given most of the say these common commands which are useful for a common user and beside that if you want to use any specialized function, then they have given the commands for that specialized task inside a library. For example, in case if we want to fit the times series model, so, there will be a library for the time series. Suppose if we

want to fit a fit a special data model, so there will be another library that is dealing with the spatial data.



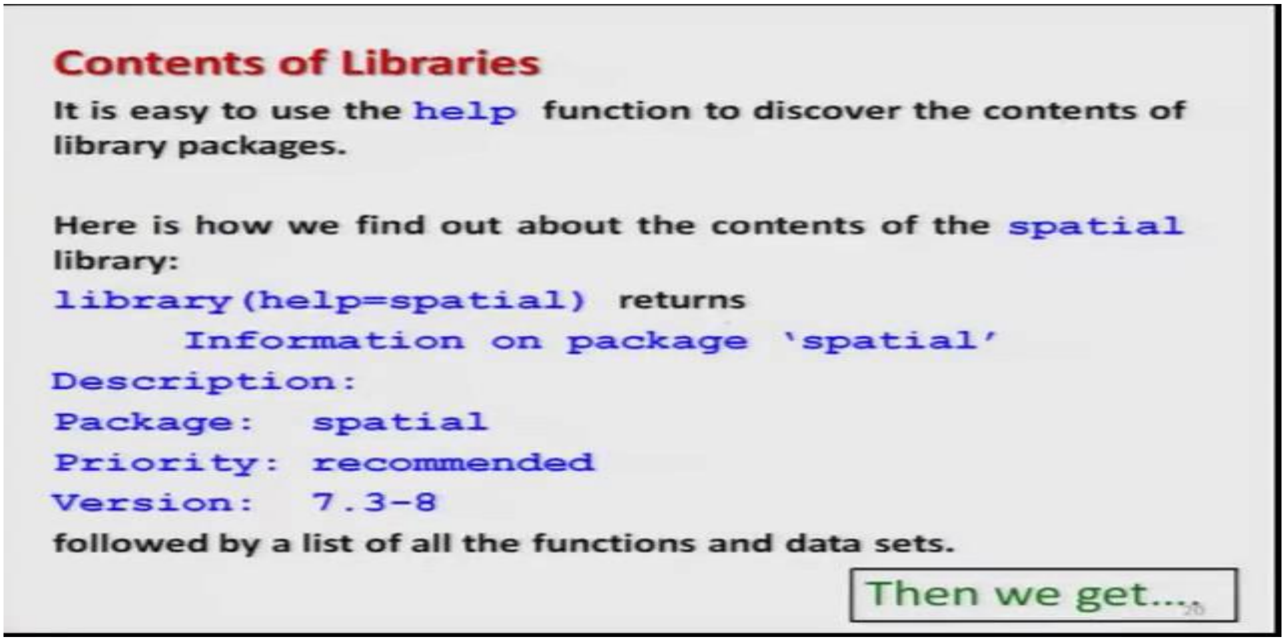
Libraries in R

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

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

mgcv : generalized additive models.

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



Contents of Libraries

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

Here is how we find out about the contents of the spatial library:

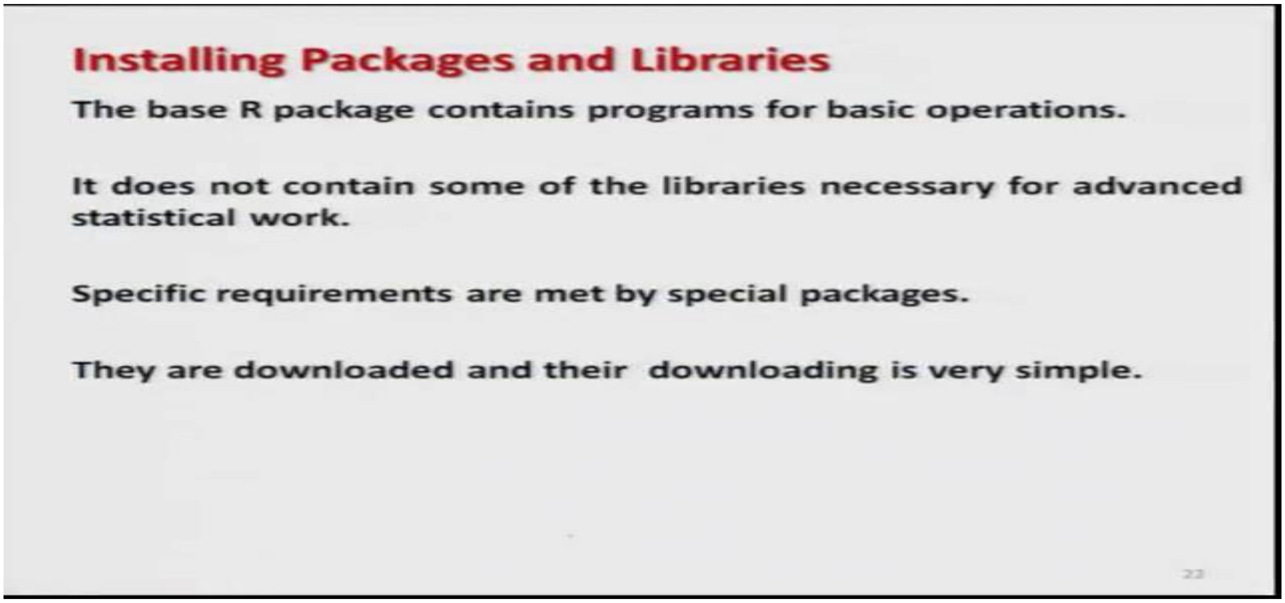
```
library(help=spatial)  returns
      Information on package 'spatial'
Description:
Package:   spatial
Priority:  recommended
Version:   7.3-8
```

followed by a list of all the functions and data sets.

Then we get...

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

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



Installing Packages and Libraries

The base R package contains programs for basic operations.

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

Specific requirements are met by special packages.

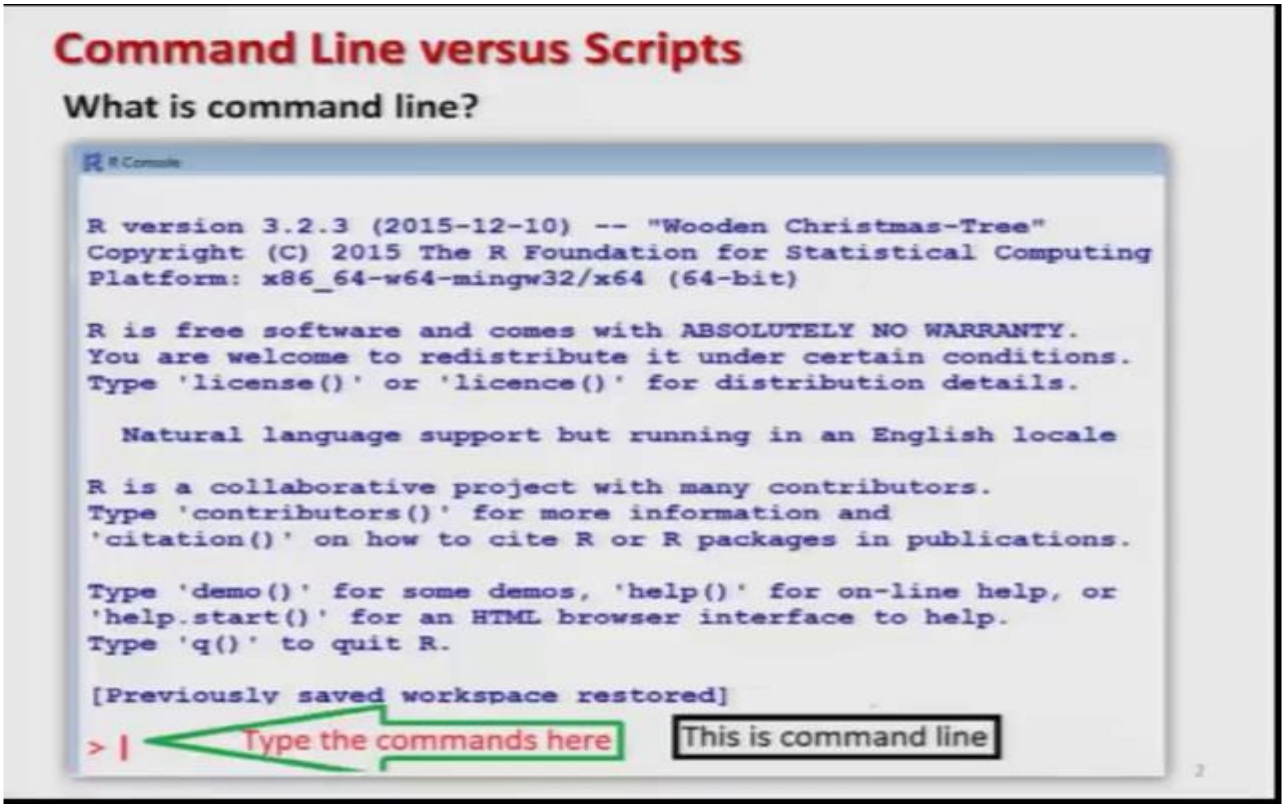
They are downloaded and their downloading is very simple.

22

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

12.5. Command line, Data Editor and R Studio.

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



Command Line versus Scripts

What is command line?

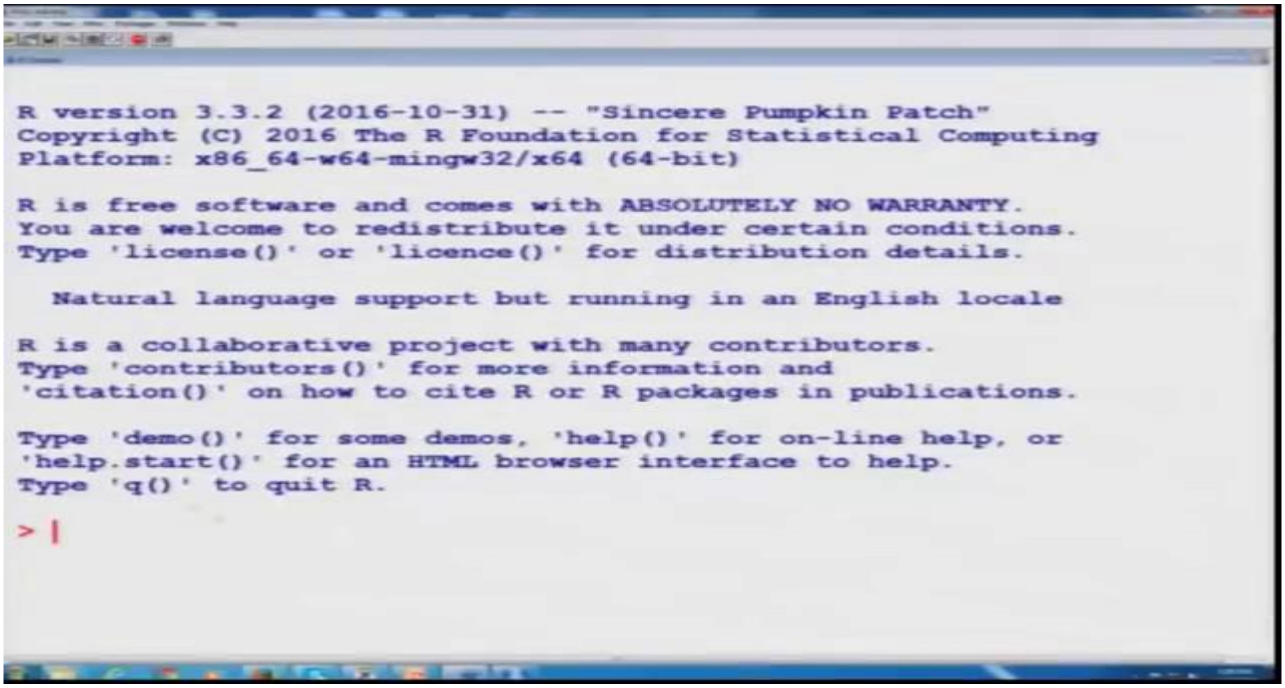
```
R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

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

  Natural language support but running in an English locale

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

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

[Previously saved workspace restored]

> |
```

Type the commands here

This is command line

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

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



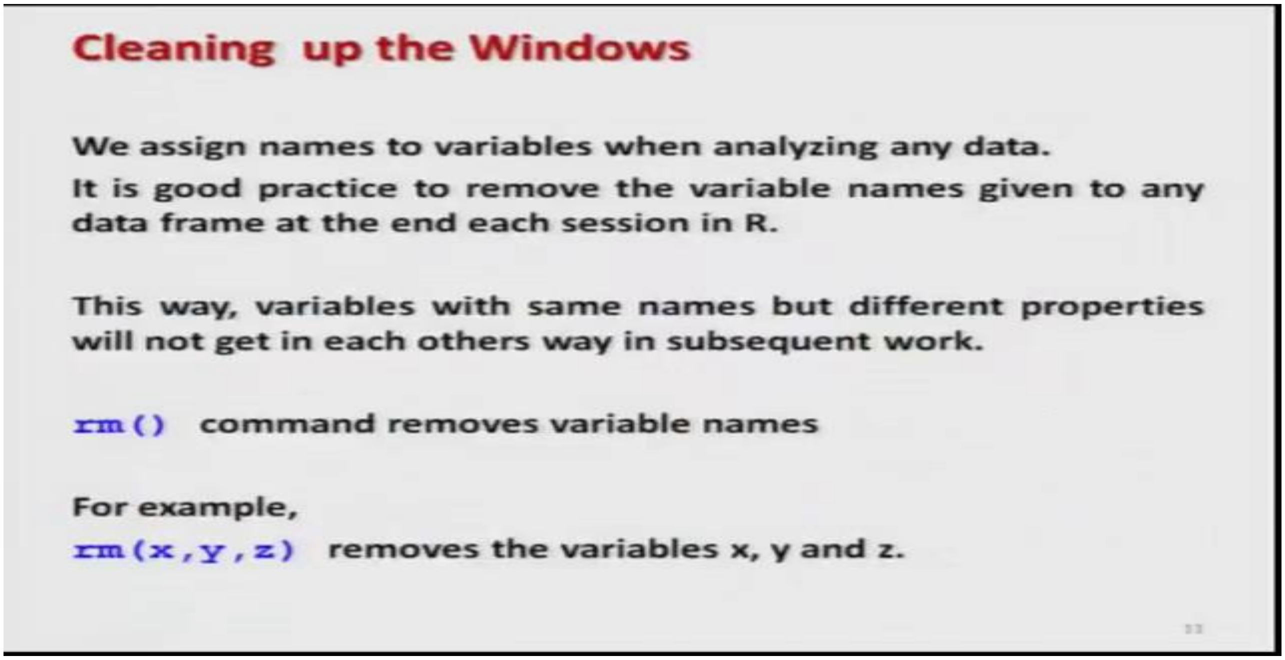
```
R version 3.3.2 (2016-10-31) -- "Sincere Pumpkin Patch"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

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

  Natural language support but running in an English locale

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

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

> |
```

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

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

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

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



Cleaning up the Windows

We assign names to variables when analyzing any data.

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

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

`rm()` command removes variable names

For example,

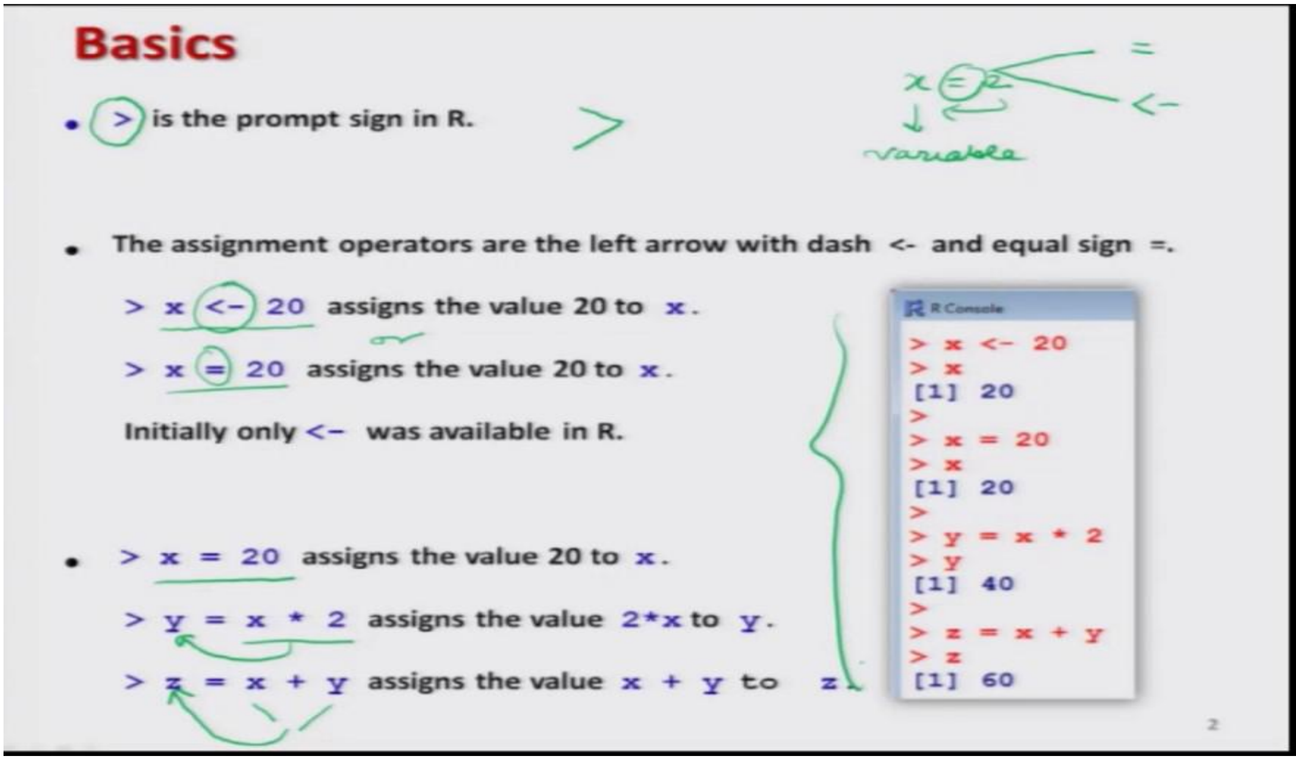
`rm(x,y,z)` removes the variables x, y and z.

Now obviously, once you have done a program, you will get ready for the next program. So, it is always better that you first clean up all the windows, whatever variable names you have defined, whatever information you have defined, and that should be cleaned up. For example, suppose we are writing a program and in which we have used a variable name say age and by age we are trying to denote the ages of some older person. Now we are trying to use another program in which we want to find the age of some children. So obviously, we will try to define as my natural instinct the variable name to be age and we will try to enter my data, but then there can be some contradiction that when we are trying to run the program, possibly it might be using the information contained in the earlier defined variable age that was containing the ages of some elderly persons.

12.6. Basic and R as a Calculator.

Welcome to the next lecture on introduction to R software. You may recall that in the earlier lectures, we had discussed about the basic fundamentals mainly related to how to start and how to work with R. From this lecture onwards, in the next couple of lectures, we will be talking about how to do calculations in R, and again we will say we will be concentrating on the basic fundamentals and my objective is that we should help you so that you can learn the course yourself. So, here again, what we are going to do that I have taken some simple calculations, we will try to show you it online and my request is that you also try to do the same thing yourself on your computer, and not only the examples which we are taking, but try to take more example from your area, from your subject and try to solve them, the more you practice, better you we will be.

One thing we can also accept that here we are trying to copy and paste the commands from my slides and, but we would request you to at least type those command yourself. The advantage of typing the commands yourself is that, you will remember where to put comma and where to put inverted comma, where to put say this say full stop and where to put colon. These things can come only when you type the command yourself. A simple calculator is one where you can do addition, subtraction, multiplication, division and some bracket rules also. So, let us try to start it, but before that let us try to understand the terminology and the symbols and notation used in R.



Basics

- > is the prompt sign in R.
- The assignment operators are the left arrow with dash <- and equal sign =.

  > x <- 20 assigns the value 20 to x.

  > x = 20 assigns the value 20 to x.

  Initially only <- was available in R.

- > x = 20 assigns the value 20 to x.

  > y = x * 2 assigns the value 2*x to y.

  > z = x + y assigns the value x + y to z.

R Console

```
> x <- 20
> x
[1] 20
>
> x = 20
> x
[1] 20
>
> y = x * 2
> y
[1] 40
>
> z = x + y
> z
[1] 60
```

2

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

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

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

12.7 Import Export of Data files.

We are going to understand some concepts related to the import of data. What do we really mean by import of data? You see whenever you want to manipulate any data set, the data set has to be present on your computer and this data set may come from different sources; first option is that you can create that data set yourself. For example, you open a spreadsheet and try to enter the data and even when you are trying to save the spreadsheet, there can be different forms to save that spreadsheet. For example, command separated values, txt format and says some other thing.

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

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

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

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

12.7.1 Import of Excel Data file:

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

```
> library(gdata)

> help(read.xls)

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

Alternatively, we can use the function load Workbook from the XL Connect Package to read the entire workbook, and then load the worksheets with read Worksheet. The XL Connect package requires Java to be pre-installed.

```
> library(XLConnect)

> wk = loadWorkbook("mydata.xls")

> df = readWorksheet(wk,sheet="Sheet1")
```

12.7.2 Import of Minitab Data file

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

```
> library(foreign)

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

12.7.3 Import of SPSS Data file

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

instead.

```
> library(foreign)

> help(read.spss)

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

12.7.4 Import of Table Data File

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

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

> mydata
```

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

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

12.7.5 Import of CSV Data File

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

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

> mydata
```

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

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

12.7.6 Working Directory

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

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

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

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

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

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

12.8 Export of Data files (Importing data Files of Other Software and Redirecting Output)

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

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

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

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

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

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

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

In case if you want to have some more description on the data import and export that can be also found in the R manual which is located at this data set on the website of the R project. So means if you want to have a specific thing, you can always read it from here. So, after learning that how we can read the data from different sources or in different formats, the next objective is that whenever we are trying to run the program, how that program can be saved?

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

12.9 Transformation of Data.

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

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

12.9.1. Using Arrange() method

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

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

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

12.9.2. Numeric

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

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

```
> x = 10.5 # assign a decimal

> x # print x

[1] 10.5

> class(x) # print class name

[1] "numeric"
```

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

```
> k = 1

> k # print k

[1] 1

> class(k) # print class name

[1] "numeric"
```

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

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

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

[1] FALSE
```

12.9.3. Integer

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

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

> y                           # print y

[1] 3

> class(y)                    # print class name

[1] "integer"

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

[1] TRUE
```

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

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

[1] 3
```

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

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

[1] 5
```

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

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

[1] NA

Warning message:

NAs introduced by coercion
```

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

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

```
> as.integer(TRUE)

[1] 1

> as.integer(FALSE)

[1] 0
```

12.9.4. Complex

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

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

> z # print z

[1] 1+2i

> class(z) # print class name

[1] "complex"
```

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

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

[1] NaN

Warning message:

In sqrt(-1) : NaNs produced
```

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

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

[1] 0+1i
```

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

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

[1] 0+1i
```

12.9.5.Logical

A logical value is often created via comparison between variables.

```
> x = 1; y = 2 # sample values
> z = x > y # is x larger?
> z # print result
[1] FALSE
> class(z) # print class
[1] "logical"
```

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

```
> u = TRUE; v = FALSE
> u & v # u AND v
[1] FALSE
> u | v # u OR v
[1] TRUE
> !u # negation of u
[1] FALSE
```

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

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

12.9.6. Character

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

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

We can find out its length with the function nchar.

```
> nchar(s)

[1] 27
```

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

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

> x # print x

[1] "3.14"

> class(x) # print class name

[1] "character"
```

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

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

> paste(fname, lname)

[1] "Joe Smith"
```

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

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

+ "Sam", 100)

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

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

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

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

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

```
> sub("little", "big",
+ "Mary has a little lamb.")
[1] "Mary has a big lamb."
```

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

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

12.9.7. Factor

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

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

The following confirms that the object is indeed a factor.

```
> class(a)
[1] "factor"
```

In particular, we can create factors from numbers:

```
> x = factor(1)
> y = factor(2)
```

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

```
> x + y
[1] NA
Warning message:
In Ops.factor(x, y) : + not meaningful for factors
```

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

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

12.9.8 Vector

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

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

[1] 2 3 5
```

And here is a vector of logical values.

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

[1] TRUE FALSE TRUE FALSE FALSE
```

A vector can contain character strings.

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

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

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

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

[1] 4
```

12.9.10. Combining Vectors

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

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

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

> c(n, s)

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

12.9.11 Value Coercion

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

12.9.13 Vector Arithmetic’s

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

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

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

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

multiplied by 5.

```
> 5 * a

[1] 5 15 25 35
```

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

```
> a + b

[1] 2 5 9 15
```

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

```
> a - b

[1] 0 1 1 -1

> a * b

[1] 1 6 20 56

> a / b

[1] 1.000 1.500 1.250 0.875
```

12.9.14. Logical Index Vector

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

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

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

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

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

> s[L]

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

The code can be abbreviated into a single line.

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

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

12.9.15 Matrix

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

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

12.9.16 Matrix Elements

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

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

> A

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

[1,]    2       4        3
```

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

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

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

> B

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

[1,]     2      1      4

[2,]     5      3      7
```

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

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

[1] 7
```

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

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

[1] 1 5 7
```

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

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

[1] 3 7
```

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

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

[,1]

[1,]    3

[2,]     7
```

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

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

[,1]     [,2]

[1,]     2       3

[2,]    1         7
```

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

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

> A

          col1     col2    col3

row1              2        4         3

row2             1         5         7

> A["row2", "col3"]

[1] 7
```

12.9.17 Data frame

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

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

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

> b = c(TRUE, FALSE, TRUE)

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

12.10 Application of Transformation of Data.

12.10.1 Matrix Construction

There are various ways to construct a matrix.

12.10.2 Transpose

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

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

> B

        [,1]    [,2]

[1,]            2     1

[2,]            4     5

[3,]            3     7
```

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

```
> t(B) # transpose

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

[1,]            2      4      3

[2,]           1      5       7
```

12.10.3 Combining Matrices

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

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

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

> C

[,1]

[1,]    7

[2,]    4

[3,]     2
```

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

```
> cbind(B, C)

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

[1,]             2       1       7

[2,]             4       5       4

[3,]              3      7       2
```

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

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

> D # D has 2 columns

         [,1]    [,2]

[1,]             6        2

> rbind(B, D)

[,1]     [,2]

[1,]     2       1

[2,]      4      5

[3,]     3       7

[4,]     6       2
```

12.10.4 Matrix Arithmetic

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

1.10.6 Addition and Subtraction

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

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

```
[,1]     [,2]
[1,]     1         4
[2,]     2         5
[3,]     3         6
> B = matrix(5:10, nrow=3); B
[,1]      [,2]
[1,]      5        8
[2,]     6         9
[3,]     7         10
```

Then we can compute their **sum**:

```
> A + B
[,1]      [,2]
[1,]      6        12
[2,]      8        14
[3,]      10       16
```

Similarly, we can compute their **difference**:

```
> A - B
[,1]      [,2]
[1,]     -4        -4
[2,]     -4       -4
[3,]     -4       -4
```

12.10.5 Matrix Multiplication

We can multiply two matrices together if the column dimension of the first matrix is the same as the row dimension of the second matrix.

More specifically, if the dimension of the first matrix is *M*x*N*, and the dimension of the second matrix is *N*x*K*, then their matrix product will be of dimension *M*x*K*.

Furthermore, the element at the *m*th row and *n*th column of the product is the *dot product* of the *m*th row of the first matrix with the *n*th column of the second matrix.

For example, consider the following two matrices C and D. As C is a 3x4 matrix, and D is a 4x5 matrix, the column dimension of C matches the row dimension of D.

Hence we can define the matrix product of C and D.

```
> C = matrix(1:12, nrow=3); C

        [,1]    [,2]    [,3]    [,4]

[1,]            1       4        7       10

[2,]            2       5       8       11

[3,]            3       6       9       12
```

```
> D = matrix (-4:15, nrow=4); D

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

[1,]     -4     0        4      8       12

[2,]    -3      1       5       9       13

[3,]    -2      2       6       10       14

[4,]    -1      3       7       11      15
```

We can multiply C and D together using the operator %*% in R. The product is a 3x5matrix as expected.

```
> C %*% D

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

[1,]            -40      48      136     224     312

[2,]            -50      54     158      262    366

[3,]             -60     60     180     300     420
```

12.11 Summary

R is a powerful statistical programming language commonly used for data analysis and visualization. To make the most of R for summer-related projects, consider exploring topics like climate data analysis, environmental studies, or seasonal trends. You could analyze temperature patterns, precipitation, or even create visualizations depicting changes over the summer months. R offers various packages for data manipulation and visualization, such as ggplot2. R Commander is a graphical user interface (GUI) for R that facilitates statistical analysis and data visualization. It provides a point-and-click interface, making it more user-friendly for those who are not comfortable with command-line operations in R. On the other hand, R Studio is an integrated development environment (IDE) for R that includes a console, syntax-highlighting editor, and tools for plotting, history, and workspace management. It doesn't have the same point-and-click interface as R Commander, but it offers a more comprehensive environment for coding in R. Both R Commander and R Studio serve different purposes – R Commander for those who prefer a GUI, and R Studio for a more comprehensive coding experience. Data files play a crucial role in various aspects of computing, analysis, and decision-making. Here are some key points highlighting the importance of data files. Data files are used to store information in a structured format. They provide a means to organize, save, and retrieve data efficiently. Researchers, analysts, and scientists use data files to store datasets for analysis. This includes exploring patterns, trends, and relationships within the data to derive insights

12.12. Self Assessment Question

Q. 1. What is the Basics of R software, discuss it.

Answer: ---

Q. 2. Discuss the R Studio and R-Commander with examples.

Answer: --

Q. 3. What do you about Creation of data files?

Answer: --

Q. 4. Discuss the Command line, Data Editor and R Studio in brief.

Answer: --

Q. 5. Give the suitable example of Basic and R as a calculation.

Answer: --

Q. 6. Discuss about the Transformation of Data with suitable examples.

Answer: --

12.13 Further Readings

1. Andrie de Vries and Joris Meys, R Programming for Dummies, 2edition, Wiley publication
2. John Paul Mueller and Luca Massaron, Machine Learning (in Python and R) for Dummies, Wiley publication
3. **https://www.stitchdata.com/resources/data-transformation/#:~:text=One%20of%20the%20major%20purposes,its%20value%20cannot%20be%20leveraged.**
4. **https://drs.icar.gov.in/Electronic-Book/module5/23Transformation%20of%20Data%20in%20Biological%20Research.pdf**